

Natalia Andrienko  
Gennady Andrienko

# Exploratory Analysis of Spatial and Temporal Data

A Systematic Approach

 Springer

# Exploratory Analysis of Spatial and Temporal Data

Natalia Andrienko · Gennady Andrienko

# Exploratory Analysis of Spatial and Temporal Data

A Systematic Approach

With 245 Figures and 34 Tables

## *Authors*

Natalia Andrienko  
Gennady Andrienko

Fraunhofer Institute AIS  
Schloss Birlinghoven  
53754 Sankt Augustin, Germany  
[gennady.andrienko@ais.fraunhofer.de](mailto:gennady.andrienko@ais.fraunhofer.de)  
<http://www.ais.fraunhofer.de/and>

Library of Congress Control Number: 2005936053

ACM Computing Classification (1998): J.2, H.3

ISBN-10 3-540-25994-5 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-25994-7 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

[springeronline.com](http://springeronline.com)

© Springer-Verlag Berlin Heidelberg 2006  
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typeset by the authors

Production: LE-TeX Jelonek, Schmidt & Vöckler GbR, Leipzig  
Cover design: KünkelLopka Werbeagentur, Heidelberg

Printed on acid-free paper 45/3142/YL - 5 4 3 2 1 0

# Preface

This book is based upon the extensive practical experience of the authors in designing and developing software tools for visualisation of spatially referenced data and applying them in various problem domains. These tools include methods for cartographic visualisation; non-spatial graphs; devices for querying, search, and classification; and computer-enhanced visual techniques. A common feature of all the tools is their high user interactivity, which is essential for exploratory data analysis. The tools can be used conveniently in various combinations; their cooperative functioning is enabled by manifold coordination mechanisms.

Typically, our ideas for new tools or extensions of existing ones have arisen from contemplating particular datasets from various domains. Understanding the properties of the data and the relationships between the components of the data triggered a vision of the appropriate ways of visualising and exploring the data. This resulted in many original techniques, which were, however, designed and implemented so as to be applicable not only to the particular dataset that had incited their development but also to other datasets with similar characteristics. For this purpose, we strove to think about the given data in terms of the generic characteristics of some broad class that the data belonged to rather than stick to their specifics.

From many practical cases of moving from data to visualisation, we gained a certain understanding of what characteristics of data are relevant for choosing proper visualisation techniques. We learned also that an essential stage on the way from data to the selection or design of proper exploratory tools is to envision the questions an analyst might seek to answer in exploring this kind of data, or, in other words, the data analysis tasks. Knowing the questions (or, rather, types of questions), one may look at familiar techniques from the perspective of whether they could help one to find answers to those questions. It may happen in some cases that there is a subset of existing tools that covers all potential question types. It may also happen that for some tasks there are no appropriate tools. In that case, the nature of the tasks gives a clue as to what kind of tool would be helpful. This is an important initial step in designing a new tool.

Having passed along the way from data through tasks to tools many times, we found it appropriate to share the knowledge that we gained from

this process with other people. We would like to describe what components may exist in spatially referenced data, how these components may relate to each other, and what effect various properties of these components and relationships between them may have on tool selection. We would also like to show how to translate the characteristics of data and structures into potential analysis tasks, and enumerate the widely accepted principles and our own heuristics that usually help us in proceeding from the tasks to the appropriate approaches to accomplishing them, and to the tools that could support this. In other words, we propose a methodological framework for the design, selection, and application of visualisation techniques and tools for exploratory analysis of spatially referenced data. Particular attention is paid to spatio-temporal data, i.e. data having both spatial and temporal components.

We expect this book to be useful to several groups of readers. People practising analysis of spatially referenced data should be interested in becoming familiar with the proposed illustrated catalogue of the state-of-the-art exploratory tools. The framework for selecting appropriate analysis tools might also be useful to them. Students (undergraduate and postgraduate) in various geography-related disciplines could gain valuable information about the possible types of spatial data, their components, and the relationships between them, as well as the impact of the characteristics of the data on the selection of appropriate visualisation methods. Students could also learn about various methods of data exploration using visual, highly interactive tools, and acknowledge the value of a conscious, systematic approach to exploratory data analysis. The book may be interesting to researchers in computer cartography, especially those imbued with the ideas of cartographic visualisation, in particular, the ideas widely disseminated by the special Commission on Visualisation of the International Cartographic Association. Our tools are in full accord with these ideas, and our data- and task-analytic approach to tool design offers a way of putting these ideas into practice. It can also be expected that the book will be interesting to researchers and practitioners dealing with any kind of visualisation, not necessarily the visualisation of spatial data. Many of the ideas and approaches presented are not restricted to only spatially referenced data, but have a more general applicability.

The topic of the book is much more general than the consideration of any particular software: we investigate the relations between the characteristics of data, exploratory tasks (questions), and data exploration techniques. We do this first on a theoretical level and then using practical examples. In the examples, we may use particular implementations of the techniques, either our own implementations or freely available demonstrators. However, the main purpose is not to instruct readers in how to use

this or that particular tool but to allow them to better understand the ideas of exploratory data analysis.

The book is intended for a broad reader community and does not require a solid background in mathematics, statistics, geography, or informatics, but only a general familiarity with these subjects. However, we hope that the book will be interesting and useful also to those who do have a solid background in any or all of these disciplines.

## Acknowledgements

This book is a result of a theoretical generalisation of our research over more than 15 years. During this period, many people helped us to establish ourselves and grow as scientists. We would like to express our gratitude to our scientific “parents” Nadezhda Chemeris, Yuri Pechersky, and Sergey Soloview, without whom our research careers would not have started. We are also grateful to our colleagues and partners who significantly influenced and encouraged our work from its early stages, namely Leonid Mikułich, Alexander Komarov, Valeri Gitis, Maria Palenova, and Hans Voss.

Since 1997 we have been working at GMD, the German National Research Centre for Information Technology, which was later transformed into the AIS (Autonomous Intelligent Systems) Fraunhofer Institute. Institute directors Thomas Christaller and Stefan Wrobel and department heads Hans Voss and Michael May always supported and approved our work. All our colleagues were always cooperative and helpful. We are especially grateful to Dietrich Wettschereck, Alexandr Savinov, Peter Gatalsky, Ivan Denisovich, Mark Ostrovsky, Simon Scheider, Vera Hernandez, Andrey Martynkin, and Willi Kloesgen for fruitful discussions and cooperation.

Our research was developed in the framework of numerous international projects. We acknowledge funding from the European Commission and the friendly support of all our partners. We owe much to Robert Peckham, Jackie Carter, Jim Petch, Oleg Chertov, Andreas Schuck, Risto Paivinen, Frits Mohren, Mauro Salvemini, and Matteo Villa. Our work was also greatly inspired by a fruitful (although informal) cooperation with Piotr Jankowski and Alexander Lotov.

Our participation in the ICA commissions on Visualisation and Virtual Environments, Maps and the Internet, and Theoretical Cartography had a strong influence on the formation and refinement of our ideas. Among all the members of these commissions, we are especially grateful to Alan MacEachren, Menno-Jan Kraak, Sara Fabrikant, Jason Dykes, David Fairbain, Terry Slocum, Mark Gahegan, Jürgen Döllner, Monica Wachowicz,

Corne van Elzakker, Michael Peterson, Georg Gartner, Alexander Volodtschenko, and Hans Schlichtmann.

Discussions with Ben Shneiderman, Antony Unwin, Robert Haining, Werner Kuhn, Jonathan Roberts, and Alfred Inselberg were a rich source of inspiration and provided apt occasions to verify our ideas. Special thanks are due to the scientists whose books were formative for our research, namely John Tukey, Jacques Bertin, George Klir, and Rudolf Arnheim.

The authors gratefully acknowledge the encouraging comments of the reviewers, the painstaking work of the copyeditor, and the friendly cooperation of Ralf Gerstner and other people of Springer-Verlag.

We thank our family for the patience during the time that we used for discussing and writing the book in the evenings, weekends, and during vacations.

Almost all of the illustrations in the book were produced using the CommonGIS system and some other research prototypes developed in our institute. Online demonstrators of these systems are available on our Web site <http://www.ais.fraunhofer.de/> and on the web site of our institute department <http://www.ais.fraunhofer.de/SPADE>. People interested in using the software should visit the site of CommonGIS, <http://www.CommonGIS.com>.

The datasets used in the book were provided by our partners in various projects.

**1. Portuguese census.** The data set was provided by CNIG (Portuguese National Centre for Geographic Information) within the EU-funded project CommonGIS (Esprit project 28983). The data were prepared by Joana Abreu, Fatima Bernardo, and Joana Hipolito.

**2. Forests in Europe.** The dataset was created within the project “Combining Geographically Referenced Earth Observation Data and Forest Statistics for Deriving a Forest Map for Europe” (15237-1999-08 F1ED ISP FI). The data were provided to us by EFI (the European Forest Institute within the project EFIS (European Forest Information System), contract number: 17186-2000-12 F1ED ISP FI).

**3. Earthquakes in Turkey.** The dataset was provided within the project SPIN! (Spatial Mining for Data of Public Interest) (IST Programme, project IST-1999-10536) by Valery Gitis and his colleagues.

**4. Migration of white storks.** The data were provided by the German Research Centre for Ornithology of the Max Planck Society within a German school project called “Naturdetektive”. The data were prepared by Peter Gatalsky.

5. **Weather in Germany.** The dataset was published by Deutscher Wetterdienst at the URL [http://www.dwd.de/de/FundE/Klima/KLIS/daten/online/nat/index\\_monatswerte.htm](http://www.dwd.de/de/FundE/Klima/KLIS/daten/online/nat/index_monatswerte.htm). Simon Scheider prepared the data for application of the tools.
6. **Crime in the USA.** The dataset was published by the US Department of Justice, URL <http://bjsdata.ojp.usdoj.gov/dataonline/>. The data were prepared by Mohammed Islam.
7. **Forest management scenarios.** The dataset was created in the project SILVICS (Silvicultural Systems for Sustainable Forest Resources Management) (INTAS EU-funded project). The data were prepared for analysis by Alexey Mikhaylov and Peter Gatalsky.
8. **Forest fires in Umbria.** The dataset was provided within the NEFIS (Network for a European Forest Information Service) project, an accompanying measure in the Quality of Life and Management of Living Resources Programme of the European Commission (contract number QLK5-CT-2002-30638). The data were collected by Regione dell’Umbria, Servizio programmazione forestale, Perugia, Italy; the survey was performed by Corpo Forestale dello Stato, Italy
9. **Health care in Idaho.** The dataset was provided by Piotr Jankowski within an informal cooperation project between GMD and the University of Idaho, Moscow, ID.

August 2005

*Sankt Augustin, Germany*

Natalia Andrienko  
Gennady Andrienko

# Contents

<b>1</b>	<b>Introduction .....</b>	<b>1</b>
1.1	What Is Data Analysis? .....	1
1.2	Objectives of the Book.....	5
1.3	Outline of the Book .....	6
1.3.1	Data .....	6
1.3.2	Tasks .....	8
1.3.3	Tools .....	10
1.3.4	General Principles .....	14
	References .....	16
<b>2</b>	<b>Data .....</b>	<b>17</b>
	Abstract.....	17
2.1	Structure of Data .....	18
2.1.1	Functional View of Data Structure .....	21
2.1.2	Other Approaches .....	25
2.2	Properties of Data.....	27
2.2.1	Other Approaches .....	31
2.3	Examples of Data .....	34
2.3.1	Portuguese Census .....	34
2.3.2	Forests in Europe .....	36
2.3.3	Earthquakes in Turkey .....	36
2.3.4	Migration of White Storks .....	38
2.3.5	Weather in Germany .....	40
2.3.6	Crime in the USA.....	41
2.3.7	Forest Management Scenarios .....	42
	Summary.....	44
	References .....	45
<b>3</b>	<b>Tasks .....</b>	<b>47</b>
	Abstract.....	47
3.1	Jacques Bertin's View of Tasks .....	49
3.2	General View of a Task.....	53

3.3	Elementary Tasks .....	60
3.3.1	Lookup and Comparison.....	61
3.3.2	Relation-Seeking.....	69
3.3.3	Recap: Elementary Tasks.....	75
3.4	Synoptic Tasks .....	81
3.4.1	General Notes.....	81
3.4.2	Behaviour and Pattern .....	83
3.4.3	Types of Patterns.....	91
3.4.3.1	Association Patterns.....	91
3.4.3.2	Differentiation Patterns.....	93
3.4.3.3	Arrangement Patterns .....	94
3.4.3.4	Distribution Summary .....	95
3.4.3.5	General Notes .....	96
3.4.4	Behaviours over Multidimensional Reference Sets .....	98
3.4.5	Pattern Search and Comparison .....	107
3.4.6	Inverse Comparison .....	112
3.4.7	Relation-Seeking.....	115
3.4.8	Recap: Synoptic Tasks .....	119
3.5	Connection Discovery .....	124
3.5.1	General Notes.....	124
3.5.2	Properties and Formalisation .....	127
3.5.3	Relation to the Former Categories .....	134
3.6	Completeness of the Framework .....	139
3.7	Relating Behaviours: a Cognitive-Psychology Perspective .....	143
3.8	Why Tasks?.....	148
3.9	Other Approaches.....	151
	Summary.....	158
	References .....	159
<b>4</b>	<b>Tools.....</b>	<b>163</b>
	Abstract.....	163
4.1	A Few Introductory Notes.....	165
4.2	The Value of Visualisation.....	166
4.3	Visualisation in a Nutshell .....	171
4.3.1	Bertin's Theory and Its Extensions .....	171
4.3.2	Dimensions and Variables of Visualisation .....	182
4.3.3	Basic Principles of Visualisation .....	189
4.3.4	Example Visualisations.....	196
4.4	Display Manipulation.....	207
4.4.1	Ordering .....	207
4.4.2	Eliminating Excessive Detail .....	214
4.4.3	Classification.....	217

---

4.4.4	Zooming and Focusing.....	231
4.4.5	Substitution of the Encoding Function.....	241
4.4.6	Visual Comparison.....	248
4.4.7	Recap: Display Manipulation.....	257
4.5	Data Manipulation.....	259
4.5.1	Attribute Transformation .....	261
4.5.1.1	“Relativisation”.....	261
4.5.1.2	Computing Changes.....	263
4.5.1.3	Accumulation.....	268
4.5.1.4	Neighbourhood-Based Attribute Transformations.....	269
4.5.2	Attribute Integration.....	276
4.5.2.1	An Example of Integration.....	278
4.5.2.2	Dynamic Integration of Attributes.....	279
4.5.3	Value Interpolation .....	288
4.5.4	Data Aggregation .....	293
4.5.4.1	Grouping Methods .....	294
4.5.4.2	Characterising Aggregates .....	297
4.5.4.3	Visualisation of Aggregate Sizes .....	300
4.5.4.4	Sizes Are Not Only Counts.....	312
4.5.4.5	Visualisation and Use of Positional Measures.....	316
4.5.4.6	Spatial Aggregation and Reaggregation .....	327
4.5.4.7	A Few Words About OLAP.....	332
4.5.4.8	Data Aggregation: a Few Concluding Remarks .....	333
4.5.5	Recap: Data Manipulation .....	335
4.6	Querying.....	336
4.6.1	Asking Questions .....	337
4.6.1.1	Spatial Queries.....	341
4.6.1.2	Temporal Queries .....	346
4.6.1.3	Asking Questions: Summary .....	349
4.6.2	Answering Questions .....	351
4.6.2.1	Filtering.....	353
4.6.2.2	Marking.....	363
4.6.2.3	Marking Versus Filtering.....	371
4.6.2.4	Relations as Query Results .....	373
4.6.3	Non-Elementary Queries.....	381
4.6.4	Recap: Querying .....	393
4.7	Computational Tools .....	395
4.7.1	A Few Words About Statistical Analysis.....	397
4.7.2	A Few Words About Data Mining .....	401
4.7.3	The General Paradigm for Using Computational Tools.....	406
4.7.4	Example: Clustering.....	407
4.7.5	Example: Classification .....	415

4.7.6 Example: Data Preparation .....	423
4.7.7 Recap: Computational Tools.....	425
4.8 Tool Combination and Coordination.....	428
4.8.1 Sequential Tool Combination .....	429
4.8.2 Concurrent Tool Combination .....	434
4.8.3 Recap: Tool Combination .....	447
4.9 Exploratory Tools and Technological Progress .....	450
Summary.....	453
References .....	454
<b>5 Principles .....</b>	<b>461</b>
Abstract.....	461
5.1 Motivation .....	463
5.2 Components of the Exploratory Process .....	465
5.3 Some Examples of Exploration.....	467
5.4 General Principles of Selection of the Methods and Tools .....	480
5.4.1 Principle 1: See the Whole.....	481
5.4.1.1 Completeness .....	483
5.4.1.2 Unification .....	494
5.4.2 Principle 2: Simplify and Abstract.....	506
5.4.3 Principle 3: Divide and Group .....	509
5.4.4 Principle 4: See in Relation.....	518
5.4.5 Principle 5: Look for Recognisable .....	530
5.4.6 Principle 6: Zoom and Focus .....	540
5.4.7 Principle 7: Attend to Particulars .....	544
5.4.8 Principle 8: Establish Linkages.....	552
5.4.9 Principle 9: Establish Structure.....	572
5.4.10 Principle 10: Involve Domain Knowledge .....	579
5.5 General Scheme of Data Exploration: Tasks, Principles, and Tools .....	584
5.5.1 Case 1: Single Referrer, Holistic View Possible.....	587
5.5.1.1 Subcase 1.1: a Homogeneous Behaviour.....	588
5.5.1.2 Subcase 1.2: a Heterogeneous Behaviour.....	590
5.5.2 Case 2: Multiple Referrers .....	593
5.5.2.1 Subcase 2.1: Holistic View Possible.....	595
5.5.2.2 Subcase 2.2: Behaviour Explored by Slices and Aspects.....	598
5.5.3 Case 3: Multiple Attributes .....	602
5.5.4 Case 4: Large Data Volume .....	606
5.5.5 Final Remarks .....	611
5.6 Applying the Scheme (an Example).....	613
Summary.....	630

References .....	632
<b>6 Conclusion .....</b>	<b>635</b>
<b>Appendix I: Major Definitions .....</b>	<b>639</b>
I.1 Data .....	639
I.2 Tasks .....	643
I.3 Tools.....	647
<b>Appendix II: A Guide to Our Major Publications Relevant to This Book .....</b>	<b>651</b>
References .....	653
<b>Appendix III: Tools for Visual Analysis of Spatio-Temporal Data Developed at the AIS Fraunhofer Institute .....</b>	<b>657</b>
References .....	658
<b>Index.....</b>	<b>659</b>

# 1 Introduction

## 1.1 What Is Data Analysis?

It seems curious that we have not found a general definition of this term in the literature. In statistics, for example, data analysis is understood as “the process of computing various summaries and derived values from the given collection of data” (Hand 1999, p. 3). It is specially stressed that the process is iterative: “One studies the data, examines it using some analytic technique, decides to look at it another way, perhaps modifying it in the process by transformation or partitioning, and then goes back to the beginning and applies another data analytic tool. This can go round and round many times. Each technique is being used to probe a slightly different aspect of the data – to ask a slightly different question of the data” (Hand 1999, p. 3).

In the area of geographic information systems (GIS), data analysis is often defined as “a process for looking at geographic patterns in your data and at relationships between features” (Mitchell 1999, p. 11). It starts with formulating the question that needs to be answered, followed by choosing a method on the basis of the question, the type of data available, and the level of information required (this may raise a need for additional data). Then the data are processed with the use of the chosen method and the results are displayed. This allows the analyst to decide whether the information obtained is valid or useful, or whether the analysis should be redone using different parameters or even a different method.

Let us look what is common to these two definitions. Both of them define data analysis as an iterative process consisting of the following activities:

- formulate questions;
- choose analysis methods;
- prepare the data for application of the methods;
- apply the methods to the data;
- interpret and evaluate the results obtained.

The difference between statistical analysis and GIS analysis seems to lie only in the types of data that they deal with and in the methods used. In both cases, data analysis appears to be driven by questions: the questions motivate one to do analysis, determine the choice of data and methods, and affect the interpretation of the results. Since the questions are so important, what are they?

Neither statistical nor GIS handbooks provide any classification of possible questions but they give instead a few examples. Here are some examples from a GIS handbook (Mitchell 1999):

- Where were most of the burglaries last month?
- How much forest is in each watershed?
- Which parcels are within 500 feet of this liquor store?

For a comparison, here are some examples from a statistical handbook for geographers (Burt and Barber 1996):

- What major explanatory variables account for the variation in individual house prices in cities?
- Are locational variables more or less important than the characteristics of the house itself or of the neighbourhood in which it is located?
- How do these results compare across cities?

It can be noticed that the example questions in the two groups have discernible flavours of the particular methods available in GIS and statistical analysis, respectively, i.e. the questions have been formulated with certain analysis methods in mind. This is natural for handbooks, which are intended to teach their readers to use methods, but how does this match the actual practice of data analysis?

We believe that questions oriented towards particular analysis methods may indeed exist in many situations, for example, when somebody performs routine analyses of data of the same type and structure. But what happens when an analyst encounters new data that do not resemble anything dealt with so far? It seems clear that the analyst needs to get acquainted with the data before he/she can formulate questions like those cited in the handbooks, i.e. questions that already imply what method to use.

“Getting acquainted with data” is the topic pursued in exploratory data analysis, or EDA. As has been said in an Internet course in statistics, “Often when working with statistics we wish to answer a specific question – such as does smoking cigars lead to an increased risk of lung cancer? Or does the number of keys carried by men exceed those carried by women? ... However sometimes we just wish to explore a data set to see what it

might tell us. When we do this we are doing Exploratory Data Analysis” (STAT 2005).

Although EDA emerged from statistics, this is not a set of specific techniques, unlike statistics itself, but rather a philosophy of how data analysis should be carried out. This philosophy was defined by John Tukey (Tukey 1977) as a counterbalance to a long-term bias in statistical research towards developing mathematical methods for hypothesis testing. As Tukey saw it, EDA was a return to the original goals of statistics, i.e. detecting and describing patterns, trends, and relationships in data. Or, in other words, EDA is about hypothesis generation rather than hypothesis testing.

The concept of EDA is strongly associated with the use of graphical representations of data. As has been said in an electronic handbook on engineering statistics, “Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out” (NIST/SEMATECH 2005).

Is the process of exploratory data analysis also question-driven, like traditional statistical analysis and GIS analysis? On the one hand, it is hardly imaginable that someone would start exploring data without having any question in mind; why then start at all? On the other hand, if any questions are asked, they must be essentially different from the examples cited above. They cannot be so specific and cannot imply what analysis method will be used. Appropriate examples can be found in George Klir’s explanation of what empirical investigation is (Klir 1985).

According to Klir, a meaningful empirical investigation implies an object of investigation, a purpose of the investigation of the object, and constraints imposed upon the investigation. “The *purpose of investigation* can be viewed as a set of questions regarding the object which the investigator (or his client) wants to answer. For example, if the object of investigation is New York City, the purpose of the investigation might be represented by questions such as ‘How can crime be reduced in the city?’ or ‘How can transportation be improved in the city?’; if the object of investigation is a computer installation, the purpose of investigation might be to answer questions ‘What are the bottlenecks in the installation?’, ‘What can be done to improve performance?’, and the like; if a hospital is investigated, the question might be ‘How can the ability to give immediate care to all emergency cases be increased?’, ‘How can the average time spent by a

patient in the hospital be reduced?”, or ‘What can be done to reduce the cost while preserving the quality of services?’; if the object of interest of a musicologist is a musical composer, say Igor Stravinsky, his question is likely to be ‘What are the basic characteristics of Stravinsky’s compositions which distinguish him from other composers?’ ” (Klir 1985, p. 83). Although Klir does not use the term “exploratory data analysis”, it is clear that exploratory analysis starts after collecting data about the object of investigation, and the questions representing the purpose of investigation remain relevant.

According to the well-known “Information Seeking Mantra” introduced by Ben Shneiderman (Shneiderman 1996), EDA can be generalised as a three-step process: “Overview first, zoom and filter, and then details-on-demand”. In the first step, an analyst needs to get an overview of the entire data collection. In this overview, the analyst identifies “items of interest”. In the second step, the analyst zooms in on the items of interest and filters out uninteresting items. In the third step, the analyst selects an item or group of items for “drilling down” and accessing more details. Again, the process is iterative, with many returns to the previous steps. Although Shneiderman does not explicitly state this, it seems natural that it is the general goal of investigation that determines what items will be found “interesting” and deserving of further examination.

On this basis, we adopt the following view of EDA. The analyst has a certain purpose of investigation, which motivates the analysis. The purpose is specified as a general question or a set of general questions. The analyst starts the analysis with looking what is interesting in the data, where “interestingness” is understood as relevance to the purpose of investigation. When something interesting is detected, new, more specific questions appear, which motivate the analyst to look for details. These questions affect what details will be viewed and in what ways. Hence, questions play an important role in EDA and can determine the choice of analysis methods. There are a few distinctions in comparison with the example questions given in textbooks on statistics and GIS:

- EDA essentially involves many different questions;
- the questions vary in their level of generality;
- most of the questions arise in the course of analysis rather than being formulated in advance.

These peculiarities make it rather difficult to formulate any guidelines for successful data exploration, any instructions concerning what methods to use in what situation. Still, we want to try.

There is an implication of the multitude and diversity of questions involved in exploratory data analysis: this kind of analysis requires multiple

tools and techniques to be used in combination, since no single tool can provide answers to all the questions. Ideally, a software system intended to support EDA must contain a set of tools that could help an analyst to answer any possible question (of course, only if the necessary information is available in the data). This ideal will, probably, never be achieved, but a designer conceiving a system or tool kit for data analysis needs to anticipate the potential questions and at least make a rational choice concerning which of them to support.

## 1.2 Objectives of the Book

This is a book about exploratory data analysis and, in particular, exploratory analysis of spatial and temporal data. The originator of EDA, John Tukey, begins his seminal book with comparing exploratory data analysis to detective work, and dwells further upon this analogy: “A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places. Equally, the analyst of data needs both tools and understanding” (Tukey 1977, p. 1).

Like Tukey, we also want to talk about *tools* and *understanding*. We want to consider current computer-based tools suitable for exploratory analysis of spatial and spatio-temporal data. By “tools”, we do not mean primarily ready-to-use software executables; we also mean approaches, techniques, and methods that have, for example, been demonstrated on pilot prototypes but have not yet come to real implementation.

Unlike Tukey, we have not set ourselves the goal of describing each tool in detail and explaining how to use it. Instead, we aim to systemise the tools (which are quite numerous) into a sort of catalogue and thereby lead readers to an understanding of the principles of choosing appropriate tools. The ultimate goal is that an analyst can easily determine what tools would be useful in any particular case of data exploration.

The most important factors for tool selection are the data to be analysed and the question(s) to be answered by means of analysis. Hence, these two factors must form part of the basis of our systemisation, in spite of the fact that every dataset is different and the number of possible questions is infinite. To cope with this multitude, it is necessary to think about data and questions in a general, domain-independent manner. First, we need to determine what general characteristics of data are essential to the problem of choosing the right exploratory tools. We want not only to be domain-

independent but also to put aside any specifics of data collection, organisation, storage, and representation formats. Second, we need to abstract a reasonable number of general question types, or data analysis tasks, from the myriad particular questions. While any particular question is formulated in terms of a specific domain that the data under analysis are relevant to, a general task is defined in terms of structural components of the data and relations between them.

Accordingly, we start by developing a general view of data structure and characteristics and then, on this basis, build a general task typology. After that, we try to extend the generality attained to the consideration of existing methods and techniques for exploratory data analysis. We abstract from the particular tools and functions available in current software packages to types of tools and general approaches. The general tool typology uses the major concepts of the data framework and of the task typology. Throughout all this general discussion, we give many concrete examples, which should help in understanding the abstract concepts.

Although each subsequent element in the chain “data–tasks–tools” refers to the major concepts of the previous element(s), this sort of linkage does not provide explicit guidelines for choosing tools and approaches in the course of data exploration. Therefore, we complete the chain by revealing the general principles of exploratory data analysis, which include recommendations for choosing tools and methods but extend beyond this by suggesting a kit of generic procedures for data exploration and by encouraging a certain amount of discipline in dealing with data.

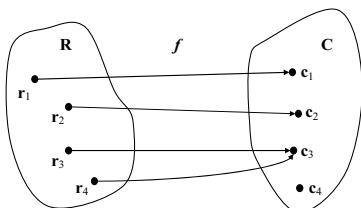
In this way, we hope to accomplish our goal: to enumerate the *tools* and to give *understanding* of how to choose and use them. In parallel, we hope to give some useful guidelines for tool designers. We expect that the general typology of data and tasks will help them to anticipate the typical questions that may arise in data exploration. In the catalogue of techniques, designers may find good solutions that could be reused. If this is not the case (we expect that our cataloguing work will expose some gaps in the data–task space which are not covered by the existing tools), the general principles and approaches should be helpful in designing new tools.

## 1.3 Outline of the Book

### 1.3.1 Data

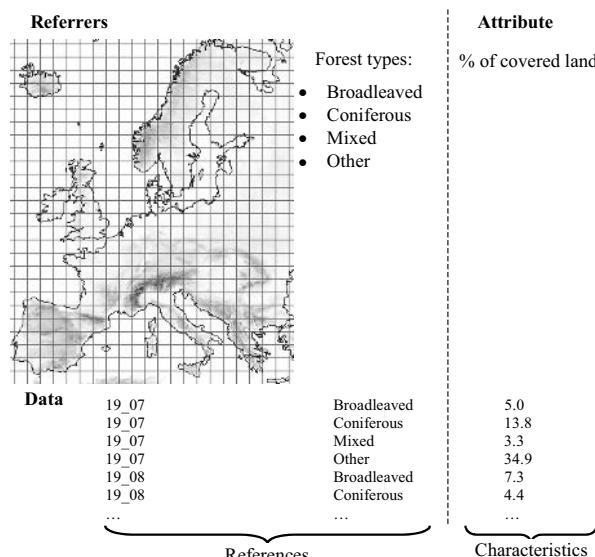
As we said earlier, we begin with introducing a general view of the structure and properties of the data; this is done in the next chapter, entitled

“Data”. The most essential point is to distinguish between characteristic and referential components of data: the former reflect observations or measurements while the latter specify the context in which the observations or measurements were made, for example place and/or time. It is proposed that we view a dataset as a function (in a mathematical sense) establishing linkages between references (i.e. particular indications of place, time, etc.) and characteristics (i.e. particular measured or observed values). The function may be represented symbolically as follows (Fig. 1.1):



**Fig. 1.1.** The functional view of a dataset

The major theoretical concepts are illustrated by examples of seven specific datasets. Pictures such as the following one (Fig. 1.2) represent visually the structural components of the data:

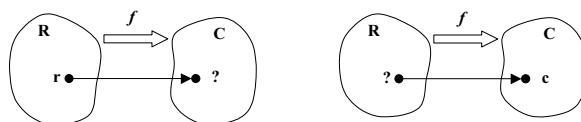


**Fig. 1.2.** A visual representation of the structure of a dataset

Those readers who tend to be bored by abstract discussions or cannot invest much time in reading may skip the theoretical part and proceed from the abstract material immediately to the examples, which, we hope, will reflect the essence of the data framework. These examples are frequently referred to throughout the book, especially those relating to the Portuguese census and the US crime statistics. If unfamiliar terms occur in the descriptions of the examples, they may be looked up in the list of major definitions in Appendix I.

### 1.3.2 Tasks

Chapter 3 is intended to propound a comprehensive typology of the possible data analysis tasks, that is, questions that need to be answered by means of data analysis. Tasks are defined in terms of data components. Thus, Fig. 1.3 represents schematically the tasks “What are the characteristics corresponding to the given reference?” and “What is the reference corresponding to the given characteristics?”

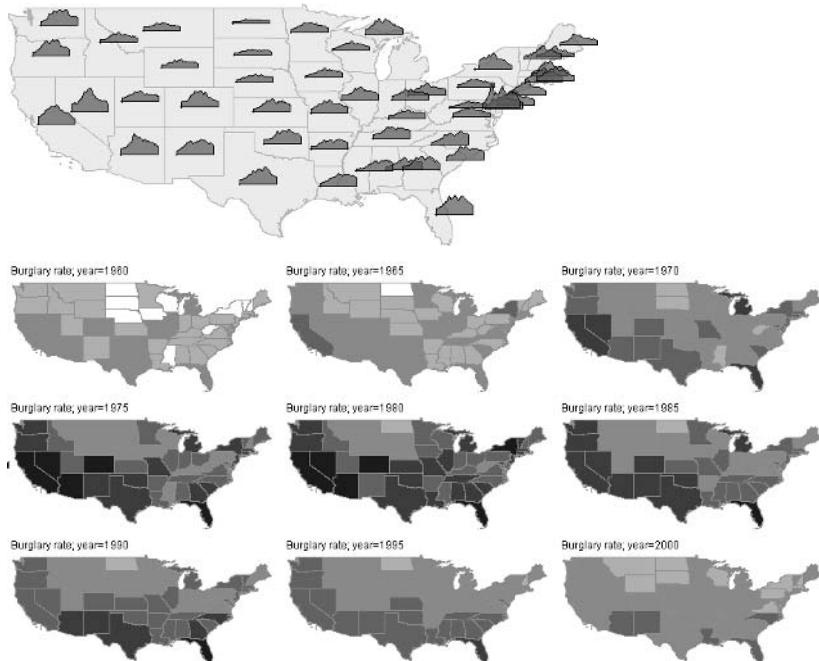


**Fig. 1.3.** Two types of tasks are represented schematically on the basis of the functional view of data

An essential point is the distinction between elementary and synoptic tasks. “Elementary” does not mean “simple”, although elementary tasks are usually simpler than synoptic ones. Elementary tasks deal with *elements* of data, i.e. individual references and characteristics. Synoptic tasks deal with sets of references and the corresponding configurations of characteristics, both being considered as unified wholes. We introduce the terms “behaviour” and “pattern”. “Behaviour” denotes a particular, objectively existing configuration of characteristics, and “pattern” denotes the way in which we see and interpret a behaviour and present it to other people. For example, we can qualify the behaviour of the midday air temperature during the first week of April as an increasing trend. Here, “increasing trend” is the pattern resulting from our perception of the behaviour.

The major goal of exploratory data analysis may be viewed generally as building an appropriate pattern from the overall behaviour defined by the entire dataset, for example, “What is the behaviour of forest structures in the territory of Europe?” or “What is the behaviour of the climate of Germany during the period from 1991 to 2003?”

We consider the complexities that arise in exploring multidimensional data, i.e. data with two or more referential components, for example space and time. Thus, in the following two images (Fig. 1.4), the same space- and time-referenced data are viewed as a spatial arrangement of local behaviours over time and as a temporal sequence of momentary behaviours over the territory:



**Fig. 1.4.** Two possible views of the same space- and time-referenced data

This demonstrates that the behaviour of multidimensional data may be viewed from different perspectives, and each perspective reveals some aspect of it, which may be called an “aspectual” behaviour. In principle, each aspectual behaviour needs to be analysed, but the number of such behaviours multiplies rapidly with increasing number of referential components: 6 behaviours in three-dimensional data, 24 in four-dimensional data, 120 in five-dimensional data, and so on.

We introduce and describe various types of elementary and synoptic tasks and give many examples. The description is rather extended, and we shall again make a recommendation for readers who wish to save time but still get the essence. At the end of the section dealing with elementary tasks, we summarise what has been said in a subsection named “Recap: Elementary Tasks”. Analogously, there is a summary of the discussion of

synoptic tasks, named “Recap: Synoptic Tasks”. Readers may proceed from the abstract of the chapter directly to the first recap and then to the second. The formal notation in the recaps may be ignored, since it encodes symbolically what has been said verbally. If unfamiliar terms are encountered, they may be looked up in Appendix I.

After the recaps, we recommend that one should read the introduction to connection discovery tasks (Sect. 3.5), which refer to relations between behaviours such as correlations, dependencies, and structural links between components of a complex behaviour. The section “Other approaches” is intended for those who are interested in knowing how our approach compares with others.

### 1.3.3 Tools

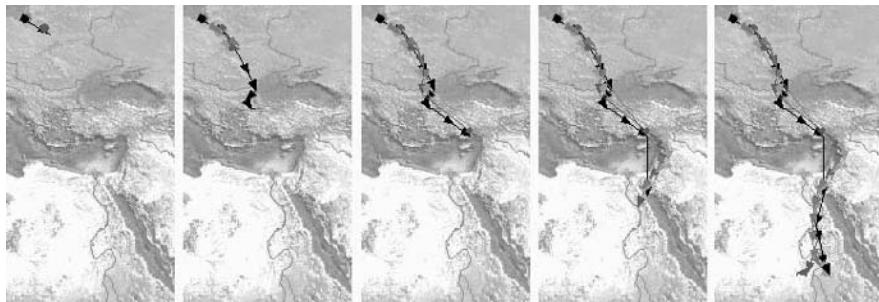
Chapter 4 systemises and describes the tools that may be used for exploratory data analysis. We divide the tools into five broad categories: visualisation, display manipulation, data manipulation, querying, and computation. We discuss the tools on a conceptual level, as “pure” ideas distilled from any specifics of the implementation, rather than describe any particular software systems or prototypes.

One of our major messages is that the main instrument of EDA is the brain of a human explorer, and that all other tools are subsidiary. Among these subsidiary tools, the most important role belongs to visualisation as providing the necessary material for the explorer’s observation and thinking. The outcomes of all other tools need to be visualised in order to be utilised by the explorer.

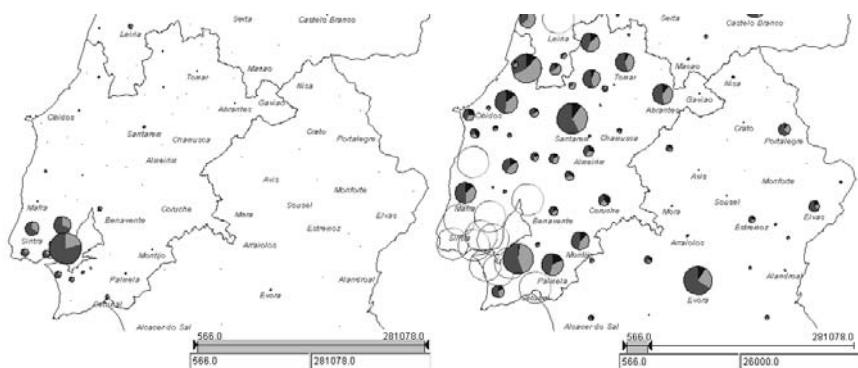
In considering visualisation tools, we formulate the general concepts and principles of data visualisation. Our treatment is based mostly upon the previous research and systemising work done in this area by other researchers, first of all Jacques Bertin. We begin with a very brief overview of that work. For those who still find this overview too long, we suggest that they skip it and go immediately to our synopsis of the basic principles of visualisation. If any unknown terms are encountered, readers may, as before, consult Appendix I.

After the overview of the general principles of visualisation, we consider several examples, such as the visualisation of the movement of white storks flying from Europe to Africa for a winter vacation (Fig. 1.5).

In the next section, we discuss display manipulation – various interactive operations that modify the encoding of data items in visual elements of a display and thereby change the appearance of the display. We are interested in such operations that can facilitate the analysis and help in



**Fig. 1.5.** A visualisation of the movement of white storks.



**Fig. 1.6.** An example of a display manipulation technique: focusing grasping general patterns or essential distinctions, rather than just “beautifying” the picture (Fig. 1.6).

Data manipulation basically means derivation of new data from existing data for more convenient or more comprehensive analysis. One of the classes of data manipulation, attribute transformation, involves deriving new attributes on the basis of existing attributes. For example, from values of a time-referenced numeric attribute, it is possible to compute absolute and relative amounts of change with respect to previous moments in time or selected moments (Fig. 1.7).

Besides new attributes, it is also possible to derive new references. We pay much attention to data aggregation, where multiple original references are substituted by groups considered as wholes. This approach allows an explorer to handle very large amounts of data. The techniques for data aggregation and for analysis on the basis of aggregation are quite numerous and diverse; here we give just a few example pictures (Fig. 1.8).

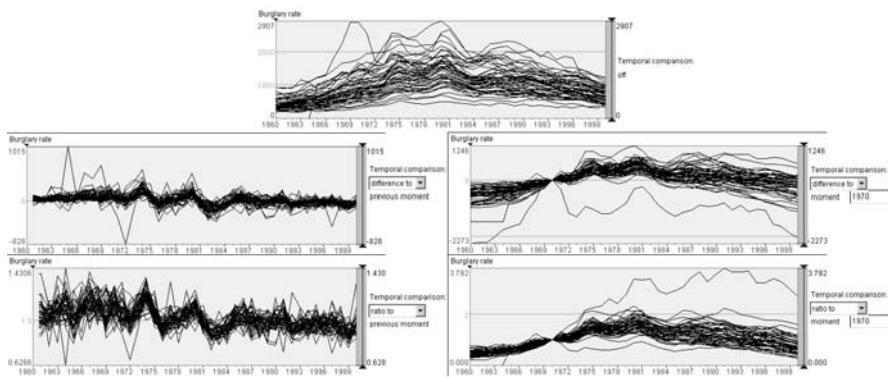


Fig. 1.7. Examples of various transformations of time-series data.

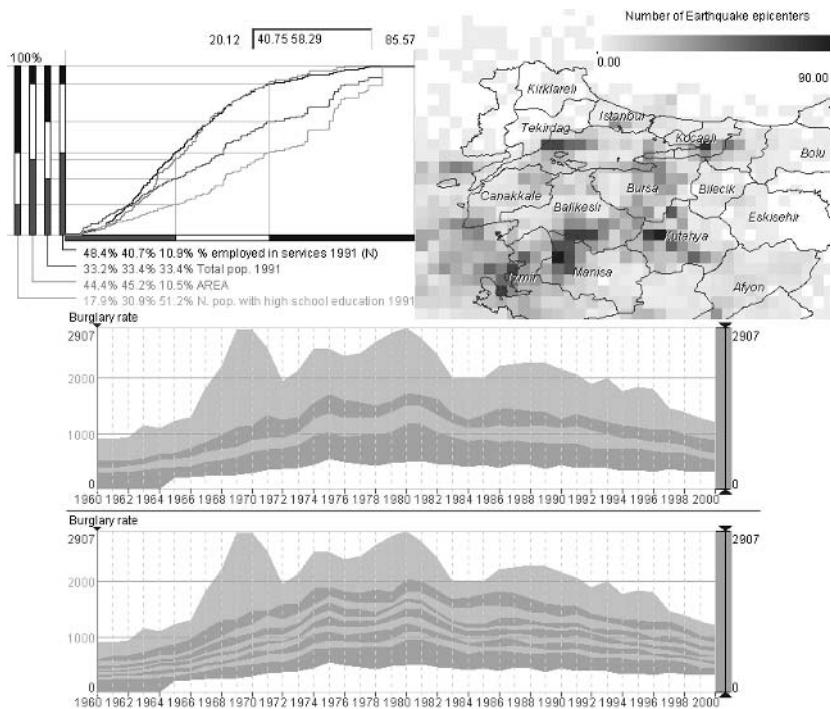
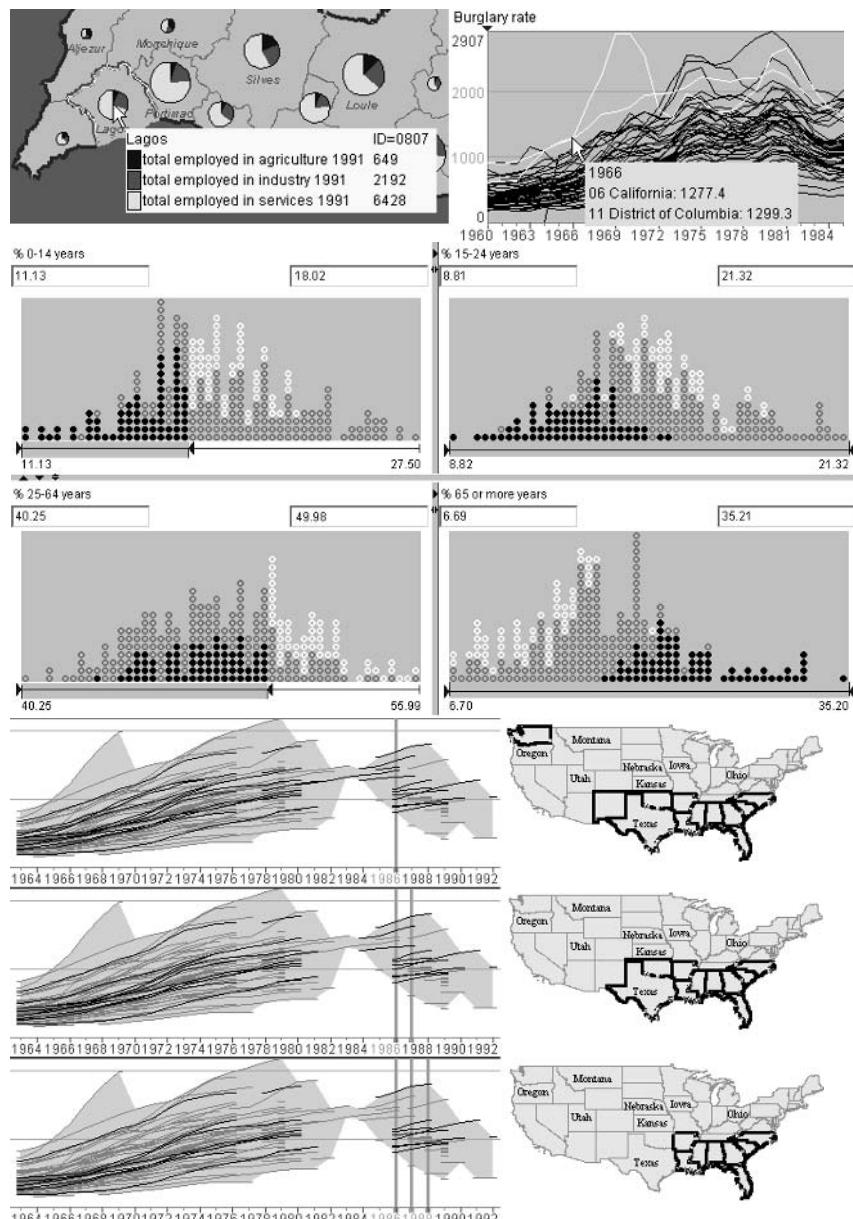


Fig. 1.8. A few examples of data aggregation

Querying tools are intended to answer various questions stated in a computer-understandable form. Among the existing querying tools, there are comprehensive ones capable of answering a wide variety of questions, which need to be formulated in special query languages. There are also dynamic querying tools that support quite a restricted range of questions

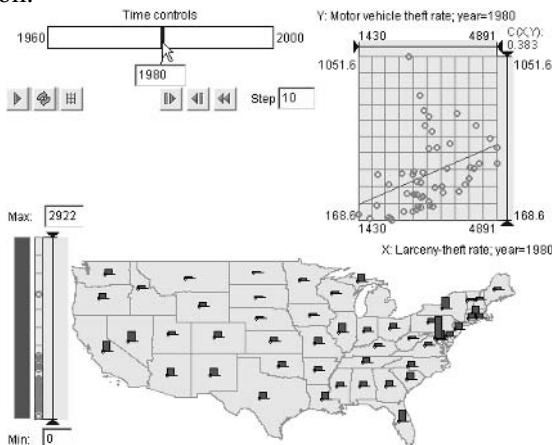
but provide a very simple and easy-to-use means for formulating questions (sometimes it is enough just to move or click the mouse) and provide a quick response to the user's actions. While both kinds of querying tools are useful, the latter kind is more exploratory by nature.



**Fig. 1.9.** Examples of dynamic querying tools

After considering querying, we briefly overview the computational techniques of data analysis, specifically, the most popular techniques from statistics and data mining. We emphasise that computational methods should always be combined with visualisation. In particular, the outcome of data mining may be hard to interpret without visualisation. Thus, in order to understand the meaning of the clusters resulting from cluster analysis, the characteristics of the members of the clusters need to be appropriately visualised.

The combining of various tools is the topic of the next section. We consider sequential tool combination, where outputs of one tool are used as inputs for other tools, and concurrent tool combination, where several tools simultaneously react in a consistent way to certain events such as querying or classification.



**Fig. 1.10.** Several tools working in combination

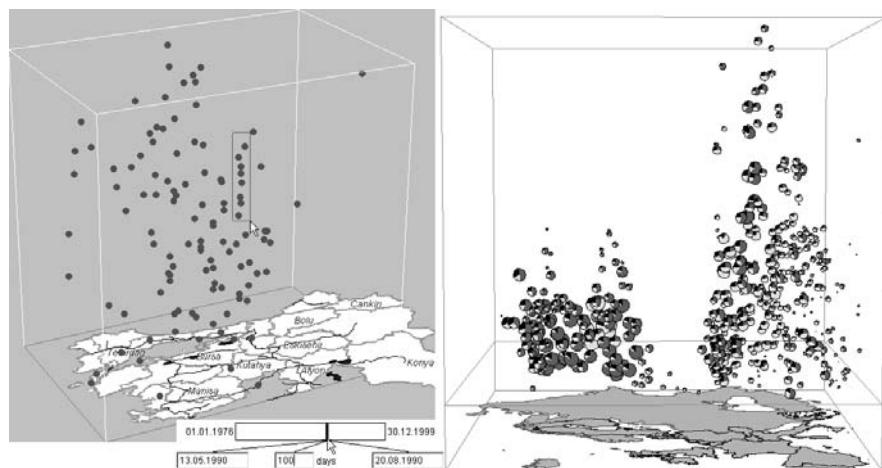
We hope that, owing to the numerous examples, this chapter about tools will not be too difficult or boring to read. The dependency between the sections is quite small, which allows readers who wish to save time to read only those sections which they are most interested in. In almost all sections, there are recaps summarising what was written concerning the respective tool category. Those who have no time or interest to read the detailed illustrated discussions may form an acquaintance with the material by reading only the recaps.

### 1.3.4 Principles

In Chap. 5, we subject our experience of designing and applying various tools for exploratory data analysis to introspection, and externalise it as a

number of general principles for data exploration and for selection of tools to be used for this purpose. The principles do not look original; most of them have been stated before by other researchers, perhaps in slightly different words. Thus, Shneiderman's mantra "Overview first, zoom and filter, and then details-on-demand" is close to our principles "see the whole", "zoom and focus", and "attend to particulars". The absence of originality does not disappoint us; on the contrary, we tend to interpret it as an indication of the general value of these principles.

The principles that we propound on the one hand explain how data exploration should be done (in our opinion), and on the other hand describe what tools could be suitable for supporting this manner of data exploration. Our intention has been to show data explorers and tool designers what they should care about in the course of data analysis and tool creation, respectively. Again, we give many examples of how our principles may be put into the practice of EDA. We refer to many illustrations from Chap. 4 and give many new ones.



**Fig. 1.11.** Illustration of some of the principles

Throughout the chapter, it can be clearly seen that the principles emphasise the primary role of visualisation in exploratory data analysis. It is quite obvious that only visualisation can allow an explorer to "see the whole", "see in relation", "look for what is recognisable", and "attend to particulars", but the other principles rely upon visualisation as well.

In the final sections we summarise the material of the book and establish explicit linkages between the principles, tools, and tasks in the form of a collection of generic procedures to be followed in the course of exploratory data analysis. We consider four cases, depending on the properties of

data under analysis: the basic case (a singlereferrer, a single attribute, and a manageable data volume), the case of multidimensional data (i.e. multiple referrers), the case of multiple attributes, and the case of a large data volume (i.e. great size of the reference set). We also give an example of the application of the procedures for choosing approaches and tools for the exploration of a specific dataset.

The above should give readers an idea of the content of this book; we hope that readers who find this content relevant to their interests will receive some value in return for the time that they will spend in reading the book.

## References

- (Burt and Barber 1996) Burt, J.E., Barber, G.M.: *Elementary Statistics for Geographers*, 2nd edn (Guilford, New York 1996)
- (Hand 1999) Hand, D.J.: Introduction. In: *Intelligent Data Analysis: an Introduction*, ed. by Berthold, M., Hand, D.J. (Springer, Berlin, Heidelberg 1999) pp.1–15
- (Klir 1985) Klir, G.J.: *Architecture of Systems Problem Solving* (Plenum, New York 1985)
- (Mitchell 1999) Mitchell, A.: *The ESRI® Guide to GIS Analysis. Vol.1: Geographic Patterns & Relationships* (Environmental Systems Research Institute, Redlands 1999)
- (NIST/SEMATECH 2005) NIST/SEMATECH e-Handbook of Statistical Methods. Chapter 1: Exploratory Data Analysis, <http://www.itl.nist.gov/div898/handbook/>. Accessed 29 Mar 2005
- (Shneiderman 1996) Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*, ed. by Burnett, M., Citrin, W. (IEEE Computer Society Press, Piscataway 1996) pp.336–343
- (STAT 2005) Wildman, P.: STAT 2005: An Internet course in statistics, <http://wind.cc.whecn.edu/~pwildman/statnew/information.htm>. Accessed 29 Mar 2005
- (Tukey 1977) Tukey, J.W.: *Exploratory Data Analysis* (Addison-Wesley, Reading MA, 1977)

## 2 Data

### Abstract

Data represent results of the observation or measurement of phenomena. By means of data analysis, people can study these phenomena. Data analysis can be regarded as seeking answers to various questions regarding the phenomena. These questions, or, in other words, data analysis tasks, are the focus of our attention. In this chapter, we attempt to develop a general view of data, which will help us to understand what data analysis tasks are potentially possible.

We distinguish two types of components of data, referrers and attributes, which can also be called independent and dependent variables. A dataset can be viewed on an abstract level as a correspondence between references, i.e. values of the referrers, and characteristics, i.e. values of the attributes. Here are a few examples:

- In a dataset containing daily prices of a stock on a stock market, the referrer is time and the attribute is the stock price. The moments of time (i.e. days) are references, and the price on each day is the characteristic corresponding to this reference.
- In a dataset containing census data of a country, the set of enumeration districts is the referrer, and various counts (e.g. the total population or the numbers of females and males in the population) are the attributes. Each district is a reference, and the corresponding counts are its characteristics.
- In a dataset containing marks received by schoolchildren in tests in various subjects (mathematics, physics, history, etc.), the set of pupils and the set of school subjects are the referrers, and the test result is the attribute. References in this case are pairs consisting of a pupil and a subject, and the respective mark is the characteristic of this reference.

As may be seen from the last example, a dataset may contain several referrers. The second example shows that a dataset may contain any number of attributes.

The examples demonstrate the three most important types of referrers:

- time (e.g. days);
- space (e.g. enumeration districts);
- population (e.g. pupils or school subjects).

The term “population” is used in an abstract sense to mean a group of any items, irrespective of their nature.

We introduce a general view of a dataset structure as a function (in the mathematical sense) defining the correspondence between the references and the characteristics.

## 2.1 Structure of Data

A set of data can be viewed as consisting of units with a common structure, i.e. it is composed of *components* having the same meaning in each of the units. We shall call such units data *records*. For example, data about total population numbers in municipalities of a country in each census year have three components in each record: the municipality, the year, and the population number. This abstract view of data is independent of any representation model.

Any item (record) of data includes two conceptually different parts: one part defines the context in which the data was obtained, while the other part represents results of measurements, observations, calculations etc. obtained in that context. The context may include the moment in time when the measurements were made, the location in space, the method of data acquisition, and the entity (or entities) the properties of which were measured (or observed, calculated, ...). Thus, in our example, the municipality and the year define the context in which the population number was measured.

We shall use the term *referential components* or *referrers* to denote data components that indicate the context, and the term *characteristic components* or *attributes* for components representing results of measurements or observations. It is convenient to assume that data components have names, such as “municipality”, “year”, or “population number”, although the names are not considered as a part of the data. The items that data records consist of, i.e. particular instances of referential and characteristic components, will be called the *values* of these components. Values of referrers will also be called *references*, and values of attributes will also be called *characteristics*.

The meaning of each component determines what items can potentially be its values and appear in the data. Thus, the values of the component “municipality” may be various existing municipalities, but not real num-

---

bers or tree species. The values of the component “year” may be various years designated by positive integer numbers, but not fractions. We shall call the set of all items that can potentially be values of some data component (but need not necessarily appear in the data) the *value domain* of this component.

Dataset components are often called *variables* in the literature; we shall use this term interchangeably with the term “components”. This does not mean, however, that we assume values of components always to be numeric. We use the term “variable” in a more general sense than “a quantity that may assume any one of a set of values” (Merriam-Webster 1999). In this definition, we replace “a quantity” by “something” and do not specify what a “value” is. The latter may be an element of a set of arbitrary nature.

We have found useful the general ideas concerning data structures and properties presented in *Architecture of Systems Problem Solving* by George Klir (Klir 1985). Klir considers the situation of studying an object through observation of its properties. The properties can be represented as *attributes* taking on various *appearances*, or manifestations. “For instance, if the attribute is the relative humidity at a certain place on the Earth, the set of appearances consists of all possible values of relative humidity (defined in some specific way) in the range from 0% to 100%”. Klir’s “appearances” correspond to our “values”.

“In a single observation, the observed attribute takes on a particular appearance. To be able to determine possible changes in its appearance, multiple observations of the attribute must be made. This requires, however, that the individual observations of the same attribute, performed according to exactly the same observation procedure, must be distinguished from each other in some way. Let any underlying property *that is actually used* to distinguish different observations of the same attribute be called a *backdrop*. The choice of this term, which may seem peculiar, is motivated by the recognition that the distinguished property, whatever it is, is in fact some sort of background against which the attribute is observed.” (Klir 1985)

Klir’s notion of a “backdrop” corresponds to what we call a referential component or referrer. According to Klir, there are three basic kinds of backdrop: *time*, *space*, and *population*. By “population” Klir means a set of any items, not only people. Some examples of Klir’s population are a set of manufactured products of the same kind, the set of words in a particular poem or story, and a group of laboratory mice.

In general, references themselves do not contain information about a phenomenon but relate items of this information (characteristics) to different places, time moments, objects, etc. Thus, the census data mentioned earlier consisting of municipalities, years, and population numbers characterise the population of a country. However, only the data component “population number” is directly related to the population and expresses

some of its properties. The other two components do not provide any information about the phenomenon. Instead, they allow us to relate specific values of the population number to corresponding time moments (years) and fragments of territory (municipalities).

There is another difference between referential and characteristic components: references can often be chosen arbitrarily, while the corresponding values of the attributes are fully determined by the choice made. Thus, in our example, the selection of the years when the population was counted was made arbitrarily by the authorities and could, in principle, be changed. The same applies to the municipalities for which the data were collected: one could decide to aggregate the data about individual people and households by smaller or larger units of territory, or to change the boundaries. At the same time, each value of a population number present in the database is inseparably linked to a specific year and a specific area. The value is completely determined by the temporal and spatial references and cannot be set arbitrarily. Hence, referrers can be viewed as independent components of data, and attributes as depending on them.

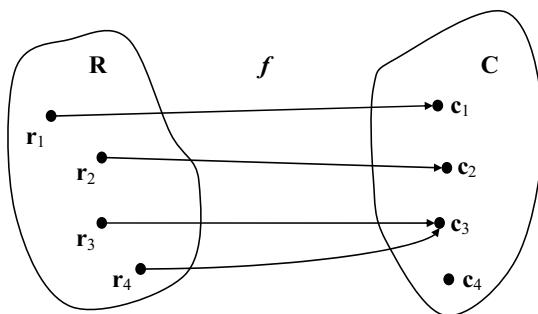
As we have mentioned, Klir distinguishes three possible types of referrers (backdrops): space, time, and population (groups of objects). However, it should not be concluded that space, time, and population are *always* used for referencing. Klir noted; “Time, space, and population, which have special significance as backdrops, may also be used as attributes. For instance, when sunrise and sunset times are observed each day at various places on the Earth, the attribute is time and its backdrops are time and space ...”. We can give more examples. In data about moving objects, such as migratory animals, the observed locations are dependent on the objects and the selected moments of observation. Hence, space is an attribute here. A set of political parties in data about the distribution of votes obtained by parties in an election is an example of a population-type referrer. However, in data showing which party won the election in each municipality, the party is an attribute.

Besides space, time, and population, other types of referrers may be encountered. Thus, the level of the water in a river is an attribute in data about daily measurements of the water level. The same attribute will be a referrer in data about the flooded area depending on the level of the water in the river. Hence, space, time, and population can be viewed as the most common types of referrers but not as the complete set of all possible types.

### 2.1.1 Functional View of Data Structure

The notion of a *function* in mathematics is a very convenient metaphor for reasoning about data. A function is a relation between two or more variables such that the values of one variable are dependent on, determined by, or correspond to values of the other variables, its arguments. In algebra and set theory, functions are often called “many-to-one” mappings. This means that, for each combination of values of the arguments, there is no more than one corresponding value of the dependent variable. In general, there is no presumption that the variables must be numeric; a function may be defined for sets of arbitrary nature.

We consider a dataset as a correspondence between referential and characteristic components (referrers and attributes) such that for each combination of values of the referential components there is no more than one combination of values of the attributes. Hence, a dataset is a function that has the referrers as arguments and has the dependent variable constructed from the attributes such that the value domain of this variable consists of all possible combinations of values of the attributes. This function will be called the *data function* in what follows.



**Fig. 2.1.** The functional view of the structure of a dataset illustrated graphically. Here,  $r_1$ ,  $r_2$ ,  $r_3$ , and  $r_4$  represent different references, i.e. combinations of values of the dataset referrers.  $R$  is the set of all references, including, among others, the references  $r_1$ ,  $r_2$ ,  $r_3$ , and  $r_4$ .  $c_1$ ,  $c_2$ ,  $c_3$ , and  $c_4$  represent different characteristics, i.e. combinations of values of attributes.  $C$  is the set of all possible characteristics, including, among others, the characteristics  $c_1$ ,  $c_2$ ,  $c_3$ , and  $c_4$ .  $f$  is the data function, which associates each reference with the corresponding characteristic

The functional view of a dataset is illustrated graphically in Fig. 2.1. The dataset structure is represented as a combination of three key components:

- **R**: The set of all references, i.e. combinations of values of the dataset referrers. This set will be called the *reference set*
- **C**: The set of all possible characteristics, i.e. combinations of values of the dataset attributes. This set will be called the *characteristic set*
- $f$ : The data function, i.e. the correspondence between each element of the reference set and a specific element of the characteristic set.

We have drawn this picture so as to demonstrate the following properties of the data function:

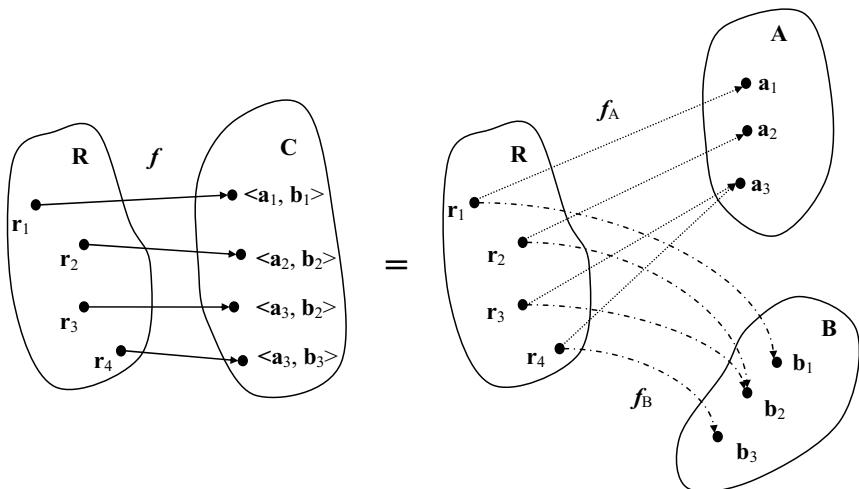
- *each* element of the reference set has a *single* corresponding element of the characteristic set;
- characteristics corresponding to different references may coincide;
- some combinations of attribute values may never occur in a dataset, i.e. there may be no references that they correspond to.

We assume that a corresponding characteristic exists for each reference present in a dataset. However, it often happens that some data in a dataset are missing. We treat such cases as incomplete information about the data function: the characteristics of some references may be unknown but they still exist, and hence can potentially be found.

In a dataset with multiple referential components, these components cannot be considered separately from each other, because only combinations of values of all of them produce complete references that uniquely determine the corresponding attribute values. In contrast, it is quite possible to consider each attribute separately from the other attributes. Hence, a dataset with  $N$  attributes may be considered both as a single function that assigns combinations of  $N$  attribute values to combinations of values of the referential components, and as  $N$  functions where each function assigns values of a single attribute to combinations of values of the referrers. These two views are equivalent. Fig. 2.2 illustrates this idea graphically by the example of two attributes.

Independent consideration of individual attributes very often takes place in the practice of data analysis since this is much easier and more convenient than dealing with multiple attributes simultaneously. However, there are cases where it is necessary to consider combinations of several attributes. For example, when the age structure of a population is studied, one may need to look at proportions of different age groups simultaneously.

The use of the notion of a function as a metaphor allows us to draw analogies between data analysis and the analysis of functions in mathematics. In particular, this will help us in defining the set of generic data analysis tasks. Nevertheless, we would like to limit the use of mathematical terms and ensure that our ideas will be understandable to people without a



**Fig. 2.2.** A dataset with  $N$  attributes may be treated in a two ways: as a single function associating the references with different combinations of values of these  $N$  attributes, or as  $N$  functions associating the references with individual values of these  $N$  attributes. This picture schematically represents a dataset with two attributes, denoted by **A** and **B**;  $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \dots$  and  $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \dots$  are different values of the attributes **A** and **B**, respectively. The characteristic set **C** consists of various pairs, each comprising one value of the attribute **A** and one value of the attribute **B**, for example  $\langle \mathbf{a}_1, \mathbf{b}_1 \rangle, \langle \mathbf{a}_1, \mathbf{b}_2 \rangle$ . The data function  $f$  associates each reference from the reference set **R** with one of these pairs. This function is equivalent to a combination of two functions, denoted by  $f_A$  and  $f_B$ . The function  $f_A$  associates each reference with a certain value of the attribute **A**, and  $f_B$  associates it with a certain value of the attribute **B**

solid mathematical background. Although we shall use a formal notation in the next chapter, we shall try to make it as simple as possible. For those who are more familiar with mathematical terminology and notation, we shall sometimes offer supplementary explanations with the use of more formal definitions and additional mathematical concepts. However, these explanations are not strongly required for an overall understanding. The symbols  $\triangleright$  and  $\triangleleft$  will be used to indicate the beginning and end of such material. Below, we present an algebraic reformulation of the functional view of data for those who want to be better prepared for the next chapter.

➤ A dataset is a mapping from a set of references onto a set of characteristics, i.e. a function

$$d: \mathbf{R} \rightarrow \mathbf{C}$$

where **R** is a set of references and **C** is a set of characteristics. By references we mean tuples (combinations) of values of referential variables, or

referrers, and by characteristics we mean tuples of values of characteristic variables, or attributes. Hence, in the general case, both sets **R** and **C** are Cartesian products of several sets, each consisting of values of one data component:

$$\begin{aligned}\mathbf{R} &= \mathbf{R}_1 \times \mathbf{R}_2 \times \dots \times \mathbf{R}_M \\ \mathbf{C} &= \mathbf{C}_1 \times \mathbf{C}_2 \times \dots \times \mathbf{C}_N\end{aligned}$$

where  $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_M$  are sets of values of the referrers, and  $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_N$  are sets of values of the attributes.

Let  $r$  be a specific element of the reference set **R**, i.e. a combination of particular elements  $r_1, r_2, \dots, r_M$  from the sets  $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_M$ , respectively. The corresponding element of the characteristic set **C** may be denoted as  $d(r)$ ; this is a combination of particular elements  $c_1, c_2, \dots, c_N$  from the sets  $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_N$ , respectively.

The mapping from the references onto the characteristics  $d: \mathbf{R} \rightarrow \mathbf{C}$  can also be represented in a slightly different way as a function of multiple variables:

$$d(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M) = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N)$$

where  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M$  are the referential (i.e. independent) variables, taking values from the sets  $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_M$ , respectively, and  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$  are the characteristic (dependent) variables taking values from  $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_N$ , respectively.

In our example of census data, the reference set **R** is a Cartesian product of the set of municipalities and the set of census years. The characteristic set **C** is a Cartesian product of the value sets of such attributes as the total population number, the numbers of females and males, and the number of children, pensioners, unemployed, and so on. This dataset can be formally represented by the expression

$$d(\mathbf{m}, \mathbf{y}) \rightarrow (\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3, \dots)$$

where the variable  $\mathbf{m}$  corresponds to municipalities,  $\mathbf{y}$  to years, and  $\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3, \dots$  to population numbers in the various population groups.

As we have already mentioned, any attribute may be considered independently of the other attributes, i.e. the function  $d(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M) \rightarrow (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N)$  having tuples as results may be decomposed into  $n$  functions, each involving one of the attributes:

$$\begin{aligned}d_1(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M) &\rightarrow \mathbf{v}_1 \\ d_2(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M) &\rightarrow \mathbf{v}_2 \\ &\dots \\ d_N(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M) &\rightarrow \mathbf{v}_N \leftarrow\end{aligned}$$

## 2.1.2 Other Approaches

Since we are particularly interested in spatial and spatio-temporal data (i.e. data having spatial and temporal components), it is appropriate to compare our view of data with that adopted in the area of spatial data handling. In cartography and geoinformatics, data about spatial phenomena are traditionally divided into spatial (geographic) and non-spatial information, the latter being also called “thematic” information or “attributes”. Here the term “attribute” is used in a different sense from what we have considered thus far: it denotes merely the non-spatial aspect of data. Recently, a need for a special consideration of time has been recognised. According to Nyerges (cited in MacEachren (1995)), any phenomenon is characterised by a “bundle of properties” that includes a *theme*, *space*, and *time*. The temporal aspect of a phenomenon includes the existence of various objects at different time moments, and changes in their properties (spatial and thematic) and relationships over time.

There is no explicit division of data into referential and characteristic components in the literature on cartography and GIS. According to our observation, space and time are typically implicitly treated as referrers and the remaining data components as referring to space and time, i.e. as attributes in our terms. The existence of the term “spatially referenced data” in the literature on GIS supports this observation. In principle, there is quite a good justification for this view. Although space and time can be characteristics as well as references, it is often possible, and useful for data analysis, to convert space and time from characteristics to references. For example, in data about occurrences of earthquakes, the locations of the earthquakes are attributes characterising the earthquakes. However, for a study of the variation of seismic characteristics over a territory, it may be appropriate to treat space as an independent container for events: each location is characterised by the presence or absence of earthquake occurrences, or by the number of occurrences. In fact, this is a transformation of the initial data (and a switch to consideration of a different phenomenon, namely the seismicity of an area rather than earthquakes themselves), although it might be done merely by indicating the locations of the earthquakes on a map. A map facilitates the perception of space as a container where some objects are placed, rather than as an attribute of these objects. Perhaps this is the reason why space is usually implicitly treated as a referer in cartography and geoinformatics.

A similar transformation may be applied to temporal attributes, such as the times of earthquake occurrences in our example. Thus, in order to produce an animated presentation of the data, a designer usually breaks the time span between the first and the last recorded earthquake occurrence

into regular intervals, e.g. days or months. At each moment of the animation, the earthquakes that occurred during one of these intervals are shown. Thereby, the time is turned from an attribute into an independent referential variable: time moments or intervals are selected arbitrarily, and the earthquake occurrences that are visible on the screen depend on the selection made. Like space, time is now treated as an independent container of events.

The possibility of treating space and time both as referrers and as attributes is reflected in the reasoning concerning the absolute and relative views of space and time (Peuquet 1994, 2002, Chrisman 1997). According to the absolute view, space and time exist independently of objects and form a framework, or a container, where objects are placed. According to the relative view, both space and time are properties attached to objects such as roads, rivers, and census tracts.

A real-life example of the dual treatment of spatial and temporal components of data can be found in Yuan and Albrecht (1995). By interviewing analysts of data about wildfires, the researchers revealed four different conceptual models of spatio-temporal data used by these people. Models are classified into location-centred and entity-centred models, according to the analyst's view of space. In the location-centred models, all information is conceptualised as attributes of spatial units delineated arbitrarily or empirically (i.e. space is a referrer). The entity-centred models represent reality by descriptions of individual entities that have, among other things, spatial properties (space is an attribute of the entities). Within these two classes, the models are further differentiated according to the view of time: either the data refer to arbitrarily defined temporal units in a universal time frame (time is a referrer) or time is described as an attribute of spatial units or entities.

For generality, we shall assume that any dataset with locations and/or time moments attached to some entities can be considered in a dual way:

- The set of entities forms a referrer of the type “population”. The locations and time moments are values of the spatial and temporal attributes characterising these entities, along with any other attributes.
- The referrers are space and time (or one of these). For specific values or combinations of values of the referrer(s), there are corresponding entities, i.e., the presence of an entity is regarded as an attribute characterising locations in space and/or moments in time. The values of all attributes characterising the entities are assumed to refer to the locations and/or time moments at which these entities exist. Hence, attribute values may not necessarily be defined for all values of the referrer(s).

Although we have a particular interest in spatio-temporal data, we still do not intend to restrict ourselves to only this kind of data, but will attempt to develop a more general framework. Therefore, we prefer to adopt the abstract view of data structure based on an explicit division of components into referrers and attributes, where both referrers and attributes may have any type, including spatial and temporal.

Our view of data structure is close to the concept of *data cubes*, which is commonly accepted as a method for abstracting and summarising relational databases and data warehouses (see, for example, Gray et al. (1998)). Data cubes categorise data components into two classes, namely dimensions and measures, corresponding to the independent and dependent variables, respectively (or, in our terms, to referrers and attributes). Data are abstractly structured as a multidimensional cube, where each axis corresponds to a dimension and consists of every possible value for that dimension. Every “cell” in the data cube corresponds to a unique combination of values for the dimensions and contains one value per measure.

## 2.2 Properties of Data

Both referential and characteristic components can be distinguished according to the mathematical properties of the underlying sets from which they take their values. According to Klir, the most important properties are the *ordering* of the set elements, the existence of *distances* between the elements, and *continuity*. Sets may be unordered, partially ordered, or fully (linearly) ordered. For example, a group of objects (a population) is a discrete set without ordering or distances. Time is a linearly ordered continuous set with distances. Space is also a continuous set with distances, but there is no natural order between its elements, i.e. locations. In space, it is possible to introduce various arbitrary orderings. Thus, a coordinate system established in space specifies a full ordering when the space has one dimension, and a partial ordering in the case of two or more dimensions. Any set of numbers is linearly ordered and has distances; the distance is the difference between the numbers concerned. A numeric set may be either discrete, as in measurements of a number of objects, or continuous, as in measurements of temperature.

We assume that the values of a component may be not only individual items but also sets (subsets of a certain basic set of individual items). For example, in a dataset describing countries, there may be attributes specifying in what international organisations each country participates, what nationalities live there, and what languages are spoken. In data about build-

ings, the locations of the buildings are not just points in space but areas, i.e. subsets of space. In data about events, such as conferences and fairs, the time frame of each event is an interval in time, i.e. again a set rather than a single time moment. A set consisting of sets is partly ordered: some of the sets may be parts of other, bigger sets.

Besides the properties of each data component taken separately, it is important to know the properties of the correspondence between the references and the characteristics, in particular, whether for each reference (i.e. combination of values of the referential components) there is (potentially) a corresponding value of each attribute. “Potentially” means that for some references the corresponding attribute values may be undetermined (not measured) even though it is known that such values exist. If all attributes characterising a phenomenon possess this property with respect to a referrer with a *continuous* value set, we call this phenomenon *continuous* with respect to this referrer. For example, a phenomenon may be continuous with respect to space (spatially continuous) and/or with respect to time (temporally continuous). It may be said that a continuous phenomenon exists everywhere over the set of values of the referrer. A phenomenon is *partly continuous* if it is continuous for one or more continuous subsets of the referrer’s value set but not for the whole set. A phenomenon is *discrete* if its characteristics exist only for a finite or countable subset of values of a continuous referrer. For example, air temperature is a spatially and temporally continuous phenomenon, clouds are partly continuous with respect to space and time, and lightning can be viewed as a spatially and temporally discrete phenomenon.

We shall also use terms such as “continuous attribute” or “discrete attribute” for denoting attributes characterising continuous or discrete phenomena.

Let us consider the case where a referrer and an attribute have value sets with distances. The attribute is called *smooth* with respect to this referrer if its values corresponding to close values of the referrer are also close, and if the smaller the distance between the referrer’s values, the closer the corresponding values of the attribute are. Air temperature is an example of a smooth phenomenon (attribute) with respect to both space and time. Room price in a hotel is smooth neither spatially nor temporally. First, one cannot expect that prices for adjacent rooms will always be closer than when the rooms are spatially distant; second, the prices may become significantly higher than usual on public holidays or during events that attract many visitors. Such attributes may be called *abrupt*.

It is possible to use our functional representation of data structure, which treats an attribute as a function, to give stricter definitions of continuous and smooth attributes, by analogy with continuous and smooth

functions in mathematics. However, we feel the informal definition to be sufficient for an understanding of the concepts.

Since a continuous set consists of an infinite and uncountable number of elements, it is impossible to determine the corresponding attribute values for each value of a referrer that has a continuous value set. Such referrers are usually handled by means of discretising, i.e. division of the value set into a finite or at least countable number of equivalence classes, i.e. value subsets in which the members are treated as being the same. These classes are then used as values of this referrer, and the values of the attributes are defined for these classes rather than for the “atomic” elements.

It is clear that any continuous reference set may be discretised in many different ways. For example, a two-dimensional geographical space (territory) may be divided into administrative units of various levels (countries, provinces or states, communes or counties, etc.) or into regular cells, with an arbitrary choice of the cell size and the origin of the grid. For time, there is a customary division down to seconds; however, it may be more meaningful to use coarser divisions such as hours, days, weeks, months, or even centuries.

It may also be of practical value to introduce equivalence classes for a referrer with a countable or finite set of values. There are various possible reasons for this. Individual elements of a set may be too numerous to allow one to consider each of them, access to data about individuals may be restricted, or an analyst may be interested in aggregate rather than individual characteristics. For example, a set of people (i.e. a population in a demographic sense) may be divided into groups by age, gender, occupation, etc. A set of biological organisms may be divided into biological populations, i.e. groups of organisms living in the same area.

When only attribute values corresponding to subsets of the reference set (rather than individual references) are available for analysis, it is important to know how these attribute values were obtained, since the applicability of some analysis methods may depend on this. The definition of values of attributes corresponding to subsets of references may be done by either aggregation or sampling. Aggregation means summarising or averaging values over the subsets, for example counting households in each district of a territorial division and finding their average income. Sampling means choosing a representative element in each subset and assuming the attribute values corresponding to this element to be valid for the whole subset. For example, one can measure air temperature each day at noon and treat this as a characteristic of the day as a whole. For a continuous referrer and a continuous, smooth attribute, it is also possible to make measurements for sample values of the referrer and then to derive the attribute value for any other reference by means of interpolation. This technique is often used

to characterise spatially continuous and smooth phenomena such as the variation of altitude or air temperature over a territory.

Since the granularity level for dividing a set of references into subsets may be chosen arbitrarily, it is possible to characterise one and the same phenomenon with different levels of detail. Moreover, if a method for deriving attribute values for larger subsets from values for smaller subsets or individual elements is defined, it is possible to vary the level of detail on which the phenomenon is considered. This corresponds to the notion of “drilling” in data analysis, which is defined as a technique for navigating through levels of data granularity or detail. The term “drill up” or “roll up” means increasing the level of abstraction, and “drill down” means descending to finer granularity and more detail.

Again, our view is close to the ideas related to the concept of the data cube mentioned earlier. In the literature explaining this concept (Gray et al. 1998, Stolte et al. 2002), it is stated that the dimensions of a data cube may have a hierarchical (or, more generally, a lattice) structure. For example, a dimension specifying locations may consist of several levels such as country, state, and county. A temporal dimension may involve days, weeks, months, quarters, and years. The main value of data cubes is the possibility to look at data on different levels of detail by means of ascending and descending along these hierarchies and thus varying the degree of data aggregation.

Some other important properties of data need to be mentioned. One of them is data *completeness*. An attribute is completely specified if some value of it can be put in correspondence with any combination built out of the values of the referrers that occur in the dataset. Incompletely specified data are said to have *missing values*. Cases of missing values must be properly handled in the course of data analysis; an analyst should be very cautious in generalising any observations to such cases.

In most real-world situations, it is practically impossible to make precise measurements. Therefore there is always some degree of *uncertainty* concerning the attribute values obtained. While minor uncertainties can often be ignored, there are many cases where data uncertainty must be taken into account.

Uncertain measurements or observations can be represented in data in different ways. Thus, for a numeric attribute, a value range rather than a single value may be specified for a combination of values of the referrers. Another approach is to specify the likelihood (probability) that a particular value is attained. While it is clear that uncertain data require special analysis methods, the current state of the art in exploratory data analysis has not yet responded adequately to the challenges related to data uncertainty.

### 2.2.1 Other Approaches

In cartography and geoinformatics, it is usual to consider spatial, temporal, and thematic aspects of data (phenomena) separately. Phenomena are classified into points, lines, areas, and volumes according to their spatial properties (Slocum 1999). Another typology of spatial phenomena is based on two orthogonal dimensions: *spatial continuity* and *spatial (in)dependence* (MacEachren 1995, Slocum 1999). Phenomena are characterised as *discrete* or *continuous* according to the first of these dimensions. Discrete phenomena occur at isolated locations, while continuous phenomena occur everywhere. Phenomena may be classified as *smooth* (adjacent locations are not independent) or *abrupt* (adjacent locations are independent) according to the second dimension. Smooth phenomena change in a gradual fashion, while abrupt phenomena change suddenly. “For instance, rainfall and sales tax rates for states are both continuous in nature, but the former is smooth, while the latter is abrupt (varying at state boundaries)” (Slocum 1999, p. 18).

From our viewpoint, the notions of spatial continuity and spatial (in)dependence are based on an implicit treatment of space as a referrer. They can easily be extended to other types of referrers with continuous value sets. For example, it is possible to think about temporal continuity and temporal (in)dependence of a phenomenon.

In the GIS literature, data properties are considered mostly from the perspective of how the data are represented in a GIS. There are two basic approaches to the representation of data about spatial phenomena: object-based and location-based (Peuquet 1994, Chrisman 1997). The object-based approach arranges all information, both spatial (coordinates) and non-spatial, as attributes of geographic objects, or features. This corresponds to the *vector model* of data representation. In the GIS area, it is conventional to distinguish between point, line, and area (or polygon) features. Features are organised into themes, or layers. In addition to the point, line, and area types, some authors consider networks consisting of connected lines as a special type of spatial data (see, for example, Verbyla (2002)). In the location-based approach, all information is stored relative to specific locations. Each location is associated with a list of objects occurring in it, and values of thematic attributes. This approach corresponds to the *raster model*, which divides a territory into multiple building blocks of the same size and shape called “grid cells” or “pixels”, which are filled with measured attribute values.

It seems apparent that the vector data model can represent discrete phenomena better, while the raster model is more suitable for continuous phenomena. In reality, however, there is no strict correspondence between the

spatial continuity of a phenomenon and the representational model used. Thus, data about a continuous phenomenon may refer to sample locations represented as point features (e.g. air temperature measurements at different weather stations) or to districts of a territorial division represented as polygons (e.g. population density by municipalities). Peuquet argues that the location-based and object-based approaches are complementary rather than contradictory: some tasks can be done better on the basis of the former approach, whereas for other tasks the latter approach is more suitable. Hence, it is beneficial to combine these two approaches (Peuquet 1994). Actually, transformation from vector to raster and from raster to vector representations is often used in data analysis.

According to the cartographic literature, the temporal aspect of a phenomenon is related to the existence of the phenomenon at different time moments and to the variation of its spatial and thematic properties over time. Spatio-temporal phenomena undergo several different types of changes (Blok 2000): existential changes (appearing and disappearing), changes in spatial properties (location, shape, size, orientation, altitude, height, gradient, and volume), and changes in thematic properties, which include qualitative changes and changes in ordinal or numeric characteristics (increases and decreases). Sometimes only one type of change takes place or is of interest to an analyst, but in many cases one needs to consider two or three types simultaneously.

Time itself can be treated in two different ways: as a *linear* continuum and as a repeating *cycle* linked to the earth's daily rotation or annual revolution (MacEachren 1995). Data can be recorded, retrieved, and processed at different *temporal precisions*, such as seconds, hours, days, or centuries.

Analogously to the spatial aspect of data, the view of the temporal aspect in the GIS literature is also representation-oriented. The two basic approaches to data representation, object-based and location-based, are extended to include time. Peuquet writes (Peuquet 2002, p. 270):

- In the “discrete” (or entity-based) view, distinct entities, such as a lake, a road, or a parcel of land, are the basis of the representation. Their spatial and temporal extents are denoted as attributes attached to these entities.
- In the “continuous” (or field-based) view, the basis of the representation is space and/or time. Individual objects are denoted as attributes attached to a given location in space–time.

We shall not go into further detail concerning the representation of time in GIS. A comprehensive study of this topic may be found, for example, in Langran (1992). We prefer to adhere to a more abstract view of data, independent as much as possible of any representational paradigm.

Attributes (i.e. thematic components of spatially referenced data) are typically distinguished according to the *levels of measurement* introduced by Stevens (1946): *nominal*, *ordinal*, *interval*, and *ratio*. Often the categories “interval” and “ratio” are united into a single category “quantitative”, or “numeric” (Bertin 1967/1983, MacEachren 1995). The traditional notion of levels of measurement can be expressed in terms of the mathematical properties of sets. The nominal level corresponds to a set without ordering or distances, the ordinal level – to a set with linear ordering but without distances, and the interval and ratio levels – to linearly ordered sets with distances. The difference between the latter two levels is that a “true zero” element exists in a set characterised by the ratio level of measurement, i.e. an element that is always the first in the order. For example, in the set of possible population numbers, 0 is the “true zero” because there cannot be a population number less than 0.

Some researchers introduce more distinctions for characterising thematic data. Thus, Roth and Mattis (1990) classify numeric attributes into *coordinates* and *amounts* in order to capture the difference between, for example, two o’clock and two hours. In our opinion, this corresponds to the distinction between the interval and ratio levels of measurement. For amounts, such as the number of hours, there is a “true zero”. As an implication, it is possible to measure how much one amount is greater or smaller than another amount, i.e. to compute ratios between amounts. This is impossible for “coordinates”, i.e. sets corresponding to the interval level of measurement, where a “zero” element can be chosen arbitrarily, such as the starting moment of time for counting the time of day.

Jung (1995) suggests a yet more detailed classification of numeric variables:

- *Amounts*: Absolute quantities
- *Measurements*: Absolute numbers representing results of measurements (e.g. distance). Along with these measurements, the corresponding units must be specified
- *Aggregated values*: Amounts or measurements summarised by areas. Such variables are always implicitly dependent on the area
- *Proportional values*, normalised by division by a fixed value
- *Densities*: Amounts or aggregated amounts divided by corresponding areas. As a result, densities do not depend on the area
- *Coordinates* that specify position in some coordinate system, e.g. on the time axis.

In relation to our view of data, some of these categories (specifically, amounts and coordinates) may be reformulated in terms of mathematical

properties of sets, as was demonstrated above. Some other categories, such as aggregated values and densities, are related to various methods of data transformation. While Jung's categories are based on a special treatment of space as a referrer, we assume that aggregated values can be derived on the basis of any referrer, and densities on the basis of any continuous referrer.

Jung considers also a number of characteristics related to *data quality*, such as reliability and exactness. While we regard data quality as an important issue, it is not the focus of our study.

In general, we aim to adhere in our study, whenever possible, to the abstract view of the structure and properties of data, without separating them into spatial, temporal, and thematic aspects. However, we recognise that space and time have their specific properties and that it is necessary to take these specifics into account in data analysis. Therefore, whenever appropriate, we shall pay special attention to data having spatial and/or temporal components, either referrers or attributes.

## 2.3 Examples of Data

Now that we have introduced a general framework for the consideration of the structure and properties of data, we shall demonstrate the use of this framework by describing a few example datasets. These datasets will be referred to throughout the book and heavily used for the purposes of illustration. Therefore, we recommend readers to familiarise themselves with these datasets by reading the following brief descriptions.

### 2.3.1 Portuguese Census

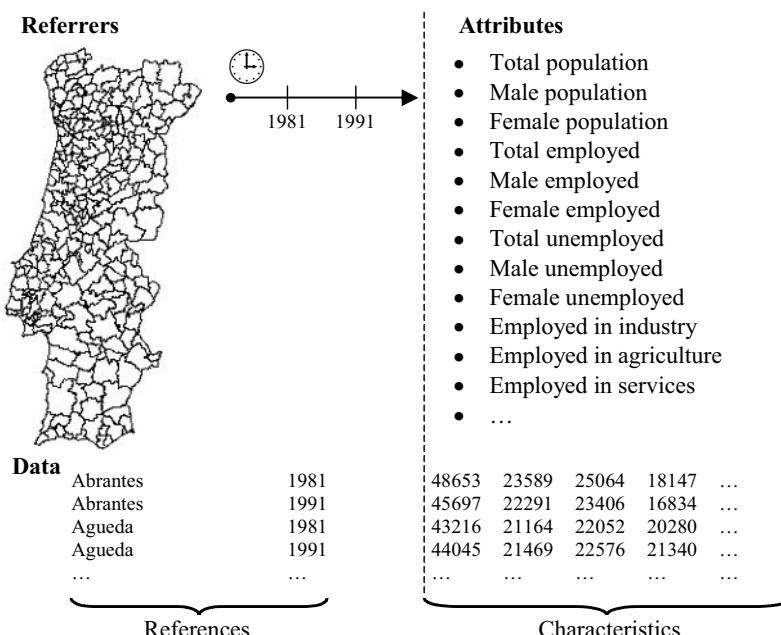
There are demographic data for the territory of Portugal, divided into 275 administrative districts, obtained from censuses in the years 1981 and 1991. The data include such attributes as population, numbers of males and females in the population, numbers of people employed in various occupations (agriculture, industry, and services), and numbers of unemployed.

In this dataset, space and time are referrers. Space has been discretised by dividing the territory of Portugal into administrative districts. The values of the attributes have been defined for these districts by summarising (aggregation) of data concerning individuals over the districts. The attributes are spatially continuous, i.e. defined everywhere on the territory of Portugal. The attributes are spatially abrupt rather than smooth because the numbers and characteristics of the population may change abruptly from

place to place (as a result of the aggregation that took place in the data preparation).

In relation to the temporal referrer, the corresponding attribute values are defined only for two sample time moments, specifically the years 1981 and 1991. The attributes are temporally continuous, since values exist at any time moment (but may not always be known). The attributes are also temporally smooth, because the population numbers and characteristics of the population usually change gradually over time.

All attributes available in the dataset have the ratio level of measurement, i.e. the value set of each of them has a “true zero”. The structure of the dataset is illustrated in Fig. 2.3.



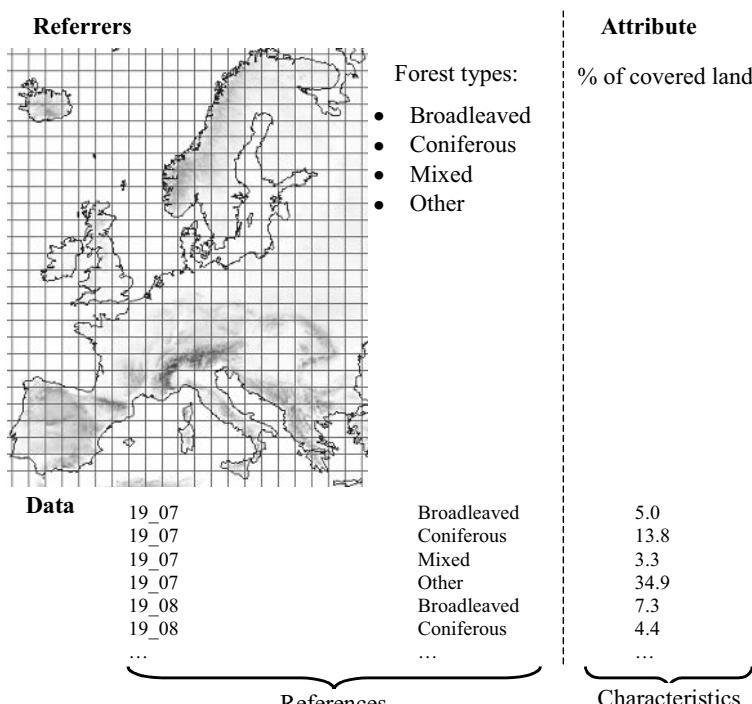
**Fig. 2.3.** The structure of the Portuguese census dataset

To make the illustrations for the book, we have often used not the original attributes expressing the absolute numbers of people in the various population groups, but derived attributes expressing the proportions of these groups in the populations of the districts. To produce these attributes, we have divided the absolute numbers by the total populations in the respective districts.

### 2.3.2 Forests in Europe

Here data about the distribution of different types of forest (coniferous, broadleaved, mixed, and other forests) over the territory of Europe are specified by numeric values referring to cells of a regular rectangular grid. The numbers represent the percentage of the area of each cell covered by the corresponding type of forest.

This dataset contains two referrers. One of the referrers is space, i.e. all locations in the territory of Europe. The spatial referrer has been discretised by dividing the territory into uniform compartments. The other referrer is the set of possible forest types consisting of four elements. This is a referrer of type “population”. There is a single attribute, the percentage of the land covered, defined for combinations of values of these two referrers. The attribute values for the compartments are defined by means of aggregation. The attribute has the ratio level of measurement.



**Fig. 2.4.** The structure of the European forest dataset

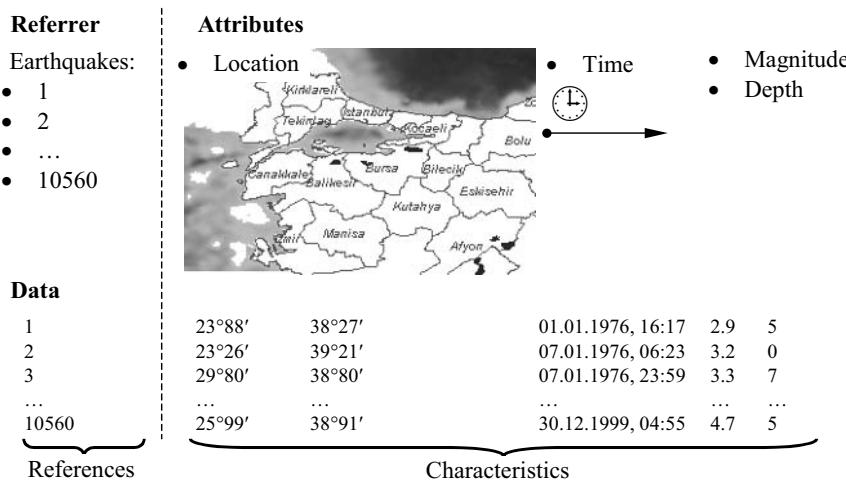
The phenomenon characterised by these data is partly continuous with respect to space: it does not exist, for example, in the areas covered by water. The phenomenon is spatially abrupt: the characteristics may change

greatly at the boundary of an urban area, for example. The structure of the dataset is illustrated in Fig. 2.4.

### 2.3.3 Earthquakes in Turkey

This is a dataset containing data about 10 560 earthquakes that occurred in the area around the Sea of Marmara in western Turkey and adjacent territories from 1 January 1976 to 30 December 1999. For each earthquake, there are the date and time of day when it occurred, and its location (longitude and latitude), magnitude, and depth. The structure of the dataset is illustrated in Fig. 2.5.

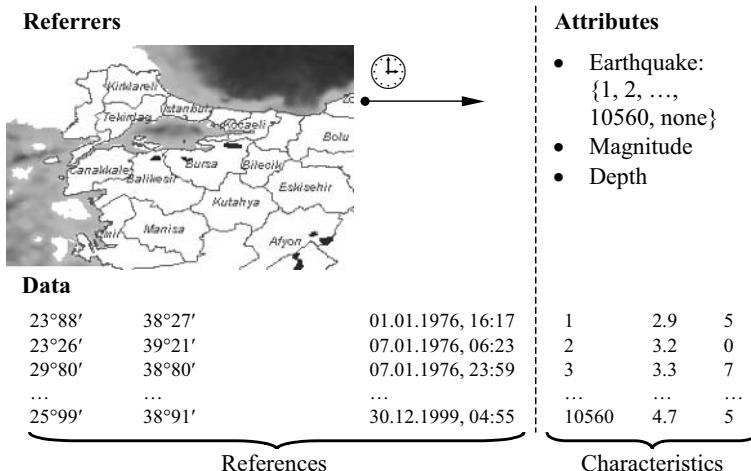
This dataset may be treated in a dual way. On the one hand, the set of earthquakes is a referrer of the type “population” with a finite number of possible values. Time (i.e. the time of occurrence of the earthquake) and space (i.e. the earthquake location), as well as the magnitude and depth, are attributes with values referring to individual earthquakes. The set of earthquakes is a discrete referrer without ordering and without distances between the elements.



**Fig. 2.5.** The structure of the dataset of earthquakes in the Marmara region (western Turkey)

On the other hand, it is possible to treat space and time as referrers and the earthquakes as a phenomenon existing in space and time. In this case, both space and time have continuous value sets with distances between elements. The earthquakes are discrete phenomena, both spatially and temporally: they exist only at specific locations and specific time mo-

ments. Hence, the attributes characterising the earthquakes (i.e. magnitude and depth) are also discrete, although each of them has a continuous value set. The structure of the dataset is illustrated in Fig. 2.6.



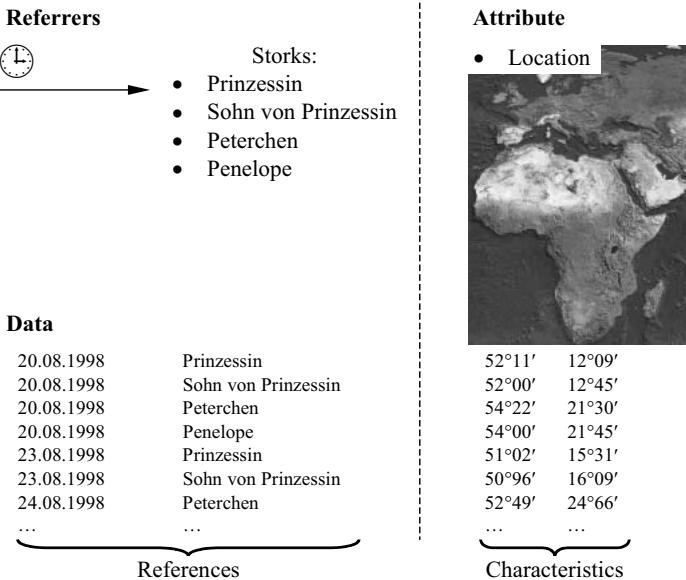
**Fig. 2.6.** Another view of the earthquake dataset. Here, space and time are treated as referrers rather than as attributes of the earthquakes

While the depth is an attribute with the ratio level of measurement, our knowledge of seismology is insufficient to make a definite judgement concerning the magnitude.

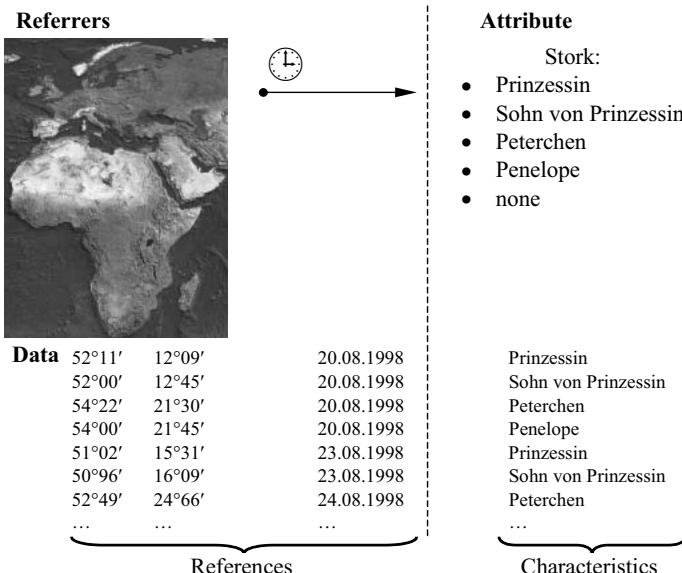
### 2.3.4 Migration of White Storks

An observation of migratory movements of four white storks was made during the period from 20 August 1998 to 1 May 1999. The data collected contain locations of the birds at various dates. In this dataset, there are two referrers, time and the storks observed; the latter is a population-type referrer with four possible values. The dataset has one attribute, location in space. The values of the attribute are defined for moments of time. With respect to time, the attribute is continuous and smooth: at any moment, each stork has a certain position in space, and this position changes gradually over time. The structure of the dataset is illustrated in Fig. 2.7.

As in the previous example, this dataset may also be viewed in another way: space and time may be treated as referrers, and the attribute reflects which stork is present in a given location at a given time moment (Fig. 2.8). In this view, the attribute is discrete.



**Fig. 2.7.** The structure of the dataset of white stork migration



**Fig. 2.8.** Another view of the white stork migration dataset. Here, space and time are treated as referrers

### 2.3.5 Weather in Germany

This dataset contains monthly climate data measured at 43 locations in Germany (weather stations) during the period from January 1991 to May 2003. The data include air temperature, rainfall, wind speed, etc.

In this dataset, there are two referrers: space and time. The phenomenon (i.e. climate) is spatially and temporally continuous, i.e. defined everywhere in space and time. However, the correspondence between the characteristics of the climate and space (more specifically, the territory of Germany) is defined by means of sampling, i.e. specifying attribute values for sample locations. The temporal referrer is discretised at the level of months. The corresponding attribute values are defined by means of aggregation; more specifically, averaging over monthly time intervals. The structure of the dataset is illustrated in Fig. 2.9.

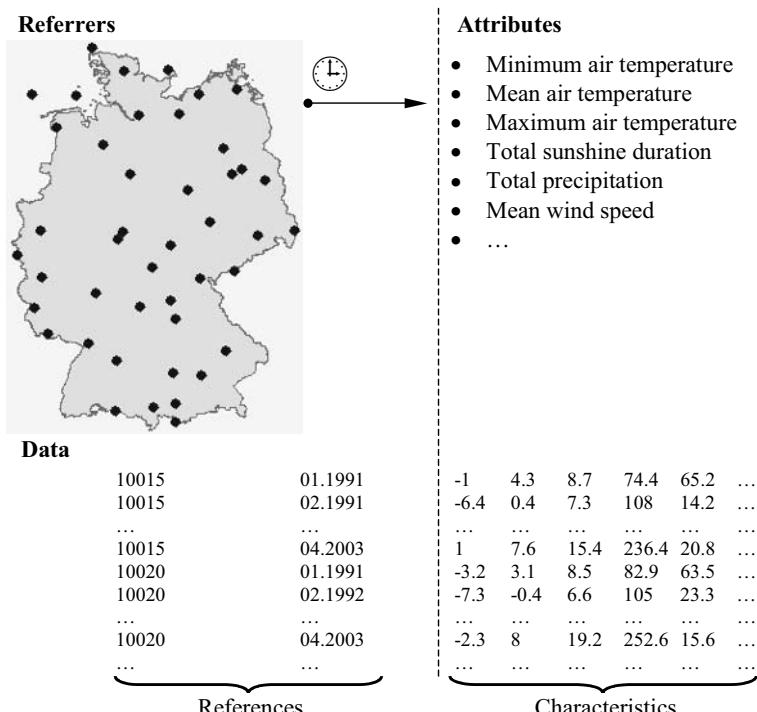


Fig. 2.9. The structure of the dataset related to the climate of Germany

The phenomenon of climate may be considered as partly smooth with respect to space. In many cases, it is possible to derive climate characteristics for locations between weather stations by interpolating the values

measured at the locations of the weather stations. However, it should be taken into account that close locations separated, for example, by a chain of mountains may have quite different climatic conditions. The climate may also be greatly influenced by other factors, for example by the closeness of a warm or cold oceanic current. Therefore, the extension of climate measurements taken at sample locations to other locations may be done correctly only on the basis of expert knowledge and additional data.

Climate is a smooth phenomenon with respect to time: although people often complain about sudden changes in the weather, the changes actually do not occur instantaneously. Thus, it usually takes some time for the temperature to increase or decrease. If the temperature was  $15^\circ$  at the moment  $t_1$  and  $20^\circ$  at the moment  $t_2$ , we may expect that there was a moment  $t$  between  $t_1$  and  $t_2$  when the temperature was  $18^\circ$ , even though the interval from  $t_1$  to  $t_2$  may sometimes be quite short. In our particular case, however, we should take into account the fact that the dataset we have at our disposal contains only aggregated (with respect to time) attribute values. These values cannot be used for interpolation, i.e. deriving values for intermediate time moments.

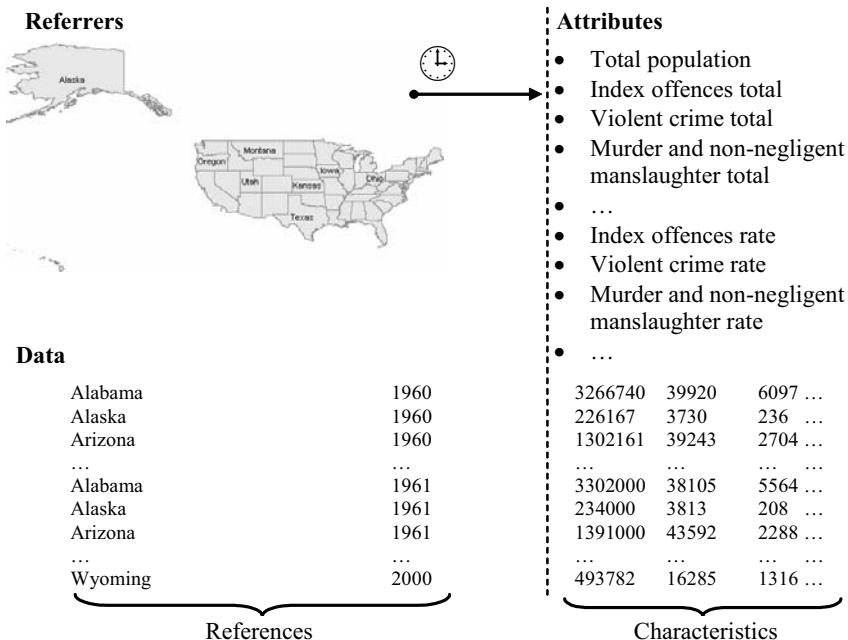
In exploring data about climate, time should not be viewed only as a linear sequence of moments; it is important also to consider the daily and yearly cycles. Since the dataset that we have does not contain diurnal data, the daily cycle is irrelevant in this case.

### 2.3.6 Crime in the USA

For each state of the USA, annual statistics concerning various types of crime are available for the 41 years from 1960 to 2000. The referrers are again space (discretised by means of division of the territory of the USA into states) and time (discretised by division into yearly intervals). The attributes are population number; the total numbers of crimes of various types, such as murder and non-negligent manslaughter, robbery, and burglary, and the crime rate for each type of crime. The attribute values are defined for each state by means of summarising (aggregation) of individual instances. The attribute values corresponding to the yearly time intervals are also defined by aggregation over the intervals. The structure of the dataset is illustrated in Fig. 2.10.

All attributes in this dataset are spatially and temporally continuous, i.e. a value exists for each place and each time moment. Since the values of the attributes result from aggregation over areas and over time intervals, these attributes should be considered as spatially and temporally abrupt. This means that changes of values from one state to another and from one year

to another are not necessarily gradual. All attributes are numeric, with the ratio level of measurement.



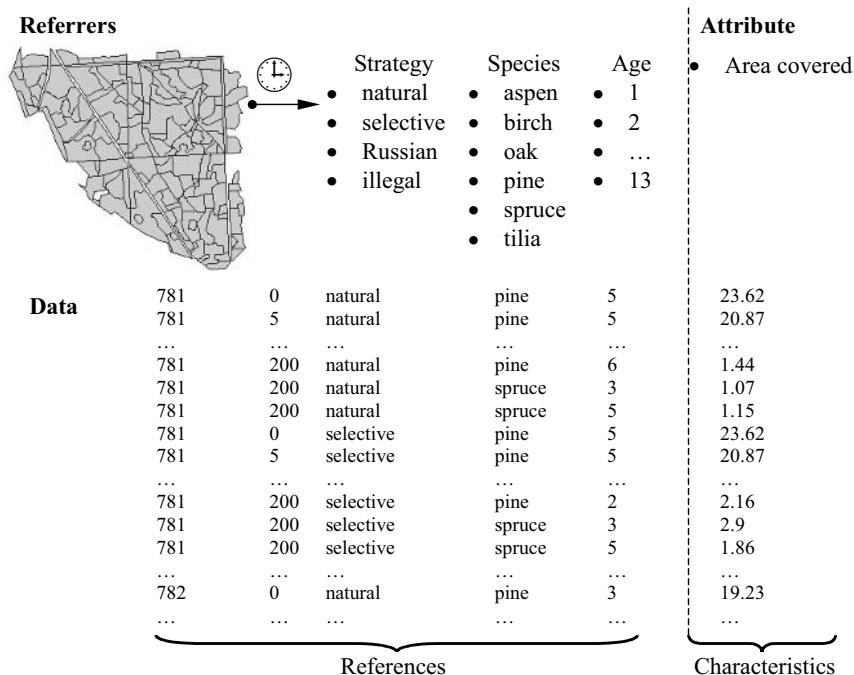
**Fig. 2.10.** The structure of the USA crime dataset

### 2.3.7 Forest Management Scenarios

This is an example of a dataset with multiple (more than two) referrers. Unlike the datasets discussed above, it contains results of simulation rather than measurements of any real-world phenomena. A simulation model was used in order to predict how a particular forest would develop if different forest management strategies were applied.

One of the referrers of this dataset is space, specifically the territory covered by the forest. The territory is divided into compartments with relatively homogeneous conditions inside them. The simulation was done for each of these compartments. Another referrer is time: the simulation covers a 200-year period, and the data are available for every fifth year. One more referrer is the forest management strategy. Four different strategies were considered: natural development without wood harvesting, selective cutting in accordance with the regulations adopted in most western European countries, cutting according to the Russian legal system, and illegal cutting to maximise immediate profit. Two more referrers are tree species,

with six different values, and the age group of the trees, with 13 different values. Hence, the dataset contains five referrers in total. The dataset contains a single attribute, the covered area, which is measured in square metres per hectare. This means that for each forest compartment, year, strategy, species, and age group the data specify the proportion of the area of that compartment that would be covered by trees of that species and that age group in the given year if the given strategy were applied. The structure of the dataset is illustrated in Fig. 2.11.



**Fig. 2.11.** The structure of the dataset of the results of forest development modelling

The three referrers other than space and time have discrete value sets. The attribute is continuous with respect to space and time, since values exist for every compartment and for every time moment. The attribute should be considered as abrupt with respect to space, since the conditions in neighbouring forest compartments may differ significantly, depending, for example, on the tree species planted in them. The attribute may be viewed as partly smooth with respect to time: during periods when no wood is harvested, all changes occur gradually; however, tree cutting results in abrupt changes.

## Summary

In this chapter, we have presented our view of data, which will be used as a basis for our further discussion. The main features of this view can be summarised as follows:

- We aim to consider data in general rather than limit our reasoning only to spatial or spatio-temporal data. At the same time, we always think about how our general ideas could be specialised to the cases of spatial and spatio-temporal data, which are the centre of our attention.
- We aim to abstract from any models of data representation and any specifics of data collection and storage. On the abstract level, data are a set of structured elements (we call them “records”). We are interested in the nature and properties of the components that the elements consist of, as well as in the relationships between these components, but not in how the components are represented and organised. Thus, it is important for us to know that one of the components of a dataset is a set of locations. However, it is not important for our study whether the locations are specified as pairs of geographical coordinates or as addresses. It is also not important whether the data are stored in a relational or object-oriented database or in a file, or whether the spatial information is kept together or separately from the other data.
- In any dataset, we distinguish the components according to their role: we have referential components (referrers), which indicate the context in which each data record was obtained, and characteristic components (attributes), i.e. values observed, measured, or derived in this context. This distinction is the keystone of our framework. We treat data as a function (in the mathematical sense) that assigns particular values of attributes to various combinations of values of referrers.
- The roles played by data components do not depend, in general, on the nature of the components. Thus, space and time may be both referrers and attributes. However, the deep-rooted view of space and time as absolute containers of things and events results in the possibility to transform spatial and temporal attributes into referrers.

The next chapter is intended to explain what we mean by data analysis. At this stage, we shall not speak about methods, i.e. how to analyse data. Before considering any methods, we want to discuss what questions need to be answered in the course of data analysis, or, in other words, what are the typical *tasks* that arise in exploratory data analysis. So, the next chapter is devoted to data analysis tasks.

## References

- (Bertin 1967/1983) Bertin, J.: *Semiology of Graphics. Diagrams, Networks, Maps* (University of Wisconsin Press, Madison 1983). Translated from Bertin, J.: *Sémiologie graphique* (Gauthier-Villars, Paris 1967)
- (Blok 2000) Blok, C.: Monitoring change: characteristics of dynamic geo-spatial phenomena for visual exploration. In: *Spatial Cognition II*, ed. by Freksa, Ch., Brauer, W., Habel., C., Wender, K.F., Lecture Notes in Artificial Intelligence, Vol.1849 (Springer, Berlin, Heidelberg 2000), pp.16–30
- (Chrisman 1997) Chrisman, N.R.: *Exploring Geographic Information Systems* (Wiley, New York 1997)
- (Gray et al. 1998) Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., Pirahesh, H.: Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. In: *Readings in Database Systems*, ed. by Stonebraker, M., Hellerstein, J.M., 3rd edn (Morgan Kaufmann, San Francisco 1998) pp.555–567
- (Jung 1995) Jung, V.: Knowledge-based visualization design for geographic information systems. In: *Proceedings of the 3rd ACM International Workshop on Advances in GIS*, ed. by Bergougnoux, P., Makki, K., Pissinou, N., Baltimore 1995 (ACM Press, New York 1995) pp.101–108
- (Klir 1985) Klir, G.J.: *Architecture of Systems Problem Solving* (Plenum, New York 1985)
- (Langran 1992) Langran, G.: *Time in Geographic Information Systems* (Taylor & Francis, London 1992)
- (MacEachren 1995) MacEachren, A.M.: *How Maps Work: Representation, Visualization, and Design* (Guilford, New York 1995)
- (Merriam-Webster 1999) Merriam-Webster's Collegiate® Dictionary, 10th edn, (Merriam-Webster, Springfield, MA 1999)
- (Peuquet 1994) Peuquet, D.J.: It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers* **84**(3), 441–461 (1994)
- (Peuquet 2001) Peuquet, D.J.: Making space for time: issues in space-time data representation. *Geoinformatica* **5**(1), 11–32 (2001)
- (Peuquet 2002) Peuquet, D.J.: *Representations of Space and Time* (Guilford, New York 2002)
- (Roth and Mattis 1990) Roth, S.M., Mattis, J.: Data characterization for intelligent graphics presentation. In: *Proceedings SIGCHI'90: Human Factors in Computing Systems*, ed. by Carrasco, J., Whiteside, J., Seattle, 1990 (ACM Press, New York 1990) pp.193–200
- (Slocum 1999) Slocum, T.A.: *Thematic Cartography and Visualization* (Prentice Hall, Upper Saddle River 1999)
- (Stevens 1946) Stevens, S.S.: On the theory of scales of measurement. *Science* **103**, 677–680 (1946)
- (Stolte et al. 2002) Stolte, C., Tang, D., Hanrahan, P.: Multiscale visualization using data cubes. In: *Proceedings of the IEEE Symposium on Information*

- Visualization 2002 InfoVis'02*, ed. by Wong. P.C., Andrews, K., Boston, October 2002 (IEEE Computer Society, Piscataway 2002) pp.7–14
- (Verbyla 2002) Verbyla, D.L.: *Practical GIS Analysis* (Taylor & Francis, London 2002)
- (Yuan and Albrecht 1995) Yuan, M., Albrecht, J.: Structural analysis of geographic information and GIS operations from a user's perspective. In: *Spatial Information Theory: a Theoretical Basis for GIS: International Conference COSIT'95, Proceedings*, ed. by Frank, A.U., Kuhn, W., Lecture Notes in Computer Science, Vol.988 (Springer, Berlin, Heidelberg 1995) pp.107–122

## 3 Tasks

### Abstract

In this chapter, we use the metaphor of a mathematical function to identify the types of tasks (questions) involved in exploratory data analysis. A task is viewed as consisting of two parts: the target, i.e. what information needs to be obtained, and the constraints, i.e. what conditions this information needs to fulfil. The target and constraints can be understood as unknown and known (specified) information, respectively; the goal is to find the initially unknown information corresponding to the specified information.

Our task typology has its origin in the ideas expressed by Jacques Bertin in his *Semiology of Graphics* (Bertin 1967/1983). Like Bertin, we distinguish tasks according to the level of data analysis (“reading level”, in Bertin’s terms) but additionally take into account the division of data components into referrers and attributes:

- *Elementary tasks* refer to individual elements of the reference set; for example, “what is the proportion of children in the enumeration district X?”
- *Synoptic tasks* involve the whole reference set or its subsets; for example, “describe the variation of the proportions of children over the whole country” (or “over the southern part of the country”).

The tasks are further divided according to the target (“question type”, in Bertin’s terms), i.e. what is the unknown information that needs to be found. At the elementary level, the target may be one or more characteristics (attribute values) or one or more references (referrer values). For example:

- What is the proportion of children in the enumeration district X? (That is, find the characteristic corresponding to the reference “district X”.)
- In what enumeration districts are the proportions of children 20% or more? (That is, find the references corresponding to the characteristics “20% and more”.)

It is important that, when a task involves several references, each of them is dealt with individually.

We have extended Bertin's ideas by explicitly considering the possible relations between references and between characteristics. Relations may also appear in a task target or may be used in task constraints. For example:

- Compare the proportions of children in district X and district Y (i.e. find the relation between the characteristics corresponding to the references "district X" and "district Y").
- Find districts where the proportion of children is higher than in district X (i.e. find references such that the corresponding characteristics are linked by the relation "higher than" to the characteristic of "district X").

At the synoptic level of analysis, we introduce the notion of a "behaviour" – the set of all characteristics corresponding to a given reference (sub)set, considered in its entirety and its particular organisation with respect to the reference sub(set). The behaviour is a generalisation of such notions as distributions, variations, and trends; for example, the variation of the proportions of children over the whole country or the trend in a stock price over a week.

Synoptic tasks involve reference (sub)sets, behaviours, and relations between (sub)sets or between behaviours. Here are a few examples:

- Describe the variation of the proportion of children over the whole country (target: the behaviour of the proportion of children; constraint: the whole country as the reference set).
- Find spatial clusters of districts with a high proportion of children (target: the reference subset(s); constraint: the behaviour specified as "spatial cluster of high values").
- Compare the distributions of the proportion of children in the north and in the south of the country (target: two behaviours and the relation between them; constraint: two reference subsets, the north and the south of the country).

Elementary tasks play a marginal role in exploratory data analysis, as compared with synoptic tasks. Among synoptic tasks, the most challenging are tasks of finding significant connections between phenomena, such as cause–effect relations or structural links, and of finding the principles of the internal organisation, functioning, and development of a single phenomenon. We call such tasks "connectional".

The main purpose of our task typology is to evaluate the existing tools and techniques for EDA in terms of their suitability for different tasks and

to try to derive operational general principles for tool selection and tool design.

### 3.1 Jacques Bertin's View of Tasks

As George Klir greatly influenced our view of data, so did Jacques Bertin concerning our understanding of how to systemise possible tasks. Therefore, we would like to begin with a discussion of Bertin's ideas.

Let us recall that we use the word "tasks" to denote typical questions that need to be answered by means of data analysis. In his fundamental book on the theory of graphical representation of data (Bertin 1967/1983), Bertin distinguishes possible questions about data first of all according to the data components that they address: "There are as many types of questions as components in the information" (Bertin 1967/1983, p. 10). He explains this concept by means of an example of a dataset containing prices of a stock day by day. This dataset has two components, date and price. Correspondingly, two types of questions are possible:

- On a given date, what is the price of stock X?
- For a given price, on what date(s) was it attained?

Within each type of question, there are three *levels of reading*, elementary, intermediate, and overall. The level of reading indicates whether a question concerns a single data element (such as a single date), a group of elements taken as a whole (e.g. a selected time interval), or all elements constituting the component (e.g. the whole time period that the available data correspond to). Bertin claims: "Any question can be defined by its type and level" (Bertin 1967/1983, p. 10).

What impresses us in this framework? First, this is a systematic approach, unlike that in many other task typologies, which simply enumerate some tasks without providing sufficient background for selecting those particular tasks and without any judgement concerning the completeness of the suggested list. We shall briefly overview such typologies later on. Second, Bertin's typology derives tasks directly from the structure of the data to be analysed. Hence, having a particular dataset, one can easily anticipate the questions that may potentially arise in the course of analysing it. Third, Bertin's framework is free from any bias towards any data analysis methods and tools, whereas some typologies appear to be greatly influenced, for example, by typical GIS functions. Fourth, the tasks are defined in a very operational way: while being tool-independent, they nevertheless provide some hints concerning what tools could support them.

For example, in order to conveniently find an answer to the question “On a given date, what is the price of stock X?” (elementary reading level), one needs a tool that would allow one to specify various dates and would respond by displaying the corresponding values of the price. The question “During the entire period, what was the trend of the price?” (overall reading level) requires some representation of the entire time period and of the price values at each moment. Moreover, the representation of the prices needs to prompt their integrated perception as a single image which will give an analyst a feeling of a “trend” as opposed to just a collection of individual prices.

These features of Bertin’s framework make it especially appropriate for our purposes of providing guidance in choosing EDA tools. So, can we simply take it and use it as it is? Not really. While this approach looks so easy and convincing at first glance, a careful examination reveals some problems that need to be coped with.

The first problem arises already when we analyse thoroughly Bertin’s example of stock prices. Bertin introduces two types of question that are possible for this example dataset and states that, within each type, there are three levels of reading, elementary, intermediate, and overall. For the first type, Bertin formulates possible questions belonging to these levels of reading:

- *Elementary*. On a given date, what is the price of stock X?
- *Intermediate*. In the first three days, what was the trend of the price?
- *Overall*. During the entire period, what was the trend of the price?

However, Bertin does not give analogous examples of the second type of question. Only an elementary-level question is cited: “For a given price, on what date(s) was it attained?”

When we try to extend this question to the other two levels, as was done for the first question, this suddenly turns out to be quite difficult to do; at least, we could not find any sensible formulations. The reason for such an asymmetry might be that the components “time” and “price” are not equivalent in terms of their roles in the data. According to our view, time is a referrer, while the price is an attribute. It seems that the notion of reading levels applies only to referrers and not to attributes.

This hypothesis could be justified in the following way. An intermediate- or overall-level question can be generalised as “What is happening to component A when component B assumes several different values from set S?” Or, for more than two components, “What is happening to component A when components B, C, ... assume several different values from sets S<sub>1</sub>, S<sub>2</sub>, ...?” To obtain an answer to this question, an analyst would take various values from sets S<sub>1</sub>, S<sub>2</sub>, ..., assign them to components B, C,

..., and determine the corresponding values of component A. Thus, in Bertin's examples, one would take consecutive time moments from some interval and observe how the price changes.

As we have explained before, a dataset consists of independent components (referrers), which can potentially assume arbitrary values, and dependent components (attributes), whose values are determined by the values assumed by the referrers. If one were to try to assign values from an arbitrary set S to a dependent component, it might occur that for many values there are no corresponding references (i.e. these values have never been attained), while for other values there are multiple references (i.e. these values have been attained more than once). In general, if Y depends on X, it seems quite strange to ask "What happens to X if Y varies within the set S?", while the opposite question sounds quite natural. This gives us some grounds for regarding intermediate- and overall-level questions as questions concerning changes of attribute values when referrers vary within (i.e. take various values from) some arbitrarily chosen value sets.

Further problems arise when we try to apply Bertin's schema to datasets with a more complex structure than just two components. Koussoulakou and Kraak (1992) made an observation concerning spatio-temporal data that the distinction according to reading level can be applied independently to the spatial and temporal dimensions of the data. For example, the question "When do the maximum values occur at location *l*?" belongs to the elementary (local) level with respect to the spatial component of the data and to the overall level with respect to the temporal component. In total, nine combinations of reading levels for space and time are possible. Example questions corresponding to these combinations are given in Table 3.1.

Let us consider a quite different example, with no space and time explicitly involved. Suppose that the salaries of the employees in some company vary depending on the job and the employee's age. Then, it is possible to ask questions such as:

- For a given job, how does the salary depend on age?
- For a given age, what is the range of salaries for all possible jobs?

The first question belongs to the elementary level with respect to the job and to the overall level with respect to age, while the opposite is true for the second question.

In general, when the number of referential components is two or more, it seems that the reading level for each of them may be chosen independently of the others. This results in  $3^N$  possible combinations of reading levels for a dataset with  $N$  referrers.

**Table 3.1.** Reading levels and example questions for spatio-temporal data, according to Koussoulakou and Kraak (1992)

Space Time	Elementary level	Intermediate level	Overall level
<b>Elementary level</b>	What is the population density at location P at time $t_i$ ?	In which neighbourhoods is the population density $d_2$ at time $t_i$ ?	Where does the highest population density occur at time $t_i$ ?
<b>Intermediate level</b>	How does the population density develop at location P from time $t_i$ to time $t_j$ ?	In which neighbourhoods is the population density $d_2$ during the time period from $t_i$ to $t_j$ ?	Where does the highest population density occur during the time period from $t_i$ to $t_j$ ?
<b>Overall level</b>	What is the trend in population density at location P over the whole time?	Which are the neighbourhoods where the population density remains at $d_2$ during the whole time?	What is the trend in high population densities over the whole time?

Furthermore, even if we choose the same intermediate or overall reading level for two or more referrers, we shall still be able to formulate questions with different meanings. Thus, Koussoulakou and Kraak suggest the following question as an example of an overall-level question with respect to both space and time: “What is the trend over the area during the whole time?” By our observation, there are at least two more questions that could also be classified as overall-level questions with respect to both space and time:

- How has the spatial distribution evolved over time? A possible answer to this question would be “the cluster of high values moved from the centre of the area to the north”.
- How do the temporal trends vary over the area? A possible answer would be “the values increased in the north and decreased in the south”.

The difference between these variants of question is essential, because different analyses would be required to answer them. This means that the question typology needs to be elaborated further in order to capture this divergence.

There is one more problem with Bertin’s schema. While Bertin acknowledged the importance of comparisons in data exploration (thus, he advocated the use of graphics that enable easy comparisons), he did not explicitly include the notion of comparison in his suggested framework.

We would like to amend this framework by explicitly introducing this and related notions.

In the following sections, we present our elaboration of Bertin's schema for task (question) classification.

## 3.2 General View of a Task

Any task (question) implies two parts: a target, i.e. what information needs to be obtained, and the constraints, i.e. what conditions this information needs to fulfil. The target and constraints can also be viewed as unknown and known (specified) information, respectively; the goal is to find the initially unknown information corresponding to the specified information. The simplest example is to find the value of an attribute corresponding to a certain specified reference. This is actually a generic task that stands for a great number of specific questions, such as:

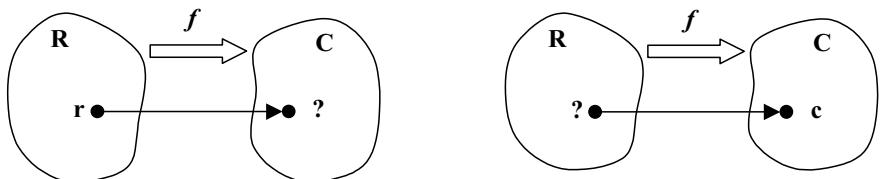
- On a given date what is the price of stock X?
- What was the population of Lisbon in the year 1991?
- In a given location in Europe, what is the proportion of land covered by broadleaved forest?
- Where was the stork Penelope on 1 September 1998?

In this group of tasks, the attribute value is the target, while the specified reference defines the constraint. This means that an arbitrary attribute value cannot be accepted as an answer; this can only be the attribute value corresponding to this particular reference. The reference is specified in terms of the values of all the referential components present in the dataset. Thus, in the question "What was the population of Lisbon in the year 1991?", the reference consists of a certain location (Lisbon) and a certain year (1991). In the question concerning the stork, the reference consists of a certain object (stork), identified by the name Penelope, and a certain time moment, specifically 1 September 1998.

A graphical illustration of this group of questions is shown in Fig. 3.1, left. Here,  $\mathbf{R}$  is the reference set of a dataset;  $\mathbf{C}$  is the characteristic set;  $f$  is the data function, which defines the correspondence between the elements of the reference and characteristic sets; and  $r$  is some specified reference, i.e. an element of the set  $\mathbf{R}$ . The corresponding characteristic, which is determined by the data function, is unknown and needs to be found. This task target is indicated in the picture by a question mark.

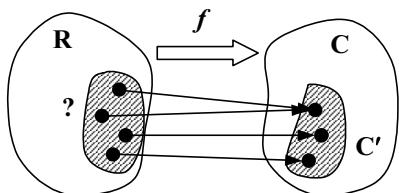
The opposite case is when some attribute value is specified and the goal is to find the corresponding reference or references. Hence, the references

are the target and the attribute value defines the constraint. This kind of task is illustrated graphically in the right part of Fig. 3.1. Here, the question mark indicates the unknown reference.



**Fig. 3.1.** Schematic representations of the task of determining the characteristic that corresponds to a given reference  $r$  according to the data function  $f$  (left), and the task of determining the reference that corresponds to a given characteristic  $c$

In tasks of the second type, inexact constraints are very often used. This means that a subset or range of attribute values is specified rather than an individual value. Therefore, the graphical representation of such a task could be modified, as shown in Fig. 3.2.



**Fig. 3.2.** In tasks of finding references corresponding to given characteristics, the characteristics may be specified imprecisely, that is, a subset (designated as  $C'$ ) rather than an individual element of the characteristic set  $C$  may play the role of a task constraint. The target comprises all references corresponding to any of the elements of the subset  $C'$

Here are some instantiations of the generic task of finding references corresponding to attribute values:

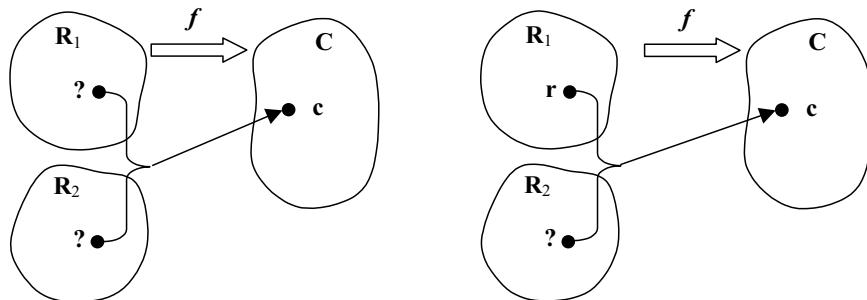
- For a given price, on what date(s) was it attained?
- Which municipalities in Portugal had 300 000 or more inhabitants, and when?
- Where in Europe is at least 80% of the area covered by broadleaved forest?
- Which stork(s) were near Lake Victoria, and when?

It may be noticed that the second and fourth examples actually contain two questions each. In the second example, the questions are “where” and

“when”, and in the fourth example, there are the questions “which” and “when”. This is an effect of the presence of two referential components in the respective datasets. The questions require one to find combinations of values of referrers corresponding to values of the attribute, i.e. the value of each referrer needs to be determined. This is illustrated in the left part of Fig. 3.3. However, nothing prevents us from formulating simpler questions, in which a value of one referrer is specified, and only the value of the other referrer needs to be found. For example:

- Which municipalities in Portugal had 300 000 or more inhabitants in the year 1991?
- In what year(s) did the municipality of Loures have 300 000 or more inhabitants?
- Which stork(s) were near Lake Victoria in February 1999?
- When was the stork Prinzessin near Lake Victoria?

In each of these examples, the specified value of one of the referrers adds one more constraint to the task, and the value of the other referrer is the target. This situation is shown schematically in the right part of Fig. 3.3.



**Fig. 3.3.** For a dataset with multiple referential components, values of either all or just some of the referrers may be task targets. This picture illustrates a case with two referrers. On the left, the values of both referrers are unknown and need to be found. On the right, the value of one referrer is specified, i.e. is a part of the task constraint, and only the value of the other referrer is the task target

Another option is to say that the value of one of the referrers is not important, i.e. this may be any possible value. For example:

- Which municipalities in Portugal had 300 000 or more inhabitants in any census year?
- In what year(s) were there any municipalities with 300 000 or more inhabitants?

- Which stork(s) were ever near Lake Victoria?
- At what times were any of the storks near Lake Victoria?

In order to bring some kind of order into all this variety, let us introduce a very simple formal notation, which can be understood without significant mathematical background. The worth of this notation is that it allows us to represent compactly the possible types of questions, or general tasks. For the notation, we use the metaphor of a mathematical function: a set of data may be represented by a function that assigns particular values of attributes to various references. References are specified as combinations of values of all referrers. If there is a single referrer, its values constitute the possible references. The set of all references present in a dataset will be called the reference set of this dataset. The set of all possible combinations of values of the attributes will be called the characteristic set.

So, a dataset may be represented by a formula such as

$$f(x) = y \quad (3.1)$$

where  $f$  is a function symbol,  $x$  is the independent variable, and  $y$  is the dependent variable. The variable  $x$  may have various elements of the reference set as its values, and the function  $f$  assigns the corresponding elements of the characteristic set to the dependent variable  $y$ .

In the general case, the values of both  $x$  and  $y$  are combinations:  $x$  assumes combinations of values of the referrers, and  $y$  is assigned combinations of values of the attributes. Therefore, the formula  $f(x) = y$  can be rewritten in a more “detailed” manner to represent the structure of the dataset, i.e. the number of referential and characteristic components:

$$f(x_1, x_2, \dots, x_M) = (y_1, y_2, \dots, y_N) \quad (3.2)$$

where  $M$  is the number of referrers,  $N$  is the number of attributes in the dataset, the independent variables  $x_1, x_2, \dots, x_M$  stand for the referrers, and the dependent variables  $y_1, y_2, \dots, y_N$  stand for the attributes.

Since the values acquired by the attributes are determined only by the references and not by values of other attributes, it is possible to consider any attribute independently of the other attributes. Therefore, the formula (3.2) can be rewritten in a more usual way, as a set of formulas with only one dependent variable in each of them:

$$\begin{aligned} f_1(x_1, x_2, \dots, x_M) &= y_1 \\ f_2(x_1, x_2, \dots, x_M) &= y_2 \\ &\dots \\ f_N(x_1, x_2, \dots, x_M) &= y_N \end{aligned} \quad (3.3)$$

Here, we have split the initial function  $f$  into  $N$  functions  $f_1, f_2, \dots, f_N$ . Each of these functions defines values of one of the attributes on the basis of the values of the referrers.

Let us now consider the task of finding the characteristic (i.e. the value of a single attribute or a combination of values of several attributes) corresponding to a specified reference. This can be viewed as substituting the variable  $x$  in the formula  $f(x) = y$  by a specific value  $\mathbf{r}$  and determining the corresponding value of  $y$  using the function  $f$ . We shall represent this task compactly as follows:

$$\text{?}y: f(\mathbf{r}) = y \quad (3.4)$$

where  $\mathbf{r}$  is a particular element of the reference set, i.e. a constant rather than a variable. The expression “? $y$ ” denotes the question target, which is, in this case, the characteristic corresponding to the reference  $\mathbf{r}$ , i.e. some value of the dependent variable  $y$ . If the reference  $\mathbf{r}$  is actually a combination of values of  $M$  referrers, the task may be represented in a more detailed fashion as

$$\text{?}y: f(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M) = y \quad (3.5)$$

where  $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M$  are specific values of the individual referrers.

The expressions (3.4) and (3.5) demonstrate our general approach to representing tasks by formulas. At the beginning, the task target is indicated, and then, after a colon, the constraint(s) are specified. The target is represented as a variable and labelled by a question mark to emphasise that this is unknown information, which needs to be found. The constraints are expressed through constants that are substituted for the appropriate variables.

If we apply the same scheme to the task of finding the reference(s) corresponding to a specified characteristic, we obtain the following:

$$\text{?}x: f(x) = \mathbf{c} \quad (3.6)$$

where  $\mathbf{c}$  is a particular element of the set of possible characteristics and  $x$  is the independent variable, i.e. the variable corresponding to the referential component of the data. In the case of multiple referrers, this can be rewritten as

$$\text{?}x_1, x_2, \dots, x_M: f(x_1, x_2, \dots, x_M) = \mathbf{c} \quad (3.7)$$

Equation (3.7) contains as many variables in the question target as there are referrers in the dataset. This is the situation that we had in the following example questions:

- Which districts in Portugal had 300 000 or more inhabitants, and when?

- Which stork(s) were near Lake Victoria, and when?

Each of these examples contains, in fact, two questions, corresponding to the number of referrers in the data. However, as we have demonstrated, we can introduce additional constraints by choosing specific values of one of the referrers while the other referrer remains the question target. If we have more than two referrers, we can choose particular values of some of them and let the other referrers be the targets. For example, the formula (3.8) encodes a task in which the value of the first referrer of  $M$  referrers of a dataset needs to be determined, while the values of the remaining  $M-1$  referrers, as well as the corresponding characteristic, are specified as task constraints:

$$\text{?}x_1: f(x_1, \mathbf{r}_2, \dots, \mathbf{r}_M) = \mathbf{c} \quad (3.8)$$

Here,  $\mathbf{r}_2, \dots, \mathbf{r}_M$  are some specific values selected for the 2nd, ...,  $M$ th referrers, respectively. Here, the variable  $x_1$  denotes the target of the task, which is the value of the first referrer. Quite analogously, we could encode tasks with the second, third, ..., or  $M$ th referrer as the target. Similarly, any two referrers may be chosen as targets, and so on. In general, for a dataset with  $M$  referrers there are  $2^M - 1$  different variants of the choice of target referrer(s).  $2^M$  is the number of possible subsets of a set containing  $M$  elements. We decrease this number by one since we exclude the variant where the values of all referrers are specified, i.e. where there is no real target.

As we have mentioned earlier, in tasks of finding references corresponding to specified characteristics, the characteristics are often defined imprecisely. For example, in the task “Find municipalities with 300 000 or more inhabitants in 1991”, a value interval of a numeric attribute rather than an exact value is used as a constraint. The task “When was the stork Prinzessin near Lake Victoria?” includes an approximate indication of a location, which is a value of a spatial attribute. In general, a subset of characteristics is often included in the constraints of a task rather than a single characteristic. The expected answer to such a task consists of all references where the corresponding characteristics are contained in the subset specified. Thus, the answer to the question concerning the municipalities would include Porto, with a population of 302 000, Loures, with 322 000, and Lisbon, with 663 000. All these values are contained in the subset of attribute values specified as “300 000 or more”. The answer to the question concerning the stork Prinzessin does not depend on whether the stork was to the south or to the north of Lake Victoria: all locations around the lake satisfy the task constraints.

Hence, (3.6) can be generalised as follows:

$$\exists x: f(x) \in \mathbf{C}' \quad (3.9)$$

where  $\mathbf{C}'$  is a certain subset of the set of all possible characteristics and  $\in$  is the symbol denoting membership in a set. In the case of multiple referrers, (3.9) can be “unfolded” as was demonstrated in (3.7) and (3.8). In our further discussion, we shall mostly use the “folded” notations, as in (3.1), (3.4), and (3.9), but it should be kept in mind that these may be abbreviated forms of equations containing multiple independent variables.

However, let us briefly return to (3.8). In this equation, we have put some constants (specific values) in place of the independent variables  $x_2, \dots, x_M$  representing the referrers from the second to the  $M$ th of the dataset. This means that we have imposed precise constraints on the values of these referrers. However, as for attributes, imprecise constraints could also be used for referrers. In other words, one could constrain areferrer by specifying a set of possible values rather than an individual value. Such a constraint means that any value from the specified set is acceptable in the answer. For example, the question “Which storks were near Lake Victoria in February 1999?” contains an imprecise specification of the time: February 1999. The granularity of the temporal referrer in this dataset is one day; consequently, the phrase “February 1999” describes a set consisting of 28 time moments. Any of these 28 time moments is suitable for an analyst seeking an answer to this question.

Therefore, we extend (3.8) and the like by allowing sets to be put in places corresponding to any referrers, as well as to attributes. For generality, we shall always assume that sets, which can consist of one or more elements, are used. For example;

$$\exists x_1: f(x_1, \mathbf{R}'_2, \dots, \mathbf{R}'_M) \in \mathbf{C}' \quad (3.10)$$

Here,  $\mathbf{R}'_2, \dots, \mathbf{R}'_M$  are some specified subsets of the value domains of the referential components from the second to the  $M$ th. Moreover, nothing prevents us from using the full sets rather than subsets for some of the components. This is what is actually done in cases when the values of these components are of no importance, as in the questions

- Which municipalities in Portugal had 300 000 or more inhabitants in any census year?
- In what year(s) were there any municipalities with 300 000 or more inhabitants?
- Which stork(s) were ever near Lake Victoria?
- At what times were any of the storks near Lake Victoria?

In each of these questions, it is assumed that one of the referrers (time, space, or stork) can take any value from its value domain.

It seems that all of the example questions cited at the beginning of this section now fit into our schema of task representation. Let us now summarise our approach. The key idea is to describe generic tasks, or task types, by using the notion of a mathematical function as a metaphor for the data. We represent tasks by quasi-algebraic formulae, in which we use question marks to indicate the unknown information, i.e. the target of the task. We have thus far considered only rather simple tasks, but the same idea can be used for more complex tasks. In the following sections, we shall present our task typology in its full extent.

A note can be made concerning the relation of our notation to Bertin's notion of question type. To recall, the type of a question, according to Bertin, reflects which component is the focus of the question, i.e. contains the potential answer to the question. In our terms, this is the question's target. Hence, Bertin's type corresponds to the place in which the variable denoting the task target stands in our formal notation for the task. However, the formula (3.7), and also any other formula with more than one target variable, cannot be subsumed under Bertin's typology. While Bertin states "There are as many types of questions as components in the information", we have shown that, for a dataset with  $M$  referrers and  $N$  attributes, there are  $2^M - 1$  variants of the selection of referrers as question targets, in addition to  $N$  possible questions concerning attributes (since attributes can be considered independently of each other, a question with two or more attributes in the target is equivalent to a group of questions each targeting a single attribute). So, we have extended the system of question types introduced by Bertin. This is not the only extension we make; some other extensions will be described below.

### 3.3 Elementary Tasks

Like Bertin, we distinguish elementary tasks from tasks of higher level. All tasks considered in the previous section are elementary. We call them elementary because they address individual elements of data, i.e. individual references or individual characteristics.

This seems to contradict the statement that sets as well as individual elements can be used in task constraints, i.e. put in place of some variables in a task formula. Actually, there is no contradiction: such use of a set does not mean that the set needs to be considered as a whole. It means only that any individual element can be taken from this set. Let us examine, for ex-

ample, the question “Which storks were near Lake Victoria in February 1999?” As we have already discussed, the phrase “February 1999” specifies a time interval (i.e. a set) rather than an individual moment. However, the question does not ask about storks that lived near Lake Victoria or moved around during the whole of February. Rather, it asks about storks that happened to be near Lake Victoria on any date in February but could be elsewhere on other dates.

So, we define *elementary tasks* as tasks that do not imply dealing with sets of references or characteristics as wholes but, rather, address their elements. Tasks that do not comply with this definition will be called *synoptic tasks*, as involving a general view of a set as a whole.

### 3.3.1 Lookup and Comparison

For a general representation of an elementary task, we use the formula  $f(x)=y$ , or its extended variant  $f(x_1, x_2, \dots, x_M) = y$  in a case of multiple referrers. Some variables in these formulas can be task target(s), while the other variables can be replaced by particular values or value sets specifying task constraints. Tasks that can be represented in this way will be called *lookup tasks*. We have discussed such tasks in the previous section and will not go into further detail, beyond introducing the distinction between direct lookup tasks and inverse lookup tasks. *Direct lookup tasks* are those where references are specified and the goal is to find the corresponding characteristics. In contrast, *inverse lookup tasks* are tasks where references corresponding to specified characteristics need to be found. This includes also tasks where references are partly specified, as in (3.10), and the goal is to complete the specification.

Hence, the tasks represented by (3.4) and (3.5) are called direct lookup tasks, and (3.6)–(3.10) and the like represent inverse lookup tasks.

Let us investigate what other types of elementary tasks may exist. While the data function  $f$  relates references to characteristics, and lookup tasks deal with these relations, there are also relations within the reference and characteristic sets, i.e. between references and between characteristics, and hence there may be tasks involving these relations.

Let us give some examples of what relations we mean. For any two elements of any set, it is possible to say whether they are the same or different, or, in other words, equal or not equal. The relations “equal” and “not equal” are often denoted using the symbols  $=$  and  $\neq$ . In an ordered set, the elements are linked by ordering relations:  $<$  (less than),  $\leq$  (less than or equal),  $>$  (greater than), and  $\geq$  (greater than or equal). When we speak about time, we usually use other names for the equality and ordering rela-

tions: simultaneous, not simultaneous, earlier than, and later than. Space is, in general, not ordered. However, if a particular coordinate system is introduced into space, it is possible to consider various relations specific to this coordinate system. Thus, in geographical space, various directional relations exist, such as “south of”, “west of”, and “south-west of”.

It has been mentioned in the previous chapter that the values of a data component may be not only individual items but also subsets of some base set. The subsets can be built on the basis of a discrete set, such as nationalities living in different countries, or a continuous set, such as ranges of real numbers, areas in space, and intervals of time. Two possible relationships for sets are inclusion (all elements of one set are contained in another set) and overlapping (two sets have some common elements). For subsets of a continuous set, an important relation is adjacency, in particular, adjacency in space or time. An element may be related to a (sub)set by the relation  $\in$  (a member of) or  $\notin$  (not a member of).

For a set with distances, it is possible not only to indicate relations between elements but also to express them numerically, in terms of the distances between the elements. For numeric attributes, distances may be defined as differences between numbers. Distances in time can also be defined as differences between time moments. Distances in space can be defined in many ways. Thus, the distance may be just the Euclidean distance on a plane. For geographical space, the curvature of the surface of the Earth can be taken into account, and so can relief. Distances are often determined by measuring the length of the actual path one needs to take to get from one place to another. For example, cars can only move along roads, and ships must go around islands and keep away from shoals and reefs.

In general, for a set of arbitrary nature, any method of defining distances may be chosen in accordance with the purposes of the data analysis, providing that certain requirements are fulfilled:

1. The distance from any element to itself is zero.
2. For any two elements A and B, the distance from A to B is the same as the distance from B to A.
3. For any three elements A, B, and C, the distance from A to C is not more than the sum of the distances from A to B and from B to C.

Relations between values of attributes with the ratio level of measurement can also be expressed numerically as ratios or percentages.

One may ask various questions concerning the relations that exist between elements or subsets of a specific set. Such questions, which will be called *relational questions* or *relational tasks*, may be constructed according to one of the following general schemes:

1. How are the elements  $p$  and  $q$  (or the subsets  $P$  and  $Q$ ) of the set  $S$  related?
2. What element (or subset) of the set  $S$  is related to the element  $p$  (or subset  $P$ ) in the way  $\rho$ ?
3. What elements (or subsets) of the set  $S$  are related in the way  $\rho$ ?

However, relational questions formulated in any of these “pure” forms are not typical of data analysis: they address general properties of the sets from which the references and characteristics are taken and have no relevance to any particular dataset. Thus, it is not necessary to have any data in order to answer questions such as

- How are the numbers 1 and 2 related? How are the months September and November of the same year related?
- What number is twice as great as 1? What months precede September?
- What are the pairs of numbers where one number is twice as great as the other? What are the months with a two-month interval between them?

Here is a pair of example questions that sound more usual in the context of data analysis (many further examples will be presented later):

- Did the population of Porto in 1981 exceed 300 000?
- Where was the stork Prinzessin on 1 February 1999 in relation to Lake Victoria?

Let us try to figure out the major distinction between the latter pair of questions and the former group.

In order to answer the first question of the latter pair, it is necessary to find out the population of Porto in 1981 (i.e. to perform a lookup task) and then compare the number found with 300 000. To answer the second question, it is necessary to determine the geographical position of the stork Prinzessin on 1 February 1999 (this is again a lookup task) and then compare this position with the location of Lake Victoria. The latter is supposed to be known or, if not, must be obtained from another lookup task. Hence, these two relational tasks are compound tasks that include one or more lookup tasks. These lookup tasks may be called *subtasks* of the relational tasks.

Such a compound organisation is not a unique feature of these two example tasks but a general property of all data analysis tasks dealing with relations between characteristics, as in these examples, or between references. Generally, the *data function*, i.e. the correspondence between the references and characteristics, is involved in all data analysis tasks. Therefore, questions about relations between elements or subsets of the same set, be it the set of references or the set of characteristics, arise only when

those elements (or at least one of them) result from answering some other questions that involve the data function.

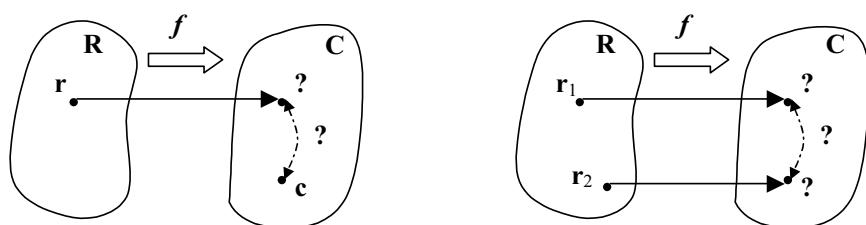
Let us return to the three general schemes of relational tasks listed above. They show us that a relation may be the target of a relational task or may be specified as a task constraint. Thus, in scheme 1, the relation between the elements  $p$  and  $q$  or the subsets  $P$  and  $Q$  needs to be found; hence, this relation is the task target. In schemes 2 and 3, a certain relation (designated as  $\rho$ ) is specified; hence, this is one of the task constraints.

We shall call tasks built according to the first scheme, i.e. having relations in their targets, *comparison tasks*. We use the word “comparison” in the general sense of determining what relation exists between two or more items. Usually, the expected type of relation is specified: equality, order, distance, ratio, direction in space, inclusion or overlapping of subsets, adjacency of continuous subsets, etc.

Tasks built according to the second or third scheme can be viewed as being inverse with respect to comparison tasks: certain relations are specified, and items that are related in the specified way need to be detected. Such tasks will be called *relation-seeking tasks*.

We are now going to focus for a while on comparison tasks and then consider relation-seeking tasks.

Comparison data analysis tasks are built according to the general scheme “How are the elements  $p$  and  $q$  (or the subsets  $P$  and  $Q$ ) of the set  $S$  related?”, where at least one of the elements  $p$  and  $q$  (or subsets  $P$  and  $Q$ ) is not specified explicitly but must result from some lookup task, or, in other words, be the target of some lookup task. This may be a direct or inverse lookup task. A graphical illustration of comparison tasks involving direct lookup tasks is given in Fig. 3.4.



**Fig. 3.4.** Graphical illustrations of two forms of a comparison task involving a direct lookup task. Left: for a given reference  $r$ , the corresponding characteristic needs to be found and then to be compared with a specified characteristic  $c$  (i.e. a constant). Right: for two given references  $r_1$  and  $r_2$ , the corresponding characteristics need to be found and then compared

The picture on the left illustrates a task that requires one to find the characteristic corresponding to a given reference  $\mathbf{r}$  and compare the characteristic found with a given characteristic  $\mathbf{c}$ . The unknown characteristic and unknown relation, which must be obtained from the comparison, are designated by question marks. This demonstrates that the task has two targets: the characteristic and the relation. The corresponding task formula may look as follows:

$$\text{?}y, \lambda: f(\mathbf{r}) = y; y\lambda\mathbf{c} \quad (3.11)$$

Here,  $\mathbf{r}$  stands for the specified reference,  $\mathbf{c}$  for the specified characteristics, and  $y$  for the unknown characteristic corresponding to  $\mathbf{r}$ ; the symbol  $\lambda$  stands for the unknown relation between  $y$  and  $\mathbf{c}$ , which needs to be determined.

The diagram on the right of Fig. 3.4 illustrates a task that requires one to find the characteristics corresponding to two given references  $\mathbf{r}_1$  and  $\mathbf{r}_2$  and compare the two characteristics found, i.e. determine how they are related. The two unknown characteristics and the unknown relation are designated by question marks. There are three question marks, which indicates that the task has three targets: two characteristics and one relation. This case may be represented by the formula

$$\text{?}y_1, y_2, \lambda: f(\mathbf{r}_1) = y_1; f(\mathbf{r}_2) = y_2; y_1\lambda y_2 \quad (3.12)$$

Here,  $\mathbf{r}_1$  and  $\mathbf{r}_2$  are two specified references. The goal is to find their characteristics, which are denoted by  $y_1$  and  $y_2$ , and then to compare these characteristics, i.e. to determine the relation between  $y_1$  and  $y_2$ , which is denoted by  $\lambda$ . Some examples of such tasks are:

- Compare the stock price on the first day with that on the last day.
- Which municipality, Porto or Loures, had more inhabitants in 1981?
- Did the population of Porto increase or decrease from 1981 to 1991?
- What were the relative positions of the storks Penelope and Peterchen on 1 February 1999?

There are also comparison tasks where the goal is to compare values of two or more different attributes corresponding to one and the same reference. For example:

- How does the number of people without primary school education in Porto in 1991 compare with the number of high school students?
- Which rate was higher in the District of Columbia in 2000, the burglary rate or motor vehicle theft rate?

In some cases, it may also make sense to compare values of different attributes corresponding to different references, for example, “compare the immigration to the USA with the emigration from Mexico”.

Tasks of comparison of values of two different attributes corresponding to the same reference or to different references may be represented by the formulae

$$\exists y_1, y_2, \lambda: f_1(\mathbf{r}) = y_1; f_2(\mathbf{r}) = y_2; y_1 \lambda y_2 \quad (3.13)$$

$$\exists y_1, y_2, \lambda: f_1(\mathbf{r}_1) = y_1; f_2(\mathbf{r}_2) = y_2; y_1 \lambda y_2 \quad (3.14)$$

In these formulae,  $f_1$  and  $f_2$  stand for two different attributes. Such tasks make sense only if the attributes are comparable, i.e. their value domains are the same or at least overlap.

As we have already said, the potential answers to comparison tasks are various relations from a certain set of possible relations, which depends on the properties of the set of items involved, such as the presence of ordering, distances, etc. Moreover, it is rather typical in data analysis that relations between items are not only designated verbally but also, whenever possible, measured numerically. For example, when an analyst wants to answer the question “Which municipality, Porto or Loures, had more inhabitants in 1981?”, he/she may be interested not only in detecting that Porto had more inhabitants than Loures but also in finding out that Porto had 50 900 more inhabitants than Loures. The question concerning the population change in Porto from 1981 to 1991 typically implies an answer such as “the population decreased by 24 900 inhabitants” rather than simply “the population decreased”. A satisfactory answer to the question concerning the relative positions of the storks could be “Peterchen was 1800 km to the west of Penelope”.

We have also mentioned that relations between values of attributes with the ratio level of measurement can be expressed numerically as ratios or percentages. For example, “the population of Porto in 1981 was 1.2 times bigger than that of Loures” or “the population of Porto decreased by 7.6% from 1981 to 1991”.

Taking all this into account, we adopt the following extended treatment of the notion of comparison: comparison means identification of the kind of relation existing between two or more elements of some set and, whenever permitted by the properties of the set, numerical specification of this relation on the basis of the distances or ratios between the elements.

In the above, we have considered comparison tasks involving direct lookup subtasks. In these tasks, relations between characteristics need to be determined. Let us now consider tasks that imply finding relations between references, i.e. values of referrers. As in the previous case, the refer-

ences are not specified explicitly but are supposed to result from some lookup tasks that have these references in their targets (i.e. inverse lookup tasks). Here are some examples:

- Which of the storks, Prinzessin or Sohn von Prinzessin, reached Lake Victoria earlier?

This task can be decomposed into three subtasks:

1. When did the stork Prinzessin reach Lake Victoria? (Answer: on 15 December 1998.)
2. When did the stork Sohn von Prinzessin reach Lake Victoria? (Answer: on 3 February 1999.)
3. Which of these two dates is earlier? (Answer: the first date is earlier, and hence Prinzessin reached Lake Victoria earlier than Sohn von Prinzessin.)

The first two questions are inverse lookup tasks and the third one is the task of determining the relation between the values of the temporal referent.

- Was the set of municipalities in Portugal with a population over 300 000 the same in 1991 as in 1981?

As in the previous example, there are two inverse lookup subtasks here

1. Which municipalities in Portugal had a population over 300 000 in the year 1981? (Answer: Lisbon and Porto.)
2. Which municipalities in Portugal had a population over 300 000 in the year 1991? (Answer: Lisbon, Porto, and Loures.)

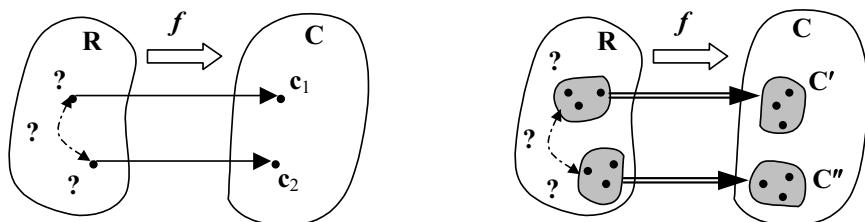
In addition, there is a subtask of comparison of two subsets of the set of municipalities, which is one of the referrers in the dataset containing the Portuguese census data:

3. Do the first and the second set consist of the same elements? (Answer: no; the second set contains one additional element, namely Loures.)

By analogy with lookup tasks, comparison tasks targeting relations between characteristics may be called *direct comparison tasks*, and comparison tasks where relations between references need to be determined may be called *inverse comparison tasks*. Inverse comparison tasks may be illustrated graphically, as is shown in Figure 3.5.

The diagram on the left represents the task variant where two particular characteristics, i.e. elements of the characteristic set, are specified as the constraints of the inverse lookup subtasks. These two characteristics are denoted by  $c_1$  and  $c_2$ . The goal is to find the corresponding references and

the relation between these references. The diagram reflects the case where there is exactly one reference corresponding to each of the characteristics  $c_1$  and  $c_2$ . However, in the general case, several references may have the same characteristics, and hence a subset of references may correspond to  $c_1$ ,  $c_2$ , or both. In this case, a relation between subsets or between a subset and an element is of interest.



**Fig. 3.5.** A graphical representation of inverse comparison tasks. On the left,  $c_1$  and  $c_2$  are two specified elements of the characteristic set. The goal is to find the corresponding references and the relation between them. On the right,  $C'$  and  $C''$  are two specified subsets of the characteristic set. The goal is to find the corresponding references or subsets of references and the relation between them. In general, subsets of references may also correspond to individual characteristics

In the right part of Fig. 3.5, subsets of characteristics are specified as the constraints. These subsets are denoted by  $C'$  and  $C''$ . The goal is to find the corresponding references or subsets of references and the relation between them. To simplify the picture, we have represented the correspondence between the references and characteristics by drawing double arrows from subsets of references to the subsets of corresponding characteristics, instead of multiple ordinary arrows from the individual references to the corresponding individual characteristics.

The formula representing inverse comparison tasks looks as follows:

$$\exists x_1, x_2, \lambda: f(x_1) \in C'; f(x_2) \in C''; x_1 \lambda x_2 \quad (3.15)$$

This notation means the following: find the reference(s) corresponding to the set of characteristics  $C'$  and the reference(s) corresponding to the set of characteristics  $C''$ , and then the relation between the former and the latter. Hence, we have explicitly represented the inverse lookup tasks  $f(x_1) \in C'$  and  $f(x_2) \in C''$  that need to be performed before the comparison operation may take place.

We shall not consider in detail all possible variants of inverse comparison tasks. Actually, the variants differ only in the form of the inverse lookup tasks involved. However, not any combination of lookup tasks is meaningful; the results of these tasks must be comparable. Thus, if the first

lookup subtask results in a spatial reference and the second in a temporal reference, these references cannot be compared. An example of such a meaningless task could be: “Compare the year when the population of Loures exceeded 300 000 with the set of municipalities that had more than 300 000 inhabitants in 1981”. If we represent the population data as  $p(m, y)$ , where  $p$  stands for the attribute “population number”,  $m$  for the municipality, and  $y$  for the year, then the lookup subtasks involved in this task could be encoded as

$$\text{?}y: p(\text{Loures}, y) \in (300\,000, \infty)$$

$$\text{?}m: p(m, 1981) \in (300\,000, \infty)$$

The targets of these two lookup tasks are referrers with different value domains; moreover, these domains consist of elements of different nature. No relations can be defined for elements taken from such sets. That is why comparison tasks are only meaningful if the targets of the lookup tasks involved are components with coincident or at least overlapping value domains.

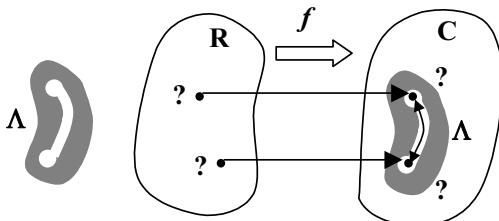
### 3.3.2 Relation-Seeking

We have defined relation-seeking tasks as tasks that imply a search for occurrences of specified relations between characteristics or between references. Here are some example tasks of this type:

- On what days did the stock price increase by more than 20% in comparison with the previous day?
- Did any earthquakes happen within 48 hours before a given earthquake?
- In which municipalities in Portugal did the population decrease from 1981 to 1991?
- In which states of the USA and in what years did the motor vehicle theft rate exceed the burglary rate?
- Find pairs of earthquakes such that the time interval between them is no more than 48 hours and the distance between their epicentres is no more than 50 km.

In an attempt to illustrate such a task graphically, we arrived at the idea of a metaphorical representation of a specified relation by something like a stencil, or mask. This stencil is to be moved over a set to find elements that fit in its holes and, hence, are related in the way specified. This metaphor is presented in Fig. 3.6. The shape on the left depicts a stencil, which represents symbolically a specified relation, denoted by  $\Lambda$ . This is assumed to be a binary relation, i.e. a relation between two elements. The goal is to find pairs of elements of the characteristic set  $C$  linked by the relation  $\Lambda$ .

Then, the references corresponding to these elements need to be determined.



**Fig. 3.6.** A graphical illustration of a relation-seeking task. A specified relation between characteristics, denoted by  $\Lambda$ , is represented as a stencil, or mask. The stencil is moved over the set of characteristics  $C$  until some elements of this set and the relation between them fit in the holes of the mask. The ultimate goal is to find what references correspond to the characteristics linked by the relation

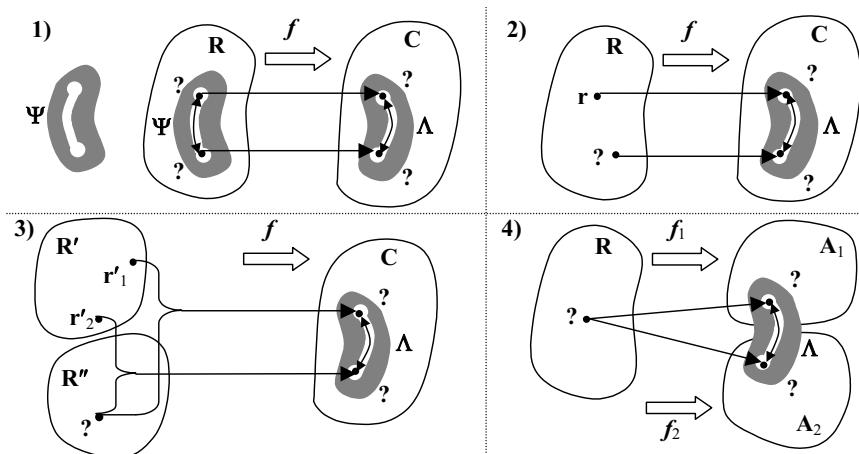
The diagram in Fig. 3.6 represents the fundamental idea of a relation-seeking task: find references such that the corresponding characteristics are related in a specific way. In this picture, only the required relation between characteristics is specified, and no other constraints are given. This situation can be represented by the formula

$$?y_1, y_2, x_1, x_2: f(x_1) = y_1; f(x_2) = y_2; y_1 \Lambda y_2 \quad (3.16)$$

Such a situation, however, rarely occurs in real relation-seeking tasks; at least, we have failed to find a more or less realistic example. The problem is that a task constructed faithfully according to this formula is perceived as underconstrained: there are four unknown items and only three constraints. Therefore, actual relation-seeking tasks typically include some additional constraints. In many cases, these are constraints concerning the references that correspond to the characteristics linked by the specified relation  $\Lambda$ . Four different ways in which the constraints on the references can be specified are represented graphically in Fig. 3.7.

*Case 1* represents the situation where the references are constrained by specifying a relation that must exist between them. Hence, the task constraints include two different relations: one relation that is expected to link characteristics and another relation that must link the corresponding references. The former relation is denoted by  $\Lambda$  and the latter relation is denoted by  $\Psi$ . The following formula corresponds to this case:

$$?y_1, y_2, x_1, x_2: f(x_1) = y_1; f(x_2) = y_2; x_1 \Psi x_2; y_1 \Lambda y_2 \quad (3.17)$$



**Fig. 3.7.** Different variants of relation-seeking tasks. (1) Not only a relation  $\Lambda$  that must exist between characteristics is specified, but also a relation  $\Psi$  that links the references corresponding to these characteristics. (2) Not only a relation  $\Lambda$  but also a reference, denoted by  $r$ , is specified in the task constraints. The goal is to find another reference such that the relation  $\Lambda$  exists between the characteristics corresponding to these two references. (3) In a dataset with two referrers, denoted by  $\mathbf{R}'$  and  $\mathbf{R}''$ , two values of one of the referrers ( $\mathbf{R}'$ ) are specified; they are denoted by  $\mathbf{r}'_1$  and  $\mathbf{r}'_2$ . The goal is to find a value of the other referrer ( $\mathbf{R}''$ ) such that the specified relation  $\Lambda$  exists between the characteristics corresponding to the combinations of this value with  $\mathbf{r}'_1$  and  $\mathbf{r}'_2$ . (4) A specified relation  $\Lambda$  is assumed to exist between the values of two different attributes corresponding to one and the same reference. The sets of values of the two attributes are denoted by  $\mathbf{A}_1$  and  $\mathbf{A}_2$

An appropriate example of this subtype of relation-seeking task is “On what days did the stock price increase by more than 20% in comparison with the previous day?” Here, the reference set is a linearly ordered set of days, and the characteristic set is the set of prices attained by the stock. Two binary relations are specified:

- a relation between characteristics, i.e. stock prices: One price must be higher than another by more than 20%;
- a relation between the corresponding references, i.e. days: The day corresponding to the higher price must immediately follow the day corresponding to the lower price.

*Case 2* is a task in which not only is a relation  $\Lambda$  between characteristics specified but also one of the references, which is denoted by  $r$ . The goal is to find another reference such that the relation  $\Lambda$  exists between the char-

acteristics corresponding to these two references. The formula for this task subtype is

$$?y_1, y_2, x: f(\mathbf{r}) = y_1; f(x_2) = y_2; y_1 \Lambda y_2 \quad (3.18)$$

An example of such a task is “Did any earthquakes happen within 48 hours before the given earthquake?” Here, the reference set is the set of earthquakes, and one element of this set, i.e. a particular earthquake, is specified. The characteristic set consists of the times of earthquake occurrences. The goal is to find earthquakes such that their occurrence times are related to the time of the specified earthquake (let this be  $t$ ) in the following way: these times are less than  $t$ , and the distances (in time) to  $t$  are no more than 48 hours.

*Case 3* illustrates the situation where a dataset has multiple (two in Fig. 3.7) referrers, and the reference set consists of combinations of values of these referrers. In a relation-seeking task, the values of some referrers may be specified while the values of the remaining referrers need to be found. This is the case for the example task “In which municipalities in Portugal did the population decrease from 1981 to 1991?” Here, one of the two referrers is time (more precisely, the set of census years), and the other one is the set of municipalities in Portugal. The characteristic set consists of the values of the attribute “population number”. For the temporal referrer, two values are specified: the years 1981 and 1991. The goal is to find all values of the other referrer such that the values of the population number corresponding to the combinations of those values of the referrer with the years 1981 and 1991 are linked by the following relation: the value corresponding to the first combination is higher than the value corresponding to the second combination. This task subtype can be encoded in the formula

$$?y_1, y_2, x: f(\mathbf{r}_1, x) = y_1; f(\mathbf{r}_2, x) = y_2; y_1 \Lambda y_2 \quad (3.19)$$

*Case 4* is the case where a specified relation is supposed to exist between values of different attributes corresponding to one and the same reference, which needs to be found. As we discussed earlier, a dataset with multiple attributes may be viewed as having multiple data functions, one per attribute. In Fig. 3.7, there are two such data functions, denoted by  $f_1$  and  $f_2$ ; the corresponding attribute value sets are denoted by  $\mathbf{A}_1$  and  $\mathbf{A}_2$ . The specified binary relation  $\Lambda$  must link some element of the set  $\mathbf{A}_1$  with the element of the set  $\mathbf{A}_2$  corresponding to the same reference as the element of the set  $\mathbf{A}_1$ . The formula for this task subtype is

$$?y_1, y_2, x: f_1(x) = y_1; f_2(x) = y_2; y_1 \Lambda y_2 \quad (3.20)$$

A representative example of such a task is “In which states of the USA and in what years did the motor vehicle theft rate exceed the burglary rate?” Here, we have two comparable<sup>1</sup> attributes: motor vehicle theft rate and burglary rate. The reference set consists of combinations of values of two referrers: states of the USA and years. The goal is to find combinations of a state and a year such that the following relation between the corresponding values of the two attributes exists: the motor vehicle theft rate exceeds the burglary rate.

We have now considered all the example tasks given at the beginning of this subsection except for the last one, “Find pairs of earthquakes such that the time interval between them is no more than 48 hours and the distance between their epicentres is no more than 50 km”. In this example, the reference set is the set of earthquakes. As in case 4 above, two attributes are involved in the task: the time of earthquake occurrence and the spatial location of the earthquake epicentre. Unlike case 4, two relations are specified as task constraints, and each relation is meant to exist between two values of one and the same attribute rather than between values of two different attributes. For the time of earthquake occurrence, the relation that we seek is “the temporal distance is no more than 48 hours”. For the epicentre locations, the relation must be “the spatial distance is no more than 50 km”. The goal is to find pairs of references such that both relations exist between the corresponding occurrence times and epicentre locations. A graphical illustration of this case is given in Fig. 3.8, and the corresponding formula (quite long) is given below:

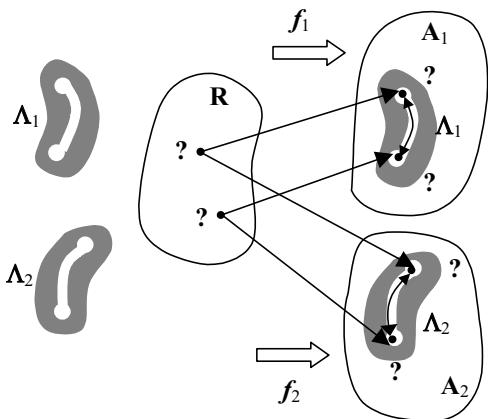
$$\begin{aligned} &?y_1, y_2, z_1, z_2, x_1, x_2: \\ &\mathbf{f}_1(x_1) = y_1; \mathbf{f}_1(x_2) = y_2; y_1 \boldsymbol{\Lambda}_1 y_2; \mathbf{f}_2(x_1) = z_1; \mathbf{f}_2(x_2) = z_2; z_1 \boldsymbol{\Lambda}_2 z_2 \end{aligned} \quad (3.21)$$

In the five variants of relation-seeking tasks that we have discussed, the basic scheme represented in Fig. 3.6 and encoded in (3.16) is modified by the introduction of additional constraints and/or by decreasing the number of unknown items.

We introduced relation-seeking tasks as inverse with respect to comparison tasks. From another viewpoint, relation-seeking tasks can be viewed as a kind of “hybrid” between lookup and comparison tasks. The targets of these tasks are references, as in inverse lookup tasks. However, the constraints are not defined by indicating particular characteristics or subsets of characteristics but rather by specifying relations that must hold between characteristics corresponding to the references. Hence, the tasks

---

<sup>1</sup> It is important that the attributes are comparable, i.e. their value sets are the same or at least overlap.



**Fig. 3.8.** A relation-seeking task may set constraints on relations that must exist between values of several attributes. Here, there are two different attributes  $f_1$  and  $f_2$ , with value sets denoted by  $A_1$  and  $A_2$ . Two relations are specified:  $\Lambda_1$ , which is expected to exist between elements of  $A_1$ , and  $\Lambda_2$  expected to exist between elements of  $A_2$ . The goal is to find two references such that the corresponding values of the attribute  $f_1$  are linked by the relation  $\Lambda_1$  and, simultaneously, the corresponding values of the attribute  $f_2$  are linked by the relation  $\Lambda_2$

involve comparison of characteristics in order to check whether the specified relation is valid.

Generally, both comparison and relation-seeking tasks are *compound* rather than atomic: any such task is built up from smaller operations, or subtasks. A compound task may be recognised from the presence of several items in its target. Thus, there are from two to six variables in the target parts of the formulas (3.11)–(3.21), which encode various types of comparison and relation-seeking tasks. The plurality of targets is also visible from the graphical illustrations of these tasks: each of them contains several question marks, which signify the unknowns.

The subtasks of compound elementary tasks may be of the following kinds:

1. For a reference, find the corresponding characteristic:  $?y: f(r) = y$ . We have called such tasks “direct lookup tasks”.
2. For a characteristic or a subset of characteristics, find the corresponding reference(s):  $?x: f(x) = c$ . Such tasks have been called “inverse lookup tasks”.
3. Compare two or more elements (either characteristics or references); that is, identify the relation existing between these elements:  $?λ: p λ q$ .

4. For a given element (either a characteristic or a reference), find other element(s) such that a specific relation exists between these elements:  
 $?p: p \Delta q$ .

In the expressions that encode these basic subtasks, bold letters are used to denote the items that are specified, i.e. constants, and italic letters (except for  $f$ ) denote the items to be found, i.e. unknowns, or variables.

Comparison tasks are built up from the basic subtasks 1 and 3 (direct comparison) or 2 and 3 (inverse comparison). A relation-seeking task implies the following operations:

- for a reference, find the corresponding characteristic (subtask 1 – direct lookup);
- for the characteristic thus found, find other characteristics linked to it by a specified relation (subtask 4);
- for the related characteristics thus found, find the corresponding references (subtask 2 – inverse lookup).

For subtask 1, the reference may be specified explicitly, or it may be required that all operations are performed repeatedly for multiple elements of the reference set. In the latter case, some additional constraints are usually specified, which limit the number of repetitions required and direct the search process through the reference set.

In general, it is possible to use these four basic subtasks to build a wide variety of compound elementary tasks, and not only the tasks we have already described and represented by formulae. The general principle of construction is to replace constants in some basic task by variables. Every such variable must be included in the task target, and some constraint(s) involving this variable must be added. Thus, in all but one of the task formulae that we have introduced so far, there are as many expressions in the constraint parts as items in the target parts. The only formula that does not comply with this rule is (3.16), which has four targets and only three constraints. Because of this disagreement, the formula (3.16) appears underconstrained, and it is hard to find a realistic example of a task that would fully correspond to this formula.

A strict general requirement for any compound task is that the data function  $f$  must appear in at least one of its constraints.

### 3.3.3 Recap: Elementary Tasks

We have introduced the notion of the *data function* and, correspondingly, the formula  $f(x) = y$  as a model of a dataset. We do not do this because of

our love of formality and obsession with encoding everything in formulae. The rationale for using the formal notation is to build a well-grounded, distinct, consistent, complete task typology, which means:

1. It must be clear where each task type comes from and why it is introduced.
2. It must be clear how one task type differs from another.
3. A common approach is used to define all task types.
4. There is a way to confirm that all potential data analysis tasks have been taken into account.

Let us explain why we rely upon the functional representation of a dataset in trying to reach our objectives.

To study an object or phenomenon comprehensively, a researcher needs to observe it from different perspectives, manipulate it, and do various experiments. When this is impossible (for example, the object is too big or too small, cannot be reached, or could be damaged), the researcher creates a model of the object and manipulates the model in order to learn the properties of the object. In our case, we could not achieve completeness by reviewing all datasets that have ever existed and collecting together all imaginable questions: both the data and the questions are incalculable. Therefore, we have created a model that could represent all possible datasets for us and now we are trying to manipulate this model in order to reveal all possible tasks. We believe that our formal model has now allowed us to enumerate all essential varieties of elementary tasks.

It must be borne in mind that a model is not identical to the object that it represents. It cannot have exactly the same properties; otherwise, the researcher could not use it, for the same reasons as the original object. A model should incorporate the most essential features of the object, and the choice of these features is made rather subjectively. Therefore, the researcher must be cautious when extending conclusions obtained from studying the model to the real object. It is good if some ways of checking these conclusions can be found. However, this is not always possible. In many cases, researchers have to be satisfied with the fact that a model seems to conform to reality and that the conclusions obtained do not contradict the observations. New observations may invalidate the conclusions, and the researchers will then have to amend the model or to create a new one.

Unfortunately, we cannot check whether our model really covers all possible elementary tasks. Thus far, it conforms to our observations obtained from our work with various data. Nevertheless, it may happen that new observations will necessitate revision of the model.

So, according to our model, a dataset consists of a set of references, a set of characteristics, and a data function, which defines the correspondence between the references and the characteristics. According to this model, three groups of relations between elements exist:

1. Relations between references.
2. Relations between characteristics.
3. Relations between references and characteristics.

While the relations between references and characteristics are defined by the data function, the relations within the set of references and within the set of characteristics depend on the nature and properties of these sets and may be diverse.

Elementary tasks are questions concerning relations between elements. The following basic questions are possible:

1. Given two (or more) elements, identify what relation exists between them.
2. Given an element and a relation, find other elements related in the specified way to the given element.

In principle, it is possible to formulate a task of the following form: given a relation, find elements linked by this relation. However, this formulation can be transformed into task 2 above, which in this case is repeated for every element.

Any task can be viewed as a combination of a target and one or more constraints. The target may include several items, which makes it also possible to speak of several targets of a task. The targets indicate the unknown information, which needs to be found. The constraints describe what is known. In our formal notation, constraints are expressions containing two parts linked by some relation, such as  $=$  (equals),  $\in$  (set membership), and  $<$  (less than).

We assume that all relations between references and characteristics are defined by the data function, or, at least, all such relations that are of interest to a data analyst. Hence, task 1 does not arise when we focus on relations between references and characteristics: there is no sense in asking what kind of relation exists between a given reference and a given characteristic. The only meaningful form of question is to find an element related to a given element according to the data function. There are two possibilities:

- for a given reference, find the corresponding characteristic;
- for a given characteristic, find the corresponding reference.

We have called the first possibility “direct lookup”, and the second possibility “inverse lookup”.

For relations between references and between characteristics, both task 1 and task 2 make sense. We call task 1 “comparison” and task 2 “relation-seeking”. However, comparison and relation-seeking tasks do not occur in data analysis in their basic form, because this form does not involve the data function. In other words, no data are needed for answering such questions, because, as we have mentioned, relations between references and between characteristics are determined by the general, invariant properties of the respective sets.

In data analysis, comparison and relation-seeking tasks appear as the basic forms modified by introducing additional targets and additional constraints, so that the data function is involved in at least one constraint. Since such tasks have several targets, they are no longer basic: the presence of several targets indicates that a task is compound, i.e. built from simpler operations.

We would like to point out that, when we speak about elementary tasks, we are not using the term “elementary” as a synonym for “simple” or “easy”. While lookup tasks might be characterised as simple, inverse comparison and relation-seeking tasks, which are also elementary, are typically compound. One of our examples of a relation-seeking task contained as many as six targets. Hence, we use the word “elementary” purely in the sense of addressing elements, i.e. individual items, rather than sets.

We have talked much about various relations between elements, but it may be noted that we have thus far considered only binary relations, i.e. relations involving two items. Relations in which more than two items participate exist as well and could be included in the suggested framework: for example, the relation “between” for an ordered set of items. However, in most cases, relations with more than two participants can be represented by collections of binary relations.

Let us now summarise the classes of elementary tasks that we have considered:

- *Lookup tasks*. In these tasks, it is necessary to find the values of some data components that correspond to given values of other data components according to the data function. Lookup tasks may be subdivided into direct and inverse lookup tasks.
  - *Direct lookup tasks*. The values of referential components are specified; the goal is to find the corresponding characteristics (values of attributes). For example, “On a given date, what is the price of stock X?”

- *Inverse lookup tasks.* The values of attributes are specified; the goal is to find the corresponding values of referrers. For example, “For a given price, on what date(s) was it attained?” If the dataset contains two or more referential components, there may be inverse lookup tasks with partly specified references (i.e. the values of some referrers are specified), where the goal is to reconstruct the complete references (i.e. to determine the unknown values of the remaining referrers). Here is an example of a case with two referrers: “Find municipalities that had 300 000 or more inhabitants in 1991”.
- *Comparison tasks.* In these tasks, the goal is to determine what relations exist between characteristics or references. At least one of the items to be compared is not specified explicitly, but must result from some lookup task. Comparison implies not only identification of the sort of relation but also, whenever possible, its numerical expression (a measure of the degree of relatedness). For numeric characterisation of a relation, distances between elements of the set of characteristics or of the set of references may be used, if such distances exist. For sets with the ratio level of measurement, ratios between elements may also be used. Depending on the type of the lookup tasks involved, comparison tasks may also be subdivided into direct and inverse comparison tasks.
  - *Direct comparison tasks* imply determining relations between elements or subsets of the set of characteristics (i.e. values of attributes). At least one of the characteristics to be compared results from a direct lookup task. Here are some examples of different variants of direct comparison tasks:
    - On a given date, did the stock price exceed €1000?
    - Compare the stock prices on the first and the last day of the week.
    - Compare the total values of the imports and exports of the given country.
    - Compare the immigration to the USA with the emigration from Mexico.
  - *Inverse comparison tasks.* In these tasks, relations between references must be determined. At least one of the references must result from an inverse lookup task. Here are some examples:
    - Did the stock price reach €1000 before or after the given date?
    - Compare the dates on which the prices €1000 and €1100 were attained.
    - Compare the location with the highest air temperature with the location with the highest humidity.

If a dataset contains two or more referential components, partial references may be involved in all three variants of comparison tasks. This means that the values of some referrers may be specified in the task constraints while the values of the other referrers may need to be determined and then compared.

- *Relation-seeking tasks.* In these tasks, the goal is to find references or pairs (or, in a more general case, groups) of references such that specified relations exist between the corresponding characteristics. Such a task consists of the following basic operations: (1) for a reference, find the corresponding characteristic (direct lookup); (2) for the characteristic thus found, find other characteristics linked to it by a specified relation; (3) for the related characteristics thus found, find the corresponding references, i.e. perform inverse lookup tasks. For the operation (1), the reference may be specified explicitly, or it may be required that this operation and the following operations are performed for multiple elements of the reference set. In the latter case, some additional constraints are usually specified, which limit the set of references to be involved and direct the search through the reference set. Here are some examples:
  - “On what days was the stock price higher than on the given day?” Here we have a case where one reference (a specific day) is explicitly specified.
  - “Find countries where the imports exceed the exports”. Here, there is an additional constraint that the values of the two different attributes must correspond to one and the same reference.
  - “Where did the population decrease from 1981 to 1991?” In this example, references consist of two components, space and time. The values of the temporal components are specified, and the goal is to find value(s) of the spatial component such that the characteristics corresponding to the full references are related in the specified way.
  - “On what dates did the price of the stock decrease in comparison with the previous date?” This task requires a search for pairs of references with a specified relation between the corresponding characteristics and, additionally, another specified relation between the references themselves.

It should be noted that elementary tasks do not play a primary role in exploratory data analysis. The goal of EDA is to discover inherent properties of a dataset as a whole. This cannot be achieved by performing only elementary tasks, which focus on individual data items and do not imply an overall view of the dataset. Nevertheless, elementary tasks cannot be ignored, since they necessarily emerge in EDA and hence require adequate

tools to support them. For example, an analyst may explore the spatial variation of a numeric attribute and notice that the value in some location looks very different from those in the neighbourhood. Naturally, the analyst would like to know what this value is, how much it differs from the values around it, and what the values of other attributes in this location are. All these are elementary tasks, which may contribute significantly to the overall understanding of the spatial phenomenon under analysis. However, in order to see the wood for the trees, one needs a higher level of abstraction than is supposed in elementary tasks.

## 3.4 Synoptic Tasks

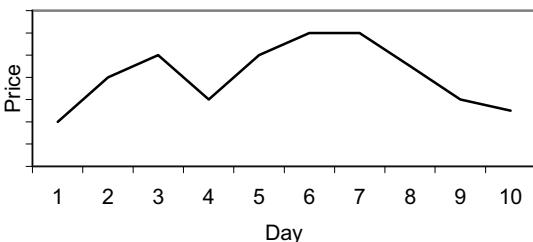
### 3.4.1 General Notes

The term “synoptic task” is the result of rather long search for a suitable name for the class of tasks that require one to deal with sets as a whole, in contrast to elementary tasks dealing with individual elements. At the beginning of the story was our uncomfortable feeling concerning Bertin’s intermediate and overall levels of reading. While it was completely clear that these two categories are different in principle from the category referred to as the “elementary level”, we could not see significant differences between those two categories themselves. According to Bertin’s definition, the only difference is that the overall level involves the whole set of possible values of some component, whereas the intermediate level refers to subsets of this set. Hence, in both cases sets are involved rather than individual elements, and the difference is only in the size of the set. In our opinion, this difference is not as important as the difference between consideration of elements and consideration of sets, and, consequently, it is not sufficient to justify the existence of two distinct categories. Besides, we have learned from our practical experience that different methods and tools are needed to support analysis on the level of elements and on the level of sets, but the same methods and tools can be used for the intermediate- and overall-level tasks. Compare, for example, the following questions:

- In the first three days, what was the trend of the price?
- During the entire period, what was the trend of the price?

These are Bertin’s examples of tasks pertaining to the intermediate and the overall level, respectively.

To find answers to these questions, one could represent the dynamics of the price as a line on a time graph, as is shown in Fig. 3.9. The first question requires one to examine the shape of the line segment corresponding to the first three days. To answer the second question, the shape of the entire line needs to be examined. Both the tool used (i.e. the time graph) and the analysis procedure (i.e. observing the shape of the line) are the same in both cases. The only difference is that for the first question one needs to delimit the relevant subset, i.e. time interval. For this delimitation, some additional tool could be used. For example, most of the current software tools for data visualisation provide opportunities for zooming and focusing on subsets of the data. However, these functions play only an auxiliary role in data analysis. Thus, for understanding the trend of the price, seeing the shape of the line is more important than the possibility of zooming into the first three-day interval.



**Fig. 3.9.** A time graph represents the dynamics of a numeric attribute over a time period. This is a suitable tool for finding answers to questions concerning the trend both during the whole period and during its subintervals

This example demonstrates that intermediate- and overall-levels have many more commonalities than differences. Therefore, our idea was to unite these two categories into a single category. The problem was to find a suitable name, which would reflect the major feature of this class of tasks: they pertain to sets, and the sets need to be treated as wholes. The options we considered were “set-related tasks”, “high-level tasks”, and “general-level tasks”, all three rather awkward. Of these three, only the first provides any clue about the idea. The term “high-level” does not express anything. Although we used the term “general-level” in our earlier publications (for example, in Andrienko et al. (2003)), we find it too ambiguous. In fact, we discuss all tasks (including elementary tasks) on a general level, and the formula  $?y: f(r) = y$  is general even though it represents a certain group of elementary tasks.

Therefore, we were quite happy to find the term “synoptic”, which seems to be much closer than those considered before to the idea we want to express. The word “synoptic” is defined in a dictionary as “pertaining to or constituting a synopsis; affording or taking a general view of the principal parts of a subject” (Random House 1996). The word “synopsis” is defined in the same source as “a brief or condensed statement giving a general view of some subject” or “a compendium of heads or short paragraphs giving a view of the whole”. This corresponds well, for example, to the task “In the first three days, what was the trend of the price?” A suitable answer would be “the price increased”. This is “a brief or condensed statement giving a general view of some subject”, i.e. a synopsis. The question “During the entire period, what was the trend of the price?” also requires a sort of synopsis rather than an enumeration of the prices for all days. Thus, one could describe the trend as “an increase, then a sudden drop followed by an increase to a higher level than before, and then a gradual decrease”. In so doing, one takes “a general view of the principal parts of a subject”: the overall trend is divided into a few principal parts characterised as an increase, a sudden drop, and a gradual decrease.

Hence, a task dealing with a set as a whole implies making a sort of synopsis concerning this set, and therefore can be called a “synoptic task”. A synopsis is not necessarily verbal; in some cases data may be summarised numerically, graphically or in the form of equations, for example.

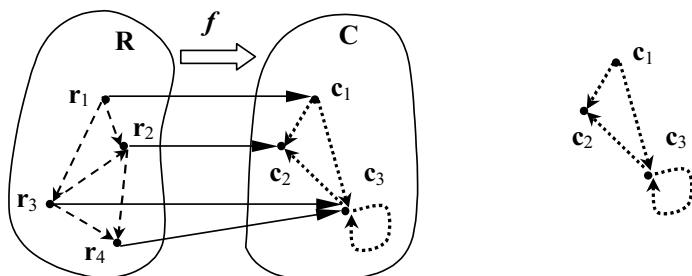
### 3.4.2 Behaviour and Pattern

In a synoptic task, a data analyst deals with a set of references as a whole, i.e. considers simultaneously all its elements, as well as the system of relations existing between these elements. For each of these elements, there is a corresponding characteristic, i.e. an element of the characteristic set. This correspondence is defined by the data function. The characteristics corresponding to the references form a *configuration* with respect to the set of references and the system of relations between the references.

The simplest example of such a configuration is a sequence of attribute values corresponding to the system of ordering relations in a linearly ordered set of references: a sequence of stock prices over a period of time, a sequence of flow speed measurements along a river, a sequence of phases in the development of an insect, etc. Another example of a configuration of characteristics corresponding to a system of relations within a reference set is a distribution of characteristics over a two-dimensional or three-dimensional space: the distribution of the population density over the territory of Portugal, the distribution of various forest structures over Europe,

the distribution of the stream direction, stream speed, and temperature over an ocean at different depths, etc.

Such configurations of characteristics, which are determined by relations between references and by the data function, are called *behaviours*. The notion of a behaviour can be illustrated graphically as shown in Fig. 3.10.



**Fig. 3.10.** A graphical illustration of the notion of a behaviour. The data function  $f$  associates each element of the reference set  $\mathbf{R}$  with the corresponding characteristic, i.e. element of the characteristic set  $\mathbf{C}$ . The references are linked by a system of relations, which are symbolised by the dashed arrows. A certain configuration, or arrangement, of the characteristics corresponds to this system of relations between the references. This configuration, which is symbolised by the dotted arrows, is the behaviour of the data function over the reference set  $\mathbf{R}$

In fact, “behaviour of a data function” is a metaphoric term. A dataset represents a phenomenon, and it is the phenomenon that “behaves”, i.e. exhibits particular characteristics under various conditions. The behaviour of a data function is a reflection of the behaviour of the underlying phenomenon. An analyst explores the behaviour of the data function in order to understand the behaviour of the phenomenon. Synoptic tasks are tasks dealing with behaviours of data functions and hence of the underlying phenomena.

To describe synoptic tasks, we shall use the following formal notation to denote the behaviour of a data function  $f$  over a reference set  $\mathbf{R}$ :

$$\beta(f(x) \mid x \in \mathbf{R}) \quad (3.22)$$

Analogously, the behaviour of the function  $f$  over any non-empty set of references  $\mathbf{R}'$  (a subset of  $\mathbf{R}$ ) is denoted as  $\beta(f(x) \mid x \in \mathbf{R}')$ . The reference set  $\mathbf{R}'$  may be called the *base* of this behaviour. We may also say that the behaviour is *based* on the set  $\mathbf{R}'$ .

For the sake of generality, we can allow the set  $\mathbf{R}'$  to contain any number of elements, in particular, one element. Such a case may be called the

“local behaviour”. The notion of the local behaviour of a function is known, for example, in mathematical analysis, where a derivative characterises “the instantaneous rate of change of one quantity in a function with respect to another” (Random House 1996). Here are some examples of local behaviours:

- On day 3, there is a peak in the stock price.
- Lisbon has the highest population density among the municipalities of Portugal.

The outcome of studying the behaviour of a function is some conception of this behaviour, some mental construct or an externalised representation of it (a description of the behaviour) that incorporates as much significant information from the data as possible, in the simplest and shortest possible way. According to Bertin, the understanding of data means “discovering combinational elements which are less numerous than the initial elements yet capable of describing all the information in a simpler form” (Bertin 1967/1983, p. 166). We shall use the term *pattern* to denote such “combinational elements”, i.e. distinctive features of a behaviour, reflected in the explorer’s mind and/or expressed in a descriptive representation.

So, we define a *pattern* as a construct reflecting essential features of a behaviour in a parsimonious manner, i.e. in a substantially shorter and simpler way than specifying every reference and the corresponding characteristics. The construct may be a description in some language (natural, formal, or graphical) or a mental image of the behaviour.

Our usage of the term “pattern” is consistent with its definition as “a combination of qualities, acts, tendencies, etc., forming a consistent or characteristic arrangement” (Random House 1996). The implied meaning is also similar to what is understood by a pattern in data mining: “a pattern is an expression  $E$  in some language  $L$  describing facts in a subset  $F_E$  of a set of facts  $F$  [i.e. a dataset, in our terms] so that  $E$  is simpler than the enumeration of all facts in  $F_E$ ” (Fayyad et al. 1996). In other words, a pattern is a parsimonious description of a subset of data. Our use of the term “pattern” is slightly broader: it is not necessarily a description but may also be a mental construct which has not yet been externalised in any language.

In data mining and statistics, the notion of a pattern is distinguished from that of a *model*: while a pattern describes a subset of data, a model is a description of the entire dataset. For our purposes, this distinction is irrelevant. We shall use the same term “pattern” irrespective of whether an entire set or a subset is considered.

Let us define a *behaviour characterisation task* as a task of revealing distinctive features of the behaviour of some phenomenon and representing them by an appropriate pattern. To denote the relation between the pattern

resulting from such a task and the behaviour under analysis, we shall use the expression “the pattern *approximates* the behaviour”, where the word “approximates” stands for “describes”, “characterises”, “summarises”, “represents”, “reflects”, etc.

Like elementary tasks, synoptic tasks also comprise targets and constraints. The target of a behaviour characterisation task is a pattern approximating the behaviour, and the constraints consist of the function (i.e. dataset) the behaviour of which is being studied, and the set of references on which the behaviour is based. We shall use the following notation for behaviour characterisation tasks:

$$\text{?}p: \beta(f(x) \mid x \in \mathbf{R}) \approx p \quad (3.23)$$

This should be read in the following way: “find a pattern  $p$  approximating the behaviour of the function  $f(x)$  over the reference set  $\mathbf{R}$ ”. We do not imply any mathematical meaning behind the symbol “ $\approx$ ”. In our notation, it means that the pattern approximates the behaviour.

We shall use the capital letter  $\mathbf{P}$  to denote a particular pattern, while  $p$  is a variable standing for an unknown pattern. Sometimes, we shall use the notation  $\mathbf{P}(\mathbf{R})$  in order to emphasise that the pattern  $\mathbf{P}$  is defined for the reference set  $\mathbf{R}$ .

The meaning of the notion of a pattern may be specialised depending on the nature and properties of the reference set considered. Thus, when the data refer to time, a typical notion is a “trend”, which can be viewed as a specialisation of the more general notion of a pattern. When spatially referenced data are explored, one looks for patterns in the spatial distribution. When the data refer to a population (i.e. a group of objects), a pattern may take the form of a statistical summary of the distribution of attribute values across the population. Table 3.2 contains examples of behaviour characterisation tasks formulated for various types of referrers and attributes.

We would like to point out the difference between the meanings of the terms “behaviour” and “pattern” as we use them. We understand a behaviour as something inherent in a phenomenon and existing objectively, independently of an observer or an analyst. A pattern, on the contrary, is something resulting from observation or analysis, an image or portrait of a behaviour that shows how the observer or analyst sees and understands it. Hence, a pattern is indispensably subjective. Different observers may understand the same behaviour differently and represent it by different patterns. Moreover, even one observer may use different patterns to describe the same behaviour depending on his/her goals. In other words, a behaviour characterisation task typically does not have a single answer. Various alternative patterns can substitute for the variable  $p$  in the expression (3.23).

**Table 3.2.** Examples of behaviour characterisation tasks for various types of referers and attributes

Attribute→ ↓Referrer	Nominal	Interval, ratio	Spatial
<b>Population</b>	Investigate the frequency distribution of the attribute values	Investigate the frequency distribution (this involves grouping the attribute values into intervals)  Find the minimum, maximum, median, mean, and mode	Investigate the spatial distribution (e.g. find concentrations, alignments, etc.)
<b>Time</b>	Investigate the frequency of changes, detect periodicity	Detect trends, investigate speed of change, detect periodicity	Investigate the direction and speed of movement or changes in size, shape, or orientation
<b>Space</b>	Investigate the patterns of the spatial distribution of the values	Detect spatial (directional) trends, and clusters of close values	Investigate the dependence of spatial characteristics of objects (size, shape, or orientation) upon location

Differences in patterns representing one and the same behaviour are, first of all, related to the desired degree of simplification. Let us imagine, for example, that the price of a stock increases steadily during a certain time period. In some cases, it is sufficient to describe this behaviour simply as “continuous growth”. In other cases, an analyst may need to take into account the speed of growth and describe the behaviour as “fast increase at the beginning, and then the speed of the growth gradually decreases”. Another possibility is to represent the behaviour by an appropriate mathematical formula. Here, “continuous growth”, “fast increase” etc., and the formula are three possible patterns representing the same behaviour. The first one is the simplest and least precise of them, while the formula is the most complex (at least for ordinary people) but definitely the most precise. Other possible variants of the simplification are to indicate the range of price variation (e.g. from €1000 to €1200) or to give just the average price, or both. It may be argued whether the word “pattern” is still applicable to such extremely simplified characterisations. However, for the sake of generality, we shall use this term for any general statement con-

cerning some behaviour, irrespective of whether it consists of a single number, an extensive description, or a formula.

There are no reasons to presume that a mathematical characterisation through a formula is superior to a verbal description or that a more precise pattern is always better. The trade-off between simplicity and precision depends on the goals of the observer. For example, annual data about air temperatures in July in Switzerland may be analysed by a person who is going to spend her/his vacation there and by a glaciologist studying the evolution of glaciers in the Alps. While for the person going on vacation it is sufficient to know the minimum, maximum, and average temperatures, the scientist would try to reveal a long-term trend in the July temperatures.

Let us return to the time series shown in Fig. 3.9. We have described its behaviour as “an increase, then a sudden drop followed by an increase to a higher level than before, and then a gradual decrease”. According to our definition, this description is a pattern approximating the behaviour of the stock price. The pattern is expressed through a verbal statement. If we look at this statement more closely, we find that it consists of four parts: (1) an increase; (2) a sudden drop; (3) an increase to a higher level than before; (4) a gradual decrease. Each of these smaller statements describes a part of the behaviour that is based on a certain subinterval. This example demonstrates a common approach to the exploration and characterisation of behaviours: if there is no simple pattern that can represent the entire behaviour (i.e. the behaviour over the whole reference set), the reference set is divided into subsets so that the behaviour over each subset can be represented by a sufficiently simple pattern. The entire behaviour is then represented by a combination of these patterns, with an indication of which pattern corresponds to which subset. We consider such a combination of patterns as a pattern also. In general, we assume that a pattern may consist of other patterns (subpatterns), and, correspondingly, an analyst may decompose a behaviour in order to represent it by sufficiently simple patterns.

A compound pattern consisting of several subpatterns may be formally represented as follows:

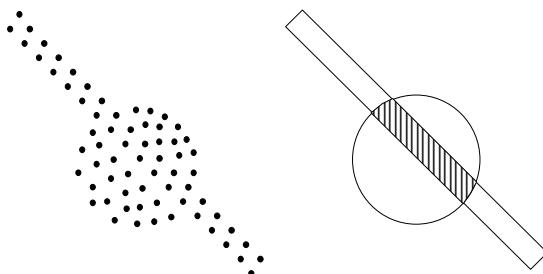
$$\mathbf{P}(\mathbf{R}) = \mathbf{P}_1(\mathbf{R}_1) \oplus \mathbf{P}_2(\mathbf{R}_2) \oplus \dots \oplus \mathbf{P}_k(\mathbf{R}_k) \quad (3.24)$$

Here,  $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_k$  are subpatterns of the compound pattern  $\mathbf{P}$ , and the symbol  $\oplus$  means pattern combination; like “ $\approx$ ”, if has no mathematical connotation. The subpatterns are defined on the reference sets  $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_k$ , respectively. Each of the sets  $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_k$  is a subset of the reference set  $\mathbf{R}$ , i.e.  $\mathbf{R}_1 \subset \mathbf{R}, \mathbf{R}_2 \subset \mathbf{R}, \dots, \mathbf{R}_k \subset \mathbf{R}$ . The union of these subsets makes  $\mathbf{R}$ :  $\mathbf{R}_1 \cup \mathbf{R}_2 \cup \dots \cup \mathbf{R}_k = \mathbf{R}$ .

A few remarks should be made concerning the process of decomposing a reference set into subsets in order to define subpatterns.

1. Dividing a set into subsets should never go down to the level of individual elements. Otherwise, no simplification will be achieved: the resulting characterisation will not contain a smaller number of items than that in the original data.
2. While a population-type reference set (i.e. a set without ordering and distances) can be divided into quite arbitrary subsets, partitioning of a reference set with ordering and/or distances is typically done with proper regard for the ordering and/or distance relations. Specifically, an ordered (fully or partially) set is usually divided into uninterrupted subsequences of elements, if it is a discrete set, or continuous subintervals, if it is a continuous set. Subsets of a set with distances are usually formed from neighbouring elements. Moreover, a continuous reference set is usually divided into continuous subsets.
3. Exceptions to the previous rule are often made when the reference set is heterogeneous, i.e. consists of qualitatively different parts. Such heterogeneity characterises, for example, geographical space in contrast to an abstract two-dimensional space: the former consists of land and water, mountains and plains, forests and arable land, etc. Another example could be a temporal referrer in some business-related dataset: vacation and holiday periods differ very much from other periods of a year, and weekends differ from workdays. An analyst may prefer to divide a heterogeneous reference set into subsets according to the character of the references rather than purely on the basis of ordering and/or distance relations. In this case, it is possible that a subset of references will consist of several disjoint parts.
4. It is not, in principle, required that the subpatterns cover the entire reference set. It should be remembered that an analyst tries to develop a synopsis, i.e. “a general view of the principal parts of a subject” (Random House 1996), and hence may deem some parts of the subject (i.e. the behaviour) not to be “principal parts”.
5. There is no particular reason to prohibit overlapping of the subsets that the subpatterns refer to. Let us consider, for example, the left part of Fig. 3.11, which represents the behaviour of some imaginary phenomenon in space, more specifically the spatial distribution of some point objects or events (the distribution has been artificially created for illustration purposes). Probably the simplest description of this behaviour would consist of two patterns: an alignment along the north-west–south-east diagonal and a round arrangement in the centre. These two patterns overlap, i.e. there is an area belonging to both of them. This area is

marked in the right part of Fig. 3.11 by hatching. All objects located in this area belong both to the diagonal alignment and to the round arrangement. Although the overlap may be viewed as an undesirable property, any attempt to get rid of it would result in much more complex patterns. Again, it depends on the goals of the analyst whether overlapping patterns are allowable or not.



**Fig. 3.11.** An artificial spatial distribution (left) and a possible pattern representing it (right)

To sum up our reasoning concerning behaviours and patterns, we understand a pattern as a synopsis of a behaviour, and, consequently, define synoptic tasks as tasks involving building, detecting, and comparing patterns representing the behaviours of phenomena over various reference sets. We have mentioned several properties of patterns:

- degree of simplification;
- precision (the property opposite to the degree of uncertainty that we encounter when we try to use a pattern to reproduce the data that it was derived from);
- coverage of the reference set (complete or partial);
- presence or absence of an overlap between subpatterns.

All other things being equal, the simplest possible patterns are typically preferred. This corresponds to the logical principle formulated by the medieval philosopher William of Occam (or Ockham) and known as “Occam’s razor” (see, for example, Heylighen (1997)). This principle states that one should not make more assumptions than the minimum needed. The principle is often called the principle of parsimony. It underlies all scientific modelling and building of theories. Moreover, it also underlies human visual perception: according to the central law of perception formulated by gestalt psychologists, people intuitively prefer the simplest, most stable of possible organisations (see, for example, Encarta (2004)).

### 3.4.3 Types of Patterns

We have considered thus far a few examples of behaviours and of patterns approximating them, such as the description of the time series shown in Fig. 3.9 and the geometrical shapes representing the artificial distribution in Fig. 3.11. We have also mentioned that a pattern may also be a mathematical formula, a range of attribute values, or just a single number such as the average temperature. Can we find something common in this variety and develop a more or less general notion of a pattern?

From these and other examples, we can extract the following basic variants of patterns:

1. *Association*: Perception or description of a (sub)set of references as a unified whole on the basis of similarity of their characteristics, i.e. close values of one or more attributes corresponding to these references.
2. *Differentiation*: Perception or description of some references or subsets of references as differing from others by their characteristics.
3. *Arrangement*: An idea or description of how characteristics are arranged, with respect to an ordering of references, for example a trend in characteristic that changes over time.
4. *Distribution summary*: A general idea or description of how characteristics are distributed over a reference set: how varied they are, what values occur most frequently, whether there are outliers (a few values greatly differing from the rest), etc.

Before considering these types of patterns in more detail, we would like to stress that we do not aspire to creating a full classification of possible patterns. Instead, we introduce and describe a few types, which are familiar to us from our own experience, mainly in order to make our concept of a pattern better understandable to readers.

#### 3.4.3.1 Association Patterns

An association pattern means that some references are unified into a whole and can be handled together. Such unification is typically done on the basis of identical or close characteristics, i.e. values of certain attributes, corresponding to these references. For example, a number of districts may be considered together as “a cluster of districts in the north of Portugal with high proportions of children”. In this example, the districts are united into a whole (a cluster) on the basis of their close characteristics in terms of the proportion of children in the population. Such an association could also be performed on the basis of multiple attributes, for example, “a cluster of

districts with high proportions of children and low proportions of elderly people”.

Another example of an association pattern is our representation of the artificial spatial distribution of points shown in Fig. 3.11. We have associated individual points into two geometric figures, first of all on the basis of the spatial proximity of these points, i.e. the closeness of the corresponding values of the attribute “spatial location”. An additional reason was that the figures were simple and readily identifiable in the visual representation of the points. As a result, we have built two new constructs, a “round cloud in the centre” and a “narrow diagonal belt”. Now we can use these new constructs in our further investigation of the phenomenon and our reasoning about it instead of the original multitude of points. Thus, we can say that the “belt” crosses the “round cloud” and that it is shifted towards the north-east with respect to the centre of the “cloud”. If we had other attributes related to the points, we could compare the characteristics of the “round cloud” and the “belt”, summarised from the characteristics of the individual points comprising these shapes.

In general, the formation of association patterns is typically accompanied by deriving summary characteristics of a union of references from the individual characteristics of its members. The easiest situation is when all the members have one and the same value of some attribute; this value becomes the characteristic of the union. In other cases, all different attribute values occurring among the characteristics of the members of the union may be listed (often with an indication of the frequencies or probabilities of their occurrence) or, for ordinal or numeric attributes, the range of variation of the values may be specified. Numeric characteristics are often summarised by means of computations that somehow aggregate individual characteristics into a single numeric value or a few values. The most common aggregate characteristic is the mean, or average, of the values of a numeric attribute. For example, “the average proportion of children in this cluster of districts is 25.2%”. It is very useful to compute, along with the mean, a measure of the variance of the values, which indicates how consistent the characteristics of the members of the union are.

Associations of references can also be characterised in a relative way, which involves comparison and differentiation between members of a union and the remaining elements or the whole reference set. For example, “the proportions of children in this cluster of districts are much higher than in the remaining part of the country” or “girls, on average, perform better in mathematics than boys”. Actually, comparison and differentiation are involved in the very process of building association patterns. Thus, a subset of references  $R'$  is considered as some unified whole if the elements of  $R'$  have characteristics closer to those of each other than to those of non-

members of the subset  $\mathbf{R}'$ . This can also be stated in other words: the variation of characteristics among elements of the subset  $\mathbf{R}'$  is considerably smaller than that in the entire reference set.

### **3.4.3.2 Differentiation Patterns**

As we have just said, differentiation is inherently involved in association: we unite some references not only because their characteristics are similar but also because they are different from the characteristics of other references. These “other references” may be all remaining elements of the reference set or the neighbourhood of the references that are united, if the reference set has ordering and/or distance relations. For example, an analyst may associate a number of districts in the north-west of Portugal with high proportions of children into a union (cluster). In other parts of the country, there may be a few other districts where the proportions of children are high, and hence close to those in the cluster. However, the analyst does not include these scattered districts into the north-western union, because of their spatial remoteness from the union. In this example, the analyst differentiates the districts with high proportions of children in the north-west from their neighbourhood, which has a lower proportion of children, rather than from all remaining districts, some of which have the same proportion of children as in the north-west.

As we have stated, association of references is typically done on the basis of close characteristics, although it involves the operation of differentiation. However, similarity of characteristics is not required for differentiation as such, and differentiation can be done even when no commonalities between references are observed: some elements or subsets of the reference set may simply be noted as having substantially different characteristics from the rest of the reference set or from the neighbourhood. In particular, an analyst may detect outliers – references with atypical characteristics, for example extraordinarily high or low values of a numeric attribute. Outliers may be “global” or “local”. Global outliers are references having atypical characteristics with respect to the whole reference set. For example, there are three districts in Portugal with population densities of 7913, 7455, and 7261 inhabitants per square km, while the densities in the remaining districts range between the values 7 and 3302. Local outliers are references that differ significantly in their characteristics from their neighbourhood but not necessarily from all other elements of the reference set. For example, the proportion of children in the population of Lisbon (14.22%) is much lower than in the surrounding districts (from 17.06% to 21.95%). The same is true for the city of Porto: 16.95% versus a range

from 20.12% to 27.50%. This does not mean, however, that the proportions of children in Lisbon and Porto are the lowest in the country.

It is not only cases where a single reference differs from all others or from its neighbourhood that may be characterised by differentiation patterns. It is also possible to differentiate a subset of references from the rest of the reference set owing to a much higher variability of characteristics in this subset than in other parts of the reference set. Such a subset may be considered as a unified whole on the basis of this extreme variability, in contrast to more coherent characteristics in other subsets. For example, an analyst may find continuous areas in Portugal with consistently high or low employment in agriculture, along with an area in the south where the districts differ very much in the values of this attribute. Another example would be a time interval where there is chaotic fluctuation of a stock price, while a more regular behaviour is observed in other intervals. This chaotic fluctuation could be characterised by a differentiation pattern, while the more regular behaviour in a time interval would typically be approximated by an arrangement pattern.

### **3.4.3.3 Arrangement Patterns**

An arrangement pattern is a perception of characteristics as being specifically ordered or organised when references are considered in a certain order. This applies first of all to “naturally” ordered reference sets, such as time. When an explorer considers characteristics corresponding to an ordered sequence of time moments, he/she can make observations such as the following:

- The values of this numeric attribute increase (or decrease) gradually.
- A gradual increase is followed by a sharp decrease.
- Gradual increases alternate with sharp drops.
- The higher the level to which the attribute increases, the deeper and more abrupt is the following drop.
- The values of this quantitative attribute appear repeatedly in the sequence  $v_1, v_2, \dots, v_k$ .
- The value  $v_n$  tends to be preserved for longer times than other values of the attribute.

Some of these types of observations are commonly called “trends” (in particular, an increasing or decreasing trend), while some others would be designated as “periodic patterns”. We suggest the term “arrangement pattern” as a more general appellation subsuming the notions of trend, periodicity, oscillation, stability, etc. – any idea concerning the sequence in

which characteristics appear when references are considered in a certain order.

Moreover, we do not limit the notion of an arrangement pattern only to linearly ordered reference sets. First, an arrangement pattern may be found with respect to an ordering arbitrarily introduced into a set of references. For example, one can put the municipalities of Portugal in order of increasing population density and observe an increase of the proportion of people employed in services and a decrease of the proportion of elderly people. Second, characteristics may be organised in a particular way with regard to a two-dimensional or three-dimensional arrangement of references. Think, for example, about the distribution of black and white squares on a chessboard. This is a highly regular arrangement of characteristics related to the arrangement of the reference set, which consists of 64 squares, in eight vertical and eight horizontal rows.

While it is common to talk about trends mostly in relation to some linear ordering of references, the phrase “spatial trend” is also widely used, although, as we have discussed earlier, there is no natural ordering between spatial locations. However, it is possible to consider various directions in space, which define certain orderings of locations. For example, one can characterise the situation with regard to crime in a city as “the crime rate increases from the centre of the city towards the periphery”. Here, a partial ordering has been introduced that arranges spatial locations according to their distance from the centre. For another spatial behaviour, a different ordering may be appropriate. For example, there may be an increasing trend from north to south or from north-west to south-east. An example of a non-numeric spatial trend could be a description such as “from the north to the south, deserts are gradually replaced by savannas, which then turn into tropical forests”.

Arrangements, and trends in particular, not only may be indicated verbally but also may be characterised by means of various numeric measures, such as the rate of change for numeric attributes or the frequency of change for qualitative attributes, the period length for periodic data, or the probability that attribute values appear in a specific order.

#### **3.4.3.4 Distribution Summary**

A distribution summary reflects the general manner in which the characteristics are distributed over a reference set. We have already mentioned such aspects as how varied the characteristics are, what values occur most frequently, and whether there are outliers.

An elaborate apparatus for summarising and otherwise characterising distributions is offered by statistics, which operates with such notions as

normal distributions, bimodal distributions, and skewed distribution. We have already mentioned the mean and variance as means of aggregation of numeric characteristics over a set of references. These measures can also be viewed as a summary of the value distribution. It is also possible to use other statistical measures, for example the median or various quantiles (percentiles). Thus, John Tukey suggested a method of summarising sets of numbers by computing their median and quartiles and representing them visually on “box-and-whiskers” plots (Tukey 1977). Handbooks on statistics describe many other possibilities for summarising numeric and non-numeric characteristics. An interesting method for summarising spatially referenced data is to compute the position of the “centre of gravity” of a spatial distribution. Thus, for the artificial distribution in Fig. 3.11, the centre of gravity would be located somewhere near the centre of the circular cloud. The idea of a centre of gravity could also be applied to a spatially referenced numeric attribute such as the population numbers for Portugal.

It is not only statistical methods that can be used to summarise distributions. Methods from the information theory can also be suitable. Thus, the main measure considered in information theory, entropy, can be used as an indicator of the heterogeneity of characteristics, for example for spatial differentiation.

We would not like to limit the notion of a distribution summary only to numerical measures or to the outcomes of computations. It is quite possible to summarise a distribution perceptually or verbally. For example, one can note that high employment in services occurs mostly along the coasts of Portugal and around big cities, and that the proportions of children in population are higher in the north of the country than in the south. As can be noted from our examples, such summarisations may involve division of a reference set into parts and association of elements on the basis of close characteristics, which results in a compound pattern consisting of several subpatterns.

### **3.4.3.5 General Notes**

There are no rules for selecting what type of pattern to look for in what situation. There are also no recipes for how to discover meaningful and useful patterns, i.e. patterns embodying some essential features of the behaviour being explored. Typically, it is not a big problem to find some common properties for a subset of references, but the problem is to find distinguishing properties, i.e. something that differentiates these references from the rest. However, even this is not enough. Thus, it is possible to make the statement ‘Districts in the west of Portugal are closer to the At-

lantic coast than are those in the east". However, this pattern is useless if the goal is to investigate the population structure or unemployment in the country. The same applies to the other types of patterns: they need to be distinctive and relevant to the goals of the investigation.

When making summaries of characteristics, it is important to keep the right level of aggregation in order to avoid worthless results such as the mean body temperature over the set of patients in a hospital. One should bear in mind the degree of variation of the characteristics pertaining to the reference set and check for the presence of outliers, i.e. a few references that differ very much from the rest in their characteristics.

Seeking to define trends is advisable first of all when the reference set is naturally ordered, for example in the case of time. In other cases, detecting a trend or other kind of arrangement may require consideration of various orderings. What ordering may be suitable in a particular case depends on the behaviour observed, but it happens very often that no ordering can be found that would allow one to detect any meaningful trend. Thus, the distribution shown in Fig. 3.11 can hardly be described as a spatial trend, but it can be described as a combination of two geometrical shapes, i.e. as a compound pattern consisting of two association patterns. Usually, it makes sense to look for possible trends when the reference set has distances and the characteristics change gradually, i.e. neighbouring elements differ less than more distant ones.

We have already discussed how reference subsets for defining subpatterns should be chosen. We would like to add a few words concerning the precision with which these reference subsets are specified. It is very common to describe them in an inexact, fuzzy manner rather than give a precise, unambiguous specification. For example, in the description "the proportion of children is higher in the north of Portugal than in the south" the notions "north of Portugal" and "south of Portugal" are not precisely defined. Like many other things in pattern definition, it depends on the analyst's goal whether a precise specification of a subset is required or whether just a fuzzy allusion would suffice.

For quick reference, we now bring examples of different types of patterns for different types of references together in table 3.3.

We have described these different types of patterns here mainly to provide a better understanding of the notion of a pattern. We shall sometimes refer to these types later, when discussing tools for exploratory data analysis. However, when we discuss the typology of data analysis tasks, which are the focus of this chapter, we shall talk about patterns in general, without regard to the distinctions between the various types.

**Table 3.3.** Examples of various types of patterns

Referrer type→ ↓Pattern type	Population	Time	Space
<b>Association</b>	A group of pupils with good marks in music and art and bad to average marks in physics	A period of westerly winds and high rainfall	A cluster of districts with high proportions of children and low proportions of elderly
<b>Differentiation</b>	Pupils with extremely high or low performance in all subjects  A group of pupils with extreme variation of marks	A day with extremely heavy rainfall  A period of highly changeable weather	A district or a few districts with low proportions of children, inside an area with mostly high proportions  A region with very high variability of the population structure
<b>Arrangement (in particular, trend)</b>	With respect to the ordering of the pupils according to their performance in mathematics, the performance in physics tends to increase	A period of increasing (or decreasing) daily temperatures  A period of alternating hot weather and thunderstorms	The proportions of elderly increase in the direction from the coast to inland  In this part of the city, built areas are separated by belts of parks and gardens
<b>Distribution summary</b>	Frequency distributions of pupils' marks in different subjects	Summers in this area are hot and dry, while winters are mild and humid	The proportions of children are high in the north-west, low inland, and close to the average in the other parts

### 3.4.4 Behaviours over Multidimensional Reference Sets

As we know, a dataset may have several referrers. We represent such data as a function of multiple variables, as is shown in (3.2) and (3.3). How can one investigate the behaviour of such a function?

For simplicity, let us first consider a function of two variables,  $f(x_1, x_2)$ , where the variable  $x_1$  takes values from some domain  $\mathbf{R}$  and the variable  $x_2$  takes values from some other domain  $\mathbf{Q}$ . Think of the US crime data as an example:  $f$  may be some of the attributes, for example, burglary rate;  $\mathbf{R}$  may stand for space; and  $\mathbf{Q}$  may stand for time.

The whole reference set of the function  $f$  consists of all possible pairs  $(\mathbf{r}, \mathbf{q})$ , where  $\mathbf{r} \in \mathbf{R}$  and  $\mathbf{q} \in \mathbf{Q}$ . In mathematics, this set is called the Cartesian product of the sets  $\mathbf{R}$  and  $\mathbf{Q}$  and denoted by  $\mathbf{R} \times \mathbf{Q}$  (in the crime dataset,  $\mathbf{R} \times \mathbf{Q}$  is the space–time continuum). The behaviour of the function on this reference set can be denoted by

$$\beta(f(x) \mid x \in \mathbf{R} \times \mathbf{Q}) \quad (3.25)$$

or

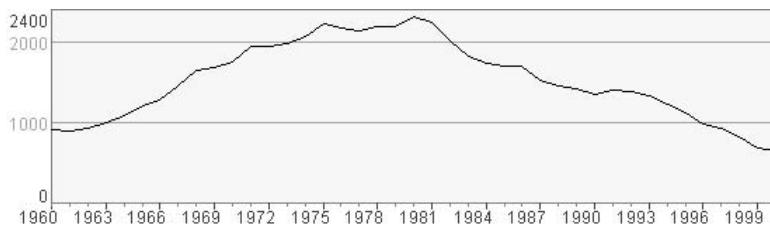
$$\beta(f(x_1, x_2) \mid x_1 \in \mathbf{R}, x_2 \in \mathbf{Q}) \quad (3.26)$$

Let us now choose some specific value  $\mathbf{r}$  of the variable  $x_1$  (using our example of crime in the USA, we choose a specific state, say, California) and allow the variable  $x_2$  to vary throughout the whole set  $\mathbf{Q}$ . We can try to explore the behaviour of the function  $f$  over  $\mathbf{Q}$  under the condition that  $x_1$  equals  $\mathbf{r}$ . This can be written formally as follows:

$$\beta_{\mathbf{Q}}(f(x_1, x_2) \mid x_1 = \mathbf{r}, x_2 \in \mathbf{Q}) \quad (3.27)$$

Here, the subscript  $\mathbf{Q}$  in  $\beta_{\mathbf{Q}}$  is used to emphasise that the behaviour is based on the set  $\mathbf{Q}$ .

In terms of our example, we can explore the temporal behaviour of the burglary rate in California over the period from 1960 to 2000. This behaviour is represented graphically in a time graph in Fig. 3.12.



**Fig. 3.12.** A graphical representation of the dynamics of the burglary rate in California during the period from 1960 to 2000

Similarly to the element  $\mathbf{r}$  (e.g. California), we can choose any other element of the set  $\mathbf{R}$  (e.g. any other state of the USA) and consider the corresponding behaviour of  $f$  over  $\mathbf{Q}$  (e.g. consider the dynamics of the burglary rates in Texas, Florida, and so on). Furthermore, we may be inter-

ested in exploring the whole set of these “partial behaviours”. Thus, in our example, we may wish to study how the dynamics of burglary rates are distributed over the territory of the USA. Actually,  $\beta_Q(f(x_1, x_2))$  may be viewed as a function of  $x_1$ , where  $x_1$  acquires various values from  $\mathbf{R}$ , and the set of values of the function consists of the possible behaviours (relevant to  $\mathbf{Q}$  and  $f$ ). Hence, we may think of the behaviour of the function  $\beta_Q$  over the set  $\mathbf{R}$ . This may be denoted by the expression

$$\beta_{\mathbf{R}}\{\beta_Q[f(x_1, x_2) \mid x_2 \in \mathbf{Q}] \mid x_1 \in \mathbf{R}\} \quad (3.28)$$

Square brackets [ and ] and braces { and } are used here instead of parentheses to provide a better understanding of how the symbols are grouped and how the parts of the expression are related.

The expression (3.28) encodes the idea of a “behaviour’s behaviour”. In our example of crime in the USA, this general idea is instantiated as “the spatial behaviour of the temporal behaviour”, or, in more conventional terms, “the spatial distribution of the temporal behaviours” (or temporal variations). A suitable graphical illustration would be the map shown in Fig. 3.13. In this map, time graphs like that shown in Fig. 3.12 are drawn in the locations of each state. We can see how various behaviours are distributed over the territory of the USA. Thus, we can observe that the states in the north-central part of the country had smaller burglary rates than the other states over the whole time period from 1960 to 2000. Another observation is that the states in the west and south-west have higher peaks in the middle of the time interval than the states in the east (with a few exceptions). It is possible to see some spatial clusters of states with similar temporal behaviours of the burglary rate; he clusters are outlined in Fig. 3.14.

We could now say that we have described the behaviour of the burglary rate over space and time. However, we recall that we have actually described  $\beta_{\mathbf{R}}\{\beta_Q[f(x_1, x_2) \mid x_2 \in \mathbf{Q}] \mid x_1 \in \mathbf{R}\}$  rather than  $\beta[f(x) \mid x \in \mathbf{R} \times \mathbf{Q}]$ . Could it be that these two things are equivalent? Let us investigate this.

When we look at the expression  $\beta_{\mathbf{R}}\{\beta_Q[f(x_1, x_2) \mid x_2 \in \mathbf{Q}] \mid x_1 \in \mathbf{R}\}$ , we notice that it imposes a specific order upon the variables  $x_1$  and  $x_2$  and the respective sets  $\mathbf{R}$  and  $\mathbf{Q}$ , while no particular order is specified in the expression  $\beta[f(x) \mid x \in \mathbf{R} \times \mathbf{Q}]$ . Does the order matter? Let us change the order in the first expression and try to interpret the result:

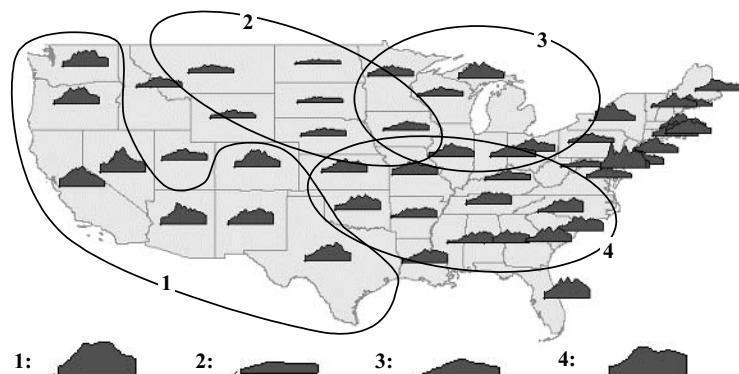
$$\beta_Q\{\beta_{\mathbf{R}}[f(x_1, x_2) \mid x_1 \in \mathbf{R}] \mid x_2 \in \mathbf{Q}\} \quad (3.29)$$

To interpret this formula, let us use our example of the burglary rates in the USA assuming, as before, that  $\mathbf{R}$  stands for space and  $\mathbf{Q}$  for time. Just as we could consider, for any particular state, the corresponding temporal behaviour of the burglary rate during the period from 1960 to 2000 (such

as that for California in Fig. 3.12), so we can consider, for any particular time moment, the spatial behaviour (distribution) of the burglary rate over the whole territory of the USA.



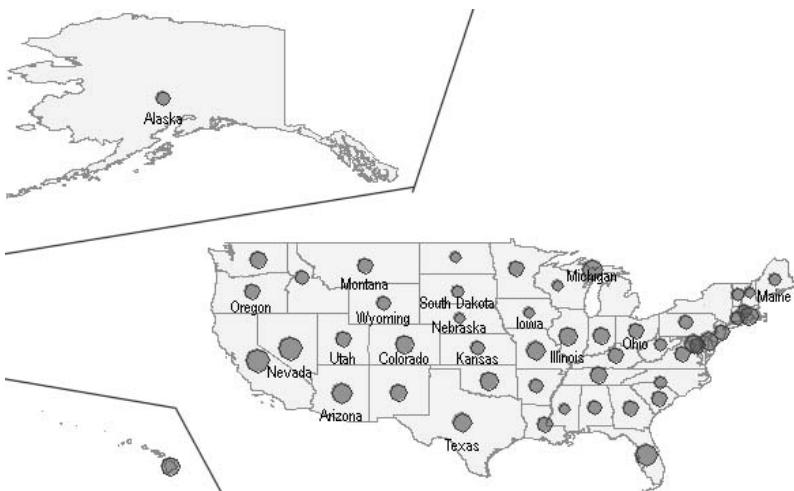
**Fig. 3.13.** A cartographic representation of the spatial distribution of the dynamics of burglary rates over the USA



**Fig. 3.14.** Spatial clusters of states with similar temporal behaviours of the burglary rate. Below the map, the typical behavioural patterns for each cluster are schematically shown

Thus, the map in Fig. 3.15 represents the spatial behaviour of the burglary rate in 1960 by graduated circles positioned in the locations of each

state. The sizes of the circles are proportional to the rate in the corresponding state.



**Fig. 3.15.** A cartographic representation of the spatial distribution of the burglary rate over the USA in 1960

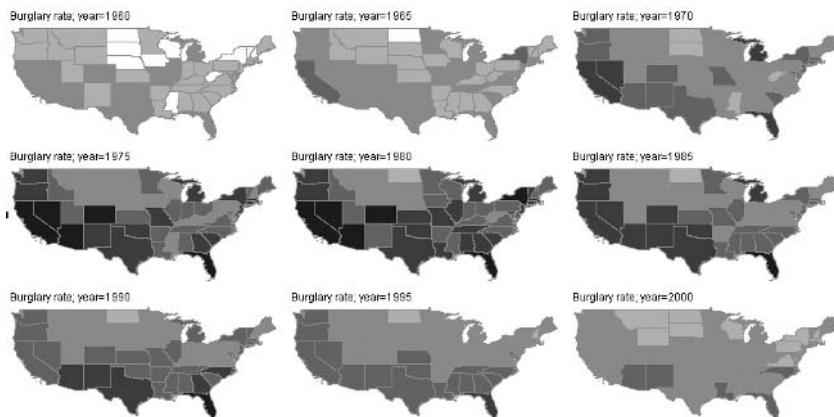
The spatial behaviour shown in Fig. 3.15 can be encoded by the formula

$$\beta_R(f(x_1, x_2) \mid x_1 \in R, x_2 = 1960) \quad (3.30)$$

Analogously, we can consider the spatial behaviour of the burglary rate in any other year. Furthermore, we can explore how the spatial behaviour changes over time during the whole period from the year 1960 to 2000. In other words, we can explore the behaviour of the function  $\beta_R$  over the set  $Q$  (i.e. time). This behaviour, which corresponds to the formula  $\beta_Q\{\beta_R[f(x_1, x_2) \mid x_1 \in R] \mid x_2 \in Q\}$ , could be represented visually in a series of 41 maps – one map per year. Since it would be hard to put all 41 maps on a page, we have limited our illustration to the nine maps shown in Fig. 3.16, one map for every fifth year. To obtain a smaller but still legible image, we have omitted the states of Alaska and Hawaii. The values of the burglary rates are represented by shades of grey in the areas corresponding to each state; the higher the value, the darker the shade.

With this series of maps, we can observe, for example, that the cluster of higher values in the south-west extended from 1960 to 1965 and merged with the cluster on the south. From 1965 to 1980, the burglary rates in this part of the territory mostly increased, as well as their variance, and the cluster of high values spread farther to the east. Starting from 1985, the values and their variance decreased, and the cluster shrank. In 1975 and

1980, a subcluster of very high values could be observed in the south-west, while in 1990 relatively higher values were found in the south of the area.



**Fig. 3.16.** A cartographic representation of the evolution of the spatial distribution of burglary rates over time

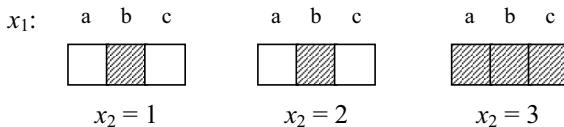
This description is quite different from that derived from Fig. 3.13. Of course, it is possible to describe one and the same behaviour in different ways. In our case, however, we do not have sufficient grounds to think that the images in Fig. 3.13 and 3.16 represent the same behaviour. Thus, one could hardly attach the second description to the first image and the first description to the second image. It is easier to accept that we have *two different behaviours* here, or, to state it better, *two different aspects of the overall behaviour* (represented by the formula  $\beta(f(x) | x \in \mathbf{R} \times \mathbf{Q})$ ). Since the two aspectual behaviours are not equivalent to each other, we cannot regard either of them as equivalent to the overall behaviour.

Can we be sure that a union of two aspectual behaviours is equivalent to the overall behaviour? Let us investigate this using a very simple artificial example. The dataset here has two referrers and one attribute. The first referrer,  $x_1$ , has the value set  $\mathbf{P}$ , containing three possible values: a, b, and c. The value set  $\mathbf{S}$  of the second referrer,  $x_2$ , also includes three values, denoted by 1, 2, and 3. The attribute  $a(x_1, x_2)$  has two possible values: yes and no. The data are shown in Table 3.4.

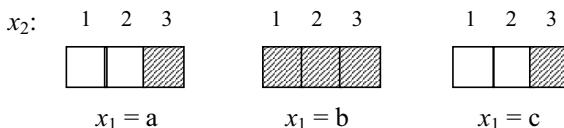
**Table 3.4.** An artificial dataset

$x_1$	a	a	a	b	b	b	c	c	c
$x_2$	1	2	3	1	2	3	1	2	3
$a(x_1, x_2)$	no	no	yes	yes	yes	yes	no	no	yes

The behaviour  $\beta_S\{\beta_P[a(x_1, x_2) \mid x_1 \in P] \mid x_2 \in S\}$  can be represented graphically as is shown in Fig. 3.17. Here, filled squares stand for the attribute value “yes” and hollow squares for “no”. In Fig. 3.18, the behaviour  $\beta_P\{\beta_S[a(x_1, x_2) \mid x_2 \in S] \mid x_1 \in P\}$  is represented graphically. It can be seen clearly that these two behaviours are not the same.

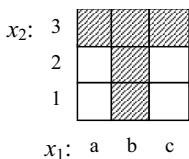


**Fig. 3.17.** A graphical representation of the behaviour  $\beta_S\{\beta_P[a(x_1, x_2) \mid x_1 \in P] \mid x_2 \in S\}$ , where  $P = \{a, b, c\}$ ,  $S = \{1, 2, 3\}$ , and  $a(x_1, x_2) \in \{\text{yes}, \text{no}\}$ . The filled squares represent the value “yes”, and the empty squares the value “no”



**Fig. 3.18.** A graphical representation of the behaviour  $\beta_P\{\beta_S[a(x_1, x_2) \mid x_2 \in S] \mid x_1 \in P\}$ , where  $P = \{a, b, c\}$ ,  $S = \{1, 2, 3\}$ , and  $a(x_1, x_2) \in \{\text{yes}, \text{no}\}$ . The filled and empty squares are used analogously to Fig. 3.17

Using the same symbolisation, we can also represent graphically the overall behaviour denoted by the formula  $\beta_{P \times S}(a(x_1, x_2) \mid x_1 \in P, x_2 \in S)$ . For this purpose, we map the values of the referrers onto two orthogonal axes. The result is shown in Fig. 3.19.



**Fig. 3.19.** A graphical representation of the behaviour  $\beta_{P \times S}(a(x_1, x_2) \mid x_1 \in P, x_2 \in S)$ . The symbolisation is the same as in Fig. 3.17 and 3.18

The image of the overall behaviour shown in Fig. 3.19 contains a feature that is not seen in Fig. 3.17 and 3.18, specifically, a T-shape. For this artificial example, it is hard to say whether this feature is important or not. It is clear, however, that the T-shape is pertinent only to the whole reference space, i.e.  $P \times S$ , whereas in Fig. 3.17 and 3.18 we see only slices of this space.

What conclusions can we draw from this example? When we have a dataset with multiple referrers (or, in the traditional terminology, a multi-dimensional dataset<sup>2</sup>), we can consider various slices of the reference space and investigate the behaviour of the underlying phenomenon on these slices. Furthermore, we can consider multiple slices simultaneously and make some observations about the variation of the behaviour over this series of slices. This gives us a *partial* understanding of the overall behaviour of the phenomenon on the whole reference set, because a series of slices can reflect only a certain aspect of the overall behaviour.

Such an aspectual behaviour may be viewed as a projection of the overall behaviour, like a two-dimensional projection of a three-dimensional object in a technical drawing. Then, just as a single projection is often insufficient for understanding the shape of an object, so consideration of a single aspectual behaviour is insufficient for understanding the overall behaviour. Of course, it would be preferable to investigate the object or a three-dimensional model of it rather than two-dimensional pictures. However, it is often impossible to use the object or a model. In this case, a sufficient number of projections and slices (the latter are especially needed when the object has internal structure) are required for understanding the shape. Moreover, even when a sufficient number of images have been provided, a significant mental effort is involved in the process of comprehension of the shape.

Analogously, for exploring a behaviour, it is preferable to consider the reference set as a whole rather than slices of it. However, this may not always be possible. In many cases, an analyst has to derive an understanding of the overall behaviour from a sufficient number of aspectual behaviours. Thus, in a case with two referrers, there are two such behaviours, and both should be investigated in order to understand the overall behaviour properly. Unfortunately, as the dimensionality increases, the number of aspectual behaviours grows dramatically. Thus, for a reference set  $\mathbf{R} \times \mathbf{Q} \times \mathbf{S}$ , we have six possible aspectual behaviours:  $\beta_{\mathbf{R}}(\beta_{\mathbf{Q}}(\beta_{\mathbf{S}}(\dots)))$ ,  $\beta_{\mathbf{R}}(\beta_{\mathbf{S}}(\beta_{\mathbf{Q}}(\dots)))$ ,  $\beta_{\mathbf{Q}}(\beta_{\mathbf{R}}(\beta_{\mathbf{S}}(\dots)))$ ,  $\beta_{\mathbf{Q}}(\beta_{\mathbf{S}}(\beta_{\mathbf{R}}(\dots)))$ ,  $\beta_{\mathbf{S}}(\beta_{\mathbf{R}}(\beta_{\mathbf{Q}}(\dots)))$ , and  $\beta_{\mathbf{S}}(\beta_{\mathbf{Q}}(\beta_{\mathbf{R}}(\dots)))$ . When the number of referrers equals 4, there are 24 such behaviours. In general, the number of aspectual behaviours for a dataset with N referrers equals  $N!$  ( $N$  factorial), i.e. the product  $N \cdot (N-1) \cdot (N-2) \cdots 2 \cdot 1$ . And it should always be remembered that additional effort is required to reconstruct the overall behaviour from aspectual behaviours, similarly to the effect of understand-

---

<sup>2</sup> We prefer not to use the traditional term “multidimensional data”, because it does not imply distinguishing between referrers and attributes. Thus, this term is often applied to a dataset with many attributes but a single referrer.

ing the shape of a three-dimensional object from two-dimensional projections. Just considering all possible aspects may, in general, not be enough.

The passage above illustrates the general truth that analysis of multidimensional data is not easy. However, data analysis does not always pursue such ambitious goals as obtaining a full understanding of the overall behaviour of a phenomenon. In many particular cases, only certain aspects of the overall behaviour are relevant to the problem to be solved or only certain aspects excite the interest of the analyst. This can make life much simpler, if the relevant aspects are correctly chosen.

Let us now recall the classification of spatio-temporal analysis tasks by Koussoulakou and Kraak (1992). In the introductory section of this chapter, we have noted that the task category “overall level with respect to both space and time” includes tasks with several different meanings:

- What is the trend over the area during the whole time?
- How did the spatial distribution evolve over time?
- How do the temporal trends vary over the area?

In this section, we have demonstrated why this is so. In our framework, the first question (to our taste, rather vaguely formulated) can be viewed as referring to the overall behaviour of a spatio-temporal phenomenon, while the other two questions ask about its aspectual behaviours.

In their paper, Koussoulakou and Kraak reported experiments on evaluating animated maps from the perspective of supporting different types of spatio-temporal analysis tasks. While animated maps are currently rather popular, it becomes clear after a close look at this technique that what it visualises is one of the aspectual behaviours rather than the overall behaviour. More precisely, an animated map shows how a spatial behaviour evolves over time, as series of maps does (for example, the maps in Fig. 3.16). This corresponds to the second question in the list above. Hence, using only an animated map is insufficient for understanding the overall behaviour.

In the experiments described by Koussoulakou and Kraak, the subjects were asked the question “What is the trend in high population densities over the whole time?” This question was chosen as representative of the task category “overall space and overall time”, and an animated map was sufficient for answering it. However, the question does not actually ask about the overall behaviour  $\beta_{SXT}(\dots)$  but rather about the aspectual behaviour  $\beta_T(\beta_S(\dots))$ , where **S** stands for space and **T** for time.

We must apologise for having digressed from the main topic of this chapter and starting to discuss tools for data analysis, which will be the subject of the next chapter. Let us return to task typology. We have dis-

cussed synoptic tasks; more precisely, behaviour characterisation tasks. We hope that this section has contributed to the understanding of the notion of behaviour, in particular, by the use of illustrated examples of behaviour characterisation.

### 3.4.5 Pattern Search and Comparison

There is a certain similarity between behaviour characterisation tasks and the elementary tasks of direct lookup. In a direct lookup task, a particular reference is specified, and the goal is to find the corresponding characteristics. In a behaviour characterisation task, a particular (sub)set of references is specified, and the goal is to find a pattern that represents (approximates) the behaviour of the characteristics appropriately over this reference set.

A similar parallel exists for inverse lookup tasks. Let us recall that in these tasks, it is necessary to find references corresponding to specified characteristics. Similarly, for a specified pattern, one may wish to find subsets of references such that the behaviour over those subsets corresponds to this pattern. Here are some examples of such tasks:

- Find the time intervals in which the stock price increased.
- Find the areas in Europe with a predominance of broadleaved forests over coniferous.
- Find the regions in Portugal with high proportions of young people.
- Find spatio-temporal clusters of earthquake occurrences.

All these examples include descriptions of certain patterns. The example concerning the stock price includes a specification of a trend. In the example concerning forests in Europe, a certain common property is specified, and the task is to find coherent areas (spatial clusters) characterised by this common property. This is an example of an absolute specification of a common property, that is, it does not involve any comparison of the areas to be found with the remaining territory. In contrast, the example concerning young people in Portugal contains a relative specification of a common property: the proportions of young people in the target regions must be higher than in the remaining territory of Portugal. In the example concerning earthquakes, the target subsets of earthquakes must have close spatial locations and close times of occurrence.

We shall call such tasks “pattern search tasks” and represent them by the general formula

$$\mathcal{P}R: \mathbf{f}(x) | x \in R \approx \mathbf{P} \quad (3.31)$$

This should be read as “for the specified pattern  $\mathbf{P}$ , find a set of references  $R$  such that the behaviour of the function  $f(x)$  over  $R$  can be approximated by the pattern  $\mathbf{P}$ ”.

Let us use the analogy with elementary tasks further and introduce synoptic tasks of behaviour comparison. We defined elementary comparison tasks as tasks that imply determining *relations between characteristics* corresponding to different *individual references*. Instead of dealing with individual references and corresponding “atomic” characteristics represented by values of attributes, synoptic tasks deal with reference sets and corresponding behaviours represented by patterns. Therefore, synoptic comparison tasks should deal with *relations between behaviours* and, accordingly, between patterns approximating these behaviours.

What are possible the relations between behaviours? Let us imagine that we are considering the behaviours of some function on two distinct reference subsets. We would probably note first of all whether those behaviours were similar or dissimilar. For example;

- Is the behaviour of the stock price during the first week similar to that during the second week?
- Is the age structure of the population in the north of Portugal similar to that in the south of Portugal?

In answer to such questions, we would say either “yes, the behaviours are similar” or “no, the behaviours are dissimilar”. Then we would probably justify our answer by recounting what the similarities or differences were. In many cases, it is impossible to say decisively whether two behaviours are similar or different. Instead, one would say that the behaviours have both similarities and differences. In such a case, an extended answer is expected, with the similarities and differences particularised. For example, “similarly to the first week, the stock price mostly increased during the second week, but the increase was slower than during the first week”. The level of detail and precision in the description of similarities and differences may vary depending on the analyst’s goals. In particular, an analyst may be interested in obtaining some numerical measures of the degree of difference. For example, he or she may wish to determine how much the rate of growth of the price decreased from the first to the second week. Another example is to measure how an actual behaviour deviates from a desired or typical behaviour, such as the deviation of the actual profit dynamics in a company from the planned dynamics or the deviation of the trajectory of a vehicle from its usual route.

Sometimes, two behaviours may be characterised as *opposite*. For example, the stock price may grow during the first week but then decrease during the second week. It may happen that the proportions of children are

high and the percentages of elderly people are low in the north of a country, whereas in the south the former are low and the latter are high.

Hence, two behaviours may be characterised as similar, opposite, or dissimilar, and this characterisation may be extended by recounting similar and distinct features or by computing quantitative measures of the degree of similarity or dissimilarity.

When discussing elementary tasks, we defined comparison tasks as tasks having a relation in their target, i.e. the goal is to find how some items are related. By analogy, we define *behaviour comparison tasks* as tasks of determining how two behaviours are related, *in terms of their similarities and differences*. This addition is important: behaviours can also be related in quite a different sense, and we would like to consider the other type of relations separately.

The general formula for a behaviour comparison task is

$$\{p_1, p_2, \lambda : \beta_1 \approx p_1; \beta_2 \approx p_2; p_1 \lambda p_2\} \quad (3.32)$$

where  $\beta_1$  and  $\beta_2$  are two behaviours,  $p_1$  and  $p_2$  are patterns that approximate these behaviours, and  $\lambda$  is the relation between the patterns (and hence the behaviours) that needs to be determined. From this formula, it may be noted that behaviour comparison tasks are compound, analogously to elementary comparison tasks. A behaviour comparison task includes one or more behaviour characterisation tasks, i.e. tasks of finding patterns to approximate certain behaviours.

Depending on what  $\beta_1$  and  $\beta_2$  in the general formula (3.32) are, behaviour comparison tasks may be divided into subcategories, quite analogously to elementary comparison tasks:

- The behaviour of a function (attribute) over a specified subset of references is compared with a specified behaviour pattern. For example, “Did the stock price increase during the first week?” Here, the behaviour of the stock price is compared with the pattern described as an “increase”. This and similar questions can be represented by the formula

$$\{p, \lambda : \beta(f(x) | x \in \mathbf{R}) \approx p; p \lambda \mathbf{P}\} \quad (3.33)$$

where  $\beta(f(x) | x \in \mathbf{R})$  is the behaviour of the function  $f(x)$  over the reference set  $\mathbf{R}$ ,  $p$  stands for an unknown pattern approximating this behaviour, and  $\mathbf{P}$  is a specified pattern.

This subcategory includes, among others, tasks of determining how characteristics over some reference set deviate from certain specified characteristics, for example inspecting the emission of pollutants from a chemical plant over a period of time with respect to the permissible threshold values.

- The behaviours of a function (attribute) over two specified reference subsets are compared. For example, “Compare the behaviours of the stock price over the first and the second week”. The appropriate formula is

$$?p_1, p_2, \lambda: \beta(f(x) | x \in R_1) \approx p_1; \beta(f(x) | x \in R_2) \approx p_2; p_1 \lambda p_2 \quad (3.34)$$

Here,  $\beta(f(x) | x \in R_1)$  is the behaviour of the function  $f(x)$  over the reference set  $R_1$ , and  $\beta(f(x) | x \in R_2)$  is the behaviour of the same function over the reference set  $R_2$ .

- The behaviours of two different functions (attributes) over a specified reference set are compared. For example, “Compare the behaviour of the stock price during this time period with the variation of the Dow Jones Industrial Average index over the same period”. The formula representing this type of task is

$$?p_1, p_2, \lambda: \beta(f_1(x) | x \in R) \approx p_1; \beta(f_2(x) | x \in R) \approx p_2; p_1 \lambda p_2 \quad (3.35)$$

Here,  $\beta(f_1(x) | x \in R)$  is the behaviour of the function  $f_1(x)$  over the reference set  $R$ , and  $\beta(f_2(x) | x \in R)$  is the behaviour of the function  $f_2(x)$  over the same reference set.

- The behaviours of two different functions (attributes) over two different reference sets are compared. For example, “Compare the spatial distribution of the proportions of children in 1991 with the spatial distribution of the birth rate in 1981”. The formula representing this type of task is

$$?p_1, p_2, \lambda: \beta(f_1(x) | x \in R_1) \approx p_1; \beta(f_2(x) | x \in R_2) \approx p_2; p_1 \lambda p_2 \quad (3.36)$$

Here,  $\beta(f_1(x) | x \in R)$  is the behaviour of the function  $f_1(x)$  over the reference set  $R_1$ , and  $\beta(f_2(x) | x \in R)$  is the behaviour of the function  $f_2(x)$  over the reference set  $R_2$ .

If the descriptions of the last two subtypes of behaviour comparison tasks are compared with the descriptions of the corresponding subtypes of elementary comparison tasks, it may be noticed that a particular phrase appears in the descriptions of the elementary tasks but does not appear when the synoptic tasks are described. This phrase is “The value domains of the attributes must coincide or overlap”. This is not an accidental omission: one can compare behaviours of attributes even when the values of the attributes are incomparable. For example, it is meaningless to compare the weight of a person with the waist size of that person. However, it would be quite meaningful to compare the variation of the body weight over a period of time with the variation of the waist size over the same period. One could even compare the behaviour of the body weight with the variation of the person’s diet over a year, for example, to see whether changes in the diet

correspond to changes in the body weight. In this case, the behaviour of a numeric attribute is compared with that of a non-numeric attribute.

A good example of comparing quite different behaviours is known from biology, where a similarity between ontogeny and phylogeny was discovered. Ontogeny is the embryonic development process of an organism of a particular species, while phylogeny is the evolutionary history of a species. These processes can be compared despite the completely different time-scales: days or months for ontogeny and millions of years for phylogeny.

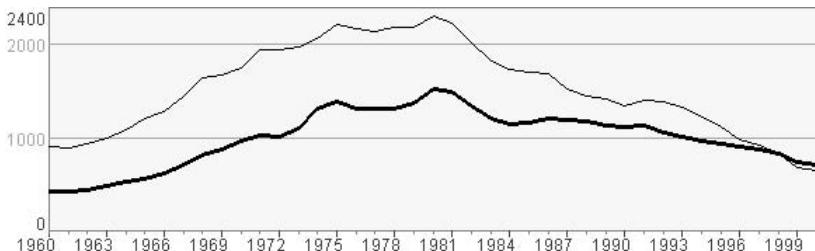
The latter example shows that the possibility of comparing behaviours does not actually require these behaviours to be based on comparable reference sets. Moreover, the base sets may be quite different in their nature, but the behaviours may still be comparable. Thus, behaviours can easily be compared if their bases have consistent properties concerning the existence of ordering and distances. For example, one can compare a process that develops with time with how a vibratory impulse is transmitted along a string. The reference set of the first behaviour is time – a linearly ordered set with distances. The reference set of the second behaviour is the extent of the string, which has a nature quite different from time while being also a linearly ordered set with distances. This commonality of properties enables a rather straightforward comparison of the respective behaviours.

In some cases, there may even be a need to compare or an interest in comparing behaviours with bases that have no common properties. Let us recall that a behaviour may be approximated by a summary pattern involving some aggregation of characteristics over the base set. Many aggregation techniques do not take into account any properties of the reference set, and hence can be applied to quite different reference sets and still produce comparable results. Thus, one can compute the average mark of a pupil over a school year and compare it with the average mark of a group of schoolchildren at the end of the year. Here, the first summary refers to a linearly ordered reference set with distances, while the second summary characterises an unordered set without distances.

Moreover, aggregation allows one even to compare behaviours based on reference sets with different dimensionalities. Thus, the variation of a pupil's performance over a school year can be compared with the variation of the performance of all pupils in a class over the year, for example by computing the average performance of the class at each time moment. Here, the former behaviour refers to time, while the latter behaviour originally refers to time and a population (i.e. a group of pupils). In order to make the comparison possible, one needs to represent the original behaviour, which is based on a two-dimensional reference set, by an appropriate one-dimensional pattern. For this purpose, an aspectual behaviour is considered instead of the overall behaviour. Then, for each value of one referrer (in

our example, time), the behaviour with respect to the other referrer (i.e. the class) is approximated by a summarising pattern. It is the behaviour of this summarising pattern that is actually compared with the one-dimensional behaviour of the individual pupil's performance.

Similarly to the above example concerning pupils, Fig. 3.20 allows us to compare the behaviour of the burglary rate in California during the period from 1960 to 2000 with the behaviour of the mean burglary rate over the whole of the USA during this period.



**Fig. 3.20.** Comparison of the dynamics of the burglary rate in California during the period from 1960 to 2000 (thin line) with the dynamics of the mean burglary rate over the USA during the same period (thick line)

### 3.4.6 Inverse Comparison

Let us use again the parallel between the types of elementary tasks and the types of synoptic tasks to define synoptic tasks of *inverse comparison*. For elementary tasks, inverse comparison means determining relations between references corresponding to specific characteristics. If we project this statement to synoptic tasks, inverse comparison means determining relations between sets of references corresponding to specific behaviours or patterns. Here are a few examples:

- Compare the durations of the periods of growth of the stock price.
- What is the time lag between the periods of growth of the stock price?
- Compare the spatial position and extent of the cluster of high unemployment rates with the position and extent of the cluster of high proportions of young people in the population.

Besides comparing two reference sets, inverse comparison tasks may also involve determining relations between reference sets and individual references. For example:

- Compare the position of a cluster of high concentrations of cholera occurrences with the locations of water pumps.

- What is the time lag between the date of introducing new anti-crime measures and the beginning of a decreasing trend in the crime rate?

When discussing elementary tasks of inverse comparison, we showed that any such task is composite and involves at least one inverse lookup task. The same applies to synoptic tasks of inverse comparison: they are composite and involve pattern search tasks. Thus, in order to compare the durations of periods of stock price growth or find the time lag between them, it is necessary first to find those periods. Here, a particular pattern, “stock price growth”, is specified. An analyst first performs the task of searching for this pattern. This results in the finding of reference subsets (i.e. time intervals) on which this pattern is observed. Then, the analyst determines the relations between these subsets. Since the subsets are time intervals, the analyst is interested in temporal relations between them.

The task concerning the proportions of young people and unemployment rates requires an analyst to find areas in space where two specified patterns referring to different attributes are observed. Then, the analyst has to determine the spatial relations between the areas, i.e. whether they coincide, overlap, or lie far from each other.

In the example concerning cholera occurrences, one needs to look for a behaviour that can be characterised as a “high concentration of cholera occurrences”. The result is some area where the behaviour is approximated by the specified pattern. Then, one needs to determine the spatial relations between this area and the positions of available water pumps. These positions result from elementary lookup tasks<sup>3</sup>. Similarly, the task concerning combating crime involves a pattern search task to find the interval of a decreasing trend in the crime rate and a lookup task to find the moment at which the new measures were introduced. Then, the temporal relation between the interval and the moment must be determined.

Hence, synoptic tasks on inverse comparison involve pattern search tasks but may also involve lookup tasks. Therefore, the variety of inverse comparison tasks is larger for synoptic tasks than for elementary tasks. Here are the relevant distinctions that need to be taken into account in defining subcategories of synoptic tasks of inverse comparison:

1. Must the base set of a pattern be compared with some specified set or reference (i.e. a constant) or with the result of a pattern search or lookup

---

<sup>3</sup> These tasks may be regarded as either direct or inverse lookup tasks, owing to the possibility of a dual treatment of space, either as one of characteristics of the water pumps (i.e. an attribute) or as a container of the water pumps (i.e. a referrer). The possibility of such a dual treatment of space and time has been discussed in Chap. 2.

task? Such distinction can be made, for example, between the tasks “How is the cluster of high concentrations of thefts situated with respect to the centre of the town?” and “How is the cluster of high concentrations of thefts situated with respect to the location of the railway station?” In the former example, “the centre of the town” may be regarded as a constant, while in the latter example “the location of the railway station” is supposed to result from a lookup task of finding the location of the railway station.

2. Must the base set of a pattern be compared with another reference set or with individual references? This kind of difference exists, for example, between the tasks “How is the period of a decrease in the crime rate related to the period of the summer vacation?” and “How is the period of a decrease in the crime rate related to moment at which the new crime-fighting measures were introduced?” In the first case, the period of a decrease in the crime rate is compared with a time period, i.e. a set, and in the second case, it is compared with a time moment, i.e. an element.
3. Are different attributes or different behavioural patterns of the same attribute involved? Thus, the task “What are the relative positions of the cluster of high unemployment rates and the cluster of high proportions of young people?” involves two different attributes, while the task “What are the relative positions of the cluster of high unemployment rates and the cluster of low unemployment rates?” includes two different patterns specified for the same attribute.

These three dichotomies can potentially yield  $2 \times 2 \times 2 = 8$  different subcategories. We do not find it necessary to consider each of them in detail and thus introduce eight additional formulae. Nevertheless, to show that these formulae may look like, let us write out one of them:

$$?R_1, R_2, \lambda: \beta(f_1(x) | x \in R_1) \approx P_1; \beta(f_2(x) | x \in R_2) \approx P_2; R_1 \lambda R_2 \quad (3.37)$$

This formula represents a generic task where the patterns  $P_1$  and  $P_2$ , specified for two different attributes ( $f_1(x)$  and  $f_2(x)$ ), need to be detected and then the corresponding reference sets ( $R_1$  and  $R_2$ ) need to be compared, i.e. the relation  $\lambda$  between them needs to be determined. It is easy to notice two pattern search tasks included in this inverse comparison task:  $?R_1: \beta(f_1(x) | x \in R_1) \approx P_1$  and  $?R_2: \beta(f_2(x) | x \in R_2) \approx P_2$  (compare these expressions with the formula (3.31) representing the class of pattern search tasks). It is also easy to observe an analogy with the formal representation of elementary tasks of inverse comparison.

It should be noted that synoptic tasks of inverse comparison do not allow such freedom as direct comparison tasks, i.e. tasks of comparing behaviours. In inverse comparison, references and reference sets are com-

pared, and they need to be comparable. Thus, one cannot compare areas in space with intervals in time, whereas it may be possible to compare corresponding behaviours: for this purpose, one needs to approximate those behaviours by comparable patterns.

### 3.4.7 Relation-Seeking

Besides lookup and comparison, we defined earlier one more category of elementary tasks, specifically tasks of finding occurrences of specified relations between characteristics and determining the corresponding references. Since all of the categories of synoptic task considered thus far have been introduced as counterparts of certain classes of elementary tasks, it should not be a surprise that the category of relation-seeking tasks will now be projected on to synoptic tasks.

Elementary tasks deal with individual references, and corresponding characteristics represented by values of attributes. On the synoptic level, we have, instead, sets of references, and corresponding behaviours represented by patterns. Hence, the definition of relation-seeking tasks on the elementary level would be translated to the synoptic level as “find occurrences of specific relations between behaviours and determine the corresponding reference sets”. The relations may be “same”, “different”, or “opposite”, or include a more precise specification of what the similarities and differences are supposed to be. Let us give a few examples:

- Are there any adjacent areas with large differences in yearly climate?
- Can any recurrent pattern be found in the behaviour of the stock price?
- During what time intervals was the trend of the stock price opposite to that during a given interval?
- In what parts of Portugal did the employment structure (i.e. the proportions of people working in industry, agriculture, and services) change dramatically from 1981 to 1991?
- Were there any storks that followed the same migration route in the specified season?
- Is there a time interval when the behaviour of the stock price was opposite to that of the Dow Jones Industrial Average index?

Just as elementary tasks of relation-seeking may differ according to the way in which the set of references to be searched through is constrained, the same can be said concerning the corresponding group of synoptic tasks. Here is the result of translating the variants of relation seeking tasks considered earlier from the elementary to the synoptic level.

*Case 1.* Search for two or more reference sets such that a specified relation exists between the behaviours of some attribute(s) over these reference sets and, additionally, another specified relation exists between the reference sets themselves. For example, “Find two contiguous time intervals with opposite trends of the stock price” or “Are there any adjacent areas with large differences in yearly climate?”

In the first example, there are two constraints. One of them concerns the relation between the behaviours over two different time intervals: these behaviours must be opposite. The other constraint specifies the relation between the time intervals themselves: they must be contiguous. Similarly, in the second example, one of the constraints says that the behaviours over two areas (subsets of geographical space) must be different, while another constraint requires that these areas are adjacent. An appropriate general formula could be

$$\begin{aligned} &?R_1, R_2, p_1, p_2: R_1 \Psi R_2; \\ &\beta(f(x) | x \in R_1) \approx p_1; \beta(f(x) | x \in R_2) \approx p_2; p_1 \Delta p_2 \end{aligned} \quad (3.38)$$

Here, the variables  $R_1$  and  $R_2$  stand for the unknown reference subsets,  $\Psi$  is the specified relation that must exist between these subsets (e.g. overlapping, not overlapping, or adjacent), and  $\Delta$  is the specified relation (e.g. similar, different, or opposite) that must exist between the behaviours of the attribute  $f(x)$  based on these two subsets. Rather than deal with the behaviours as such, an analyst would consider some patterns  $p_1$  and  $p_2$  approximating these behaviours.

The example task “Can any recurrent pattern be found in the behaviour of the stock price?” can also be subsumed under this case. The task can be reformulated as “Find time intervals in which the behaviours of the stock price are similar to each other”. Although it is not specified explicitly, certain relations are expected to exist between the time intervals to be found: they need to be non-overlapping and to have approximately the same length.

*Case 2* is a task in which not only a relation  $\Delta$  between behaviours but also one of the corresponding reference subsets is specified. The goal is to find another reference subset such that the relation  $\Delta$  exists between the behaviours based on these two reference subsets, as in the example task “During what time interval was the trend of the stock price opposite to that during a given interval?” The formula for this task subtype is

$$\begin{aligned} &?R_2, p_1, p_2: \\ &\beta(f(x) | x \in R') \approx p_1; \beta(f(x) | x \in R_2) \approx p_2; p_1 \Delta p_2 \end{aligned} \quad (3.39)$$

Here, the constant  $\mathbf{R}'$  stands for the specified reference subset and the variable  $R_2$  stands for the unknown reference subset, which needs to be found. The meaning of the remaining symbols is the same as in Case 1.

*Case 3* may arise when a dataset has two or more referrers, such as space and time. On the elementary level, it is possible to specify the values of some of the referrers, whereas the values of the remaining referrers need to be found. Similarly, on the synoptic level, values/subsets of some referrers may be specified, whereas subsets/values of the remaining referrers need to be found. This applies, for example, to the task “In what parts of Portugal did the employment structure change dramatically from 1981 to 1991?” Here, the reference set consists of two components, space and time, and the attribute is the employment structure. The values of the temporalreferrer (i.e. the years 1981 and 1991) are specified as task constraints. The target is to find an area in space, i.e. a subset of the spatialreferrer. Although this must be a single subset of values (of one of the referrers), two different reference subsets are actually involved, since two different values of the second referrer must be considered. Over these two reference sets, the behaviours of the attribute “employment structure” are required to be substantially dissimilar.

This situation can be generalised into the formula

$$\begin{aligned} & ?R, p_1, p_2: \\ & \beta(f(x_1, x_2) \mid x_1 \in R, x_2 = q_1) \approx p_1; \\ & \beta(f(x_1, x_2) \mid x_1 \in R, x_2 = q_2) \approx p_2; \\ & p_1 \Delta p_2 \end{aligned} \tag{3.40}$$

and into similar formulae varying in the number of variables and in the values of which of them are specified. In (3.40),  $f(x_1, x_2)$  is an attribute defined on a two-dimensional reference set and represented, according to our symbolism, as a function of two variables,  $x_1$  and  $x_2$ . The symbols  $q_1$  and  $q_2$  stand for two specific values of the second variable, i.e. these are constants. In the example concerning the employment structure,  $q_1$  corresponds to the year 1981 and  $q_2$  to the year 1991.  $\beta(f(x_1, x_2) \mid x_1 \in R, x_2 = q_1)$  and  $\beta(f(x_1, x_2) \mid x_1 \in R, x_2 = q_2)$  denote two “partial behaviours”: the value of one variable ( $x_2$ ) is fixed, while the other variable ( $x_1$ ) varies. The symbol  $\Delta$  stands for the specified relation between the behaviours. In our example, this corresponds to the relation “dissimilar”.

A slightly different case is the task “Were there any storks that followed the same migration route in the specified season?” Here, we again have two referrers, a group of storks and time, and a spatial attribute, the values of which are the positions of the storks at different time moments. For the temporalreferrer, a particular set (time interval) is specified: the migration

season. The goal is to find two or more individual values of the other referent. This is different from the previous example, where specific individual values of one of the referrers were fixed, and the goal was to find some set of values of the other referrer. Similarly to the previous example, two or more two-dimensional reference sets are involved; these coincide with respect to the temporal component but differ in the stork component. The task constraints specify that the behaviours of the attribute “position” based on these different reference sets (i.e. the trajectories in space followed by different storks during the specified time interval) must be similar. This and analogous tasks could be represented formally as

$$\begin{aligned} & ?q_1, q_2, p_1, p_2: \\ & \beta(f(x_1, x_2) \mid x_1 \in \mathbf{R}, x_2 = q_1) \approx p_1; \\ & \beta(f(x_1, x_2) \mid x_1 \in \mathbf{R}, x_2 = q_2) \approx p_2; \\ & p_1 \Lambda p_2 \end{aligned} \quad (3.41)$$

Here, as in the previous case,  $f(x_1, x_2)$  is an attribute defined on a two-dimensional reference set,  $\mathbf{R}$  stands for a specified set of values of the first referrer (in our example, the migration season), and  $q_1$  and  $q_2$  denote the individual values of the second referrer that need to be found (in our example, the storks).  $\beta(f(x_1, x_2) \mid x_1 \in \mathbf{R}, x_2 = q_1)$  and  $\beta(f(x_1, x_2) \mid x_1 \in \mathbf{R}, x_2 = q_2)$  are two partial behaviours, and the relation  $\Lambda$  (in our example, similarity) must exist between them.

*Case 4.* Search for a single reference set where a specified relation between the behaviours of different attributes exists. For example, “Is there a time interval when the behaviour of the stock price was opposite to that of the Dow Jones index?” The general formula for this subcategory is

$$?R, p_1, p_2: \beta(f_1(x) \mid x \in R) \approx p_1; \beta(f_2(x) \mid x \in R) \approx p_2; p_1 \Lambda p_2 \quad (3.42)$$

Here, the variable  $R$  stands for the unknown reference set, and  $\Lambda$  is the specified relation (e.g. similar, different, or opposite) that must exist between the behaviours of the attribute  $f_1(x)$  and the attribute  $f_2(x)$ .

*Case 5.* Search for two or more reference sets such that specified relations between the behaviours of two or more different attributes exist simultaneously, for example “Find pairs of time intervals with similar trends of industrial growth and opposite trends of unemployment”. We shall not write a formula for this case, since it would be rather cumbersome, like the formula for the corresponding group of elementary tasks. From the previous examples, it is easy to grasp the general principle of transforming formulae for elementary tasks into the formulae for the corresponding synoptic tasks, and hence the structure of the formula for this case should be quite understandable.

As with elementary tasks, we would like to point out the compound structure of relation-seeking synoptic tasks, which are built from the following subtasks:

1. For a reference subset, characterise the behaviour of the corresponding characteristics (i.e. approximate it with an appropriate pattern).
2. For the pattern resulting from subtask 1, define a pattern or a collection of patterns related to this pattern in a specified way (i.e. similar, dissimilar, or opposite).
3. For the pattern(s) resulting from subtask 2, find reference subsets such that the behaviours based on them can be approximated by these patterns.

The first subtask may be applied to a specified reference subset, or it may be necessary to repeat the whole sequence of operations for multiple subsets. In the latter case, some constraints are typically specified, which restrict the number of different reference subsets to be considered and somehow direct the selection of subsets for further consideration.

### 3.4.8 Recap: Synoptic Tasks

We have defined synoptic tasks as tasks requiring consideration of sets of references in their entirety, in contrast to elementary tasks, which deal with individual references (i.e. elements). The principal notion that we have to consider for synoptic tasks is that of a *behaviour*. A behaviour can be understood as a particular configuration of characteristics (i.e. values of attributes) corresponding to some reference set and, like the reference set, is considered in its entirety and in relation to the reference set.

We distinguish the notion of a *configuration* from the notion of a *subset* of characteristics. There are two principal differences:

- A data function may associate several references with one and the same characteristic. While such a characteristic is just a single element of a set of characteristics, it occurs as many times in a configuration as there are references that it corresponds to. Every such occurrence can be differentiated from other occurrences of the same characteristic.
- A configuration of characteristics reflects the structure of the corresponding set of references, that is, a certain system of relations between the elements of the reference set.<sup>4</sup> Thus, a configuration corresponding

---

<sup>4</sup> Such a structure may be “natural” or specially introduced. For example, a set of time moments has a natural linear ordering, while a set of people may be arranged according to the alphabetical order of their names.

to a linearly ordered set of references is a sequence of characteristics, where the order in which the characteristics appear is determined by the order of the references. In contrast, the relations between elements of a set are typically determined only by the nature of those elements. Thus, a set of characteristics corresponding to a linearly ordered set of references may be unordered or may have its own ordering independent of the ordering of the references.

Hence, not just subsets of characteristics are considered in synoptic tasks but configurations of characteristics corresponding to (sub)sets of references; such configurations are called “behaviours”. This reflects the dependent role of characteristics with respect to references.

A behaviour can be represented, or approximated, by a *pattern*. A pattern is a mental construct or a statement in some language representing the character and essential features of a behaviour. A pattern does not necessarily have a verbal form; it may be, for example, computational, algebraic, or even graphical.

In order to define subcategories of elementary tasks, we manipulated the formula  $f(x) = y$  representing the link between the two principal components of data, references and characteristics. We differentiate elementary tasks on the basis of which items of this expression occur in the task target (i.e. what information needs to be found) and which items participate in the task constraints (i.e. are assumed to be known).

For synoptic tasks, the fundamental formula is  $\beta(f(x) \mid x \in R) \approx P$ , where  $\beta(f(x) \mid x \in R)$  stands for the behaviour of the characteristic component (attribute or set of attributes)  $f(x)$  over the reference set  $R$ , and  $P$  denotes a pattern that approximates (i.e. describes or summarises) this behaviour. By manipulating this formula, we can derive classes of synoptic tasks. There is a correspondence between the classification of elementary tasks and that of synoptic tasks: classes of elementary tasks have their counterparts on the synoptic level. Table 3.5 is a comparative table that summarises both elementary and synoptic tasks.

This task typology elaborates Bertin’s approach to classifying tasks on the basis of the structure of data. One dimension of Bertin’s classification is the reading level, i.e. whether a task addresses individual data items (elements) or sets. This corresponds to our division of tasks into elementary and synoptic tasks. Our category of synoptic tasks embraces Bertin’s intermediate and overall levels: unlike Bertin, we do not make a distinction between tasks referring to a whole set of data items and tasks referring to its subsets. The other dimension used by Bertin is the task type corresponding to the data component referred to in the target of a task. This matches our subdivision of elementary and synoptic task categories.

**Table 3.5.** Correspondence between classes of elementary and synoptic tasks

Elementary	Synoptic
<p><b><i>Lookup</i></b></p> <ul style="list-style-type: none"> <li>• <i>Direct lookup</i></li> </ul> <p>On a given date, what is the price of stock X?</p> <p>What was the population of Loures in 1981?</p>	<p><b><i>Pattern identification</i></b></p> <ul style="list-style-type: none"> <li>• <i>Behaviour characterisation (pattern definition)</i></li> </ul> <p>During a given time interval, what was the trend of the stock price?</p> <p>How was the population distributed over Portugal in 1981?</p>
<ul style="list-style-type: none"> <li>• <i>Inverse lookup</i></li> </ul> <p>For a given price, on what date(s) was it attained?</p> <p>When and where did the population exceed 300 000?</p>	<ul style="list-style-type: none"> <li>• <i>Pattern search</i></li> </ul> <p>Find time intervals in which the stock price increased.</p> <p>Find regions in Portugal with high proportions of young people.</p>
<p><b><i>Comparison</i></b></p> <ul style="list-style-type: none"> <li>• <i>Direct comparison</i></li> </ul> <ul style="list-style-type: none"> <li>– With specified attribute values</li> </ul> <p>On a given date, did the stock price exceed €1000?</p> <p>Was the proportion of children in the population of Loures in 1991 less than 15%?</p>	<p><b><i>Behaviour (pattern) comparison</i></b></p> <ul style="list-style-type: none"> <li>• <i>Direct comparison</i></li> </ul> <ul style="list-style-type: none"> <li>– With a specified pattern</li> </ul> <p>Did the stock price increase during a given time interval?</p> <p>Are high proportions of children in the population concentrated in the north of Portugal?</p>
<ul style="list-style-type: none"> <li>– Between values of the same attribute(s) for different references</li> </ul> <p>Compare the stock prices on the first and last days of the week.</p> <p>How did the population of Loures change from 1981 to 1991?</p>	<ul style="list-style-type: none"> <li>– Between behaviours of the same attribute(s) over different reference sets</li> </ul> <p>Compare the behaviours of the stock price during the first and the second week.</p> <p>Compare the distributions of the proportion of children over Portugal in 1981 and 1991.</p>
<ul style="list-style-type: none"> <li>– Between values of different attributes for the same reference</li> </ul> <p>Compare the total values of the imports and exports of the given country.</p>	<ul style="list-style-type: none"> <li>– Between behaviours of different attributes over the same reference set</li> </ul> <p>Compare the behaviour of the stock price with the variation of the Dow</p>

<b>Elementary</b>	<b>Synoptic</b>
<p>How does the number of people without primary education in Loures in 1991 compare with the number of high school students in the same year?</p> <ul style="list-style-type: none"> <li data-bbox="74 357 519 451">– Between values of different attributes for (partly) different references</li> </ul>	<p>Jones index over the same time period.</p> <p>Compare the distributions of the proportion of children and the proportion of old people over the territory of Portugal.</p>
<p>Compare the immigration to the USA with the emigration from Mexico.</p> <ul style="list-style-type: none"> <li data-bbox="74 566 519 681">– How does the number of people without primary education in Loures in 1981 compare with the number of high school students in 1991?</li> </ul>	<ul style="list-style-type: none"> <li data-bbox="533 357 979 451">– Between behaviours of different attributes over (partly) different reference sets</li> </ul> <p>Compare the trends in the immigration to the USA and in the emigration from Mexico.</p>
<ul style="list-style-type: none"> <li data-bbox="74 566 519 681">• <i>Inverse comparison</i></li> <li data-bbox="74 746 519 779">– With specified reference(s)</li> </ul> <p>Did the stock price reach €1000 before or after a given date?</p>	<ul style="list-style-type: none"> <li data-bbox="533 566 979 681">• <i>Inverse comparison</i></li> <li data-bbox="533 746 979 779">– With specified reference sets</li> </ul> <p>Compare the spatial distribution of the proportion of children in 1991 with the spatial distribution of the birth rate in 1981.</p>
<ul style="list-style-type: none"> <li data-bbox="74 746 519 981">– Where is the district with the highest crime rate situated with respect to the town centre?</li> </ul> <p>Where is the cluster of high crime rates situated with respect to the town centre?</p>	<ul style="list-style-type: none"> <li data-bbox="533 746 979 981">– How is a decreasing trend in the stock price related to the period of the summer vacation?</li> </ul> <p>Where is the cluster of high crime rates situated with respect to the town centre?</p>
<ul style="list-style-type: none"> <li data-bbox="74 1014 519 1107">– Between references corresponding to different values of the same attribute(s)</li> </ul> <p>Compare the dates on which the prices €1000 and €1100 were attained.</p> <p>How far is the district with the highest crime rate from that with the lowest crime rate?</p>	<ul style="list-style-type: none"> <li data-bbox="533 1014 979 1107">– Between the reference sets corresponding to specified behaviours of the same attribute(s)</li> </ul> <p>How is the period of stock price growth related to that of stock price decrease?</p> <p>What are the relative positions of the clusters of high and low crime rates?</p>
<ul style="list-style-type: none"> <li data-bbox="74 1325 519 1419">– Between references corresponding to specific values of different attributes</li> </ul> <p>Compare the date when the highest stock price was attained with the date when the highest level of the NASDAQ</p>	<ul style="list-style-type: none"> <li data-bbox="533 1325 979 1419">– Between the reference sets corresponding to specified behaviours of different attributes</li> </ul> <p>How is the period of stock price growth related to that of growth of the NASDAQ index?</p>

<b>Elementary</b>	<b>Synoptic</b>
<p>index was reached.</p> <p>Where is the district with the highest crime rate situated compared with the district with the highest unemployment?</p>	<p>What are the relative positions of the cluster of high crime rates and that of high unemployment?</p>
<p><b><i>Relation-seeking</i></b></p> <ul style="list-style-type: none"> <li>– Between values of attribute(s) and, at the same time, between references</li> </ul> <p>On what dates did the price of the stock decrease in comparison with the previous date?</p> <p>Are there any two storks that ever visited the same place simultaneously?</p>	<p><b><i>Relation-seeking</i></b></p> <ul style="list-style-type: none"> <li>– Between behaviours of attribute(s) and, at the same time, between reference sets</li> </ul> <p>Find two contiguous time intervals with opposite trends in the stock price.</p> <p>Are there any two storks that followed the same migration route but during different time intervals?</p>
<ul style="list-style-type: none"> <li>– Between characteristic(s) of a specified reference and characteristics of other references</li> </ul> <p>On what date(s) did the stock price exceed the price attained on a given date?</p> <p>Which of the storks were in the same place as Prinzessin on 1 February?</p>	<ul style="list-style-type: none"> <li>– Between an attribute behaviour over a specified reference subset and attribute behaviours over other reference subsets</li> </ul> <p>In what time interval was the trend in the stock price opposite to that in the given interval?</p> <p>Is there any stork with a migration route similar to that of Prinzessin?</p>
<ul style="list-style-type: none"> <li>– Between values of the same attribute(s) for partly different references (in a dataset with multiple referrers)</li> </ul> <p>In which districts of Portugal did the population decrease from 1981 to 1991?</p> <p>On what date(s) did the storks Prinzessin and Moritz visit the same places?</p>	<ul style="list-style-type: none"> <li>– Between behaviours of the same attribute(s) over partly different reference sets (in a dataset with multiple referrers)</li> </ul> <p>In which parts of Portugal did the employment structure change dramatically from 1981 to 1991?</p> <p>During what time periods did the storks Prinzessin and Moritz move in opposite directions?</p>
<ul style="list-style-type: none"> <li>– Between values of different attributes for the same reference</li> </ul> <p>Find countries where the imports exceed exports.</p>	<ul style="list-style-type: none"> <li>– Between behaviours of different attributes over the same reference set</li> </ul> <p>Find periods when the trends in the imports and exports were the same.</p>

<b>Elementary</b>	<b>Synoptic</b>
Where and when did the motor vehicle theft rate exceed the burglary rate?	At what times did the spatial distribution of motor vehicle theft rates differ from that of burglary rates?

There are several extensions in our typology as compared with Bertin's schema. First, we make a distinction between referential and characteristic data components and treat these components differently. We assume that characteristic components play a dependent role, since their values are determined by values of referential components. Therefore, we divide tasks into elementary and synoptic tasks on the basis of how *referential* components are addressed, i.e. whether individual references or reference sets are involved. Second, we subdivide the categories of elementary and synoptic tasks depending on whether referential or characteristic components appear in the task target. Tasks with characteristic components in the target are called direct tasks, while tasks with referrers in the target are called inverse tasks. Third, unlike Bertin, we explicitly consider comparison tasks. We define comparison tasks as tasks of determining relations between characteristics (direct comparison) or between references (inverse comparison). Additionally, we introduce relation-seeking tasks, in which occurrences of specified relations need to be detected. Fourth, we pay special attention to multidimensional datasets, i.e. datasets with multiple referential components. These multiple referrers can be addressed on different levels (elementary or synoptic) independently of each other, which results in numerous varieties of tasks.

Although the task typology summarised in Table 3.5 is rather detailed, it is still incomplete. Let us recall that, when considering behaviour comparison tasks, we made the special point that these tasks have to do with similarity/difference relations between behaviours but do not touch upon other possible relations. Hence, the typology needs to be completed with other types of relations. Let us now try to do this.

## 3.5 Connection Discovery

### 3.5.1 General Notes

When studying a phenomenon, an analyst is interested not only in describing or summarising its behaviour but also in explaining it. The analyst wishes to find out the driving forces that make the phenomenon behave in the way observed. These forces may be internal or external. Internal forces

originate from the inherent structure of the phenomenon and interactions between its structural components. External forces originate from interactions between the phenomenon and other phenomena. Hence, the goal is to determine what components and/or phenomena interact and how they interact.

In most cases, it turns out to be impossible to explain a behaviour only on the basis of the data available, and an analyst needs to make use of relevant domain knowledge in trying to reason out pertinent causal links and influence mechanisms. In our study, we do not touch on these reasoning processes as such, but consider possible foundations for such processes that can be obtained by exploring data. The main value of data exploration is that it allows the analyst to spot significant interactions, which provide food for thought and give direction to further reasoning.

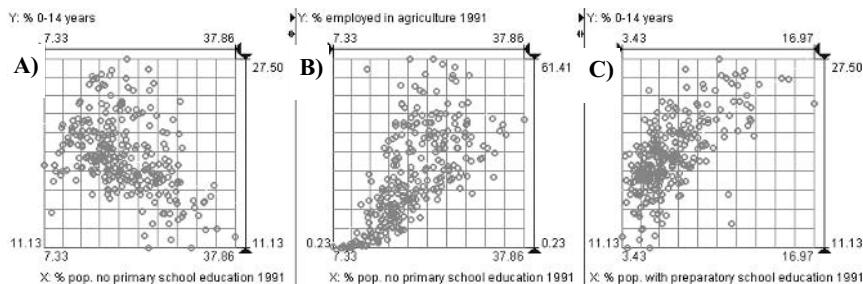
Let us illustrate our thoughts by an example. An analyst explores a dataset containing demographic data related to the census districts of a certain country. He/she notices that there is some connection between the attributes “Number of cars per capita” and “Proportion of population suffering from long-term illnesses”: the higher the number of cars per capita is, the lower the proportion of ill people. It would be wrong to try to explain this finding by an influence of either of these two attributes upon the other attribute. In this way, one could come to an absurd conclusion that using cars strengthens health or that ill people usually avoid buying cars. It is much more appropriate to refer to relevant domain knowledge and guess that both attributes may have something to do with material deprivation, and that the actual root of the observed link may lie here.

Although deriving conclusions from observations and establishing links with domain knowledge are very important and interesting topics, they deserve close attention and separate investigation, and we prefer not to consider them within our current study devoted to exploratory data analysis. Instead, we focus on what provides a basis for reasoning and gives stimulus to it, i.e. on observations that may result directly from data exploration. In this context, we investigate what indications of probable interactions between phenomena or between different aspects of a phenomenon may be present in data, and how to detect them. At the next stage, the indications detected need to be interpreted, but this stage is outside the scope of our current work.

Hence, one of the goals of EDA is to find indications of possible causal links or influences within or between phenomena. It is clear that this requires the data to be considered on the synoptic rather than elementary level : a causal link is expected to manifest itself throughout a reference set as a whole or at least a substantial part of it.

One of the indications of a possible connection may be a correspondence between behaviours, i.e. when two behaviours are either similar or opposite. However, to compare behaviours, one approximates them by patterns and then compares the patterns. There are a potentially infinite number of patterns representing the same behaviour. They differ in their type, degree of simplification, precision, coverage, etc. Not all patterns are suitable for the purposes of detecting possible interactions between phenomena or aspects of a phenomenon. Thus, one may characterise the behaviour of the proportion of children on the set of municipalities of Portugal as a statistically normal distribution with an average value of 19.1% and a median of 18.91%. Analogously, the behaviour of the proportion of people without primary school education may be described as a statistically normal distribution with an average value of 19.89% and a median of 18.92%. These descriptions are very similar; however, this should not be treated as an indication of a possible connection between the proportion of children and the proportion of people without primary school education.

Hence, not all patterns are appropriate for behaviour comparison when the goal is to discover connections rather than just to note similarities and differences. On the other hand, comparison of individual behaviour patterns is not the only possible way to spot interactions. Thus, it is usual to detect a connection (it is more customary to say “correlation”) between two numeric attributes by representing their values on a scatterplot. Examples of scatterplots are shown in Fig. 3.21.



**Fig. 3.21.** Using scatterplots for detecting correlations between attributes: (A) proportion of people without primary school education versus proportion of children; (B) proportion of people without primary school education versus proportion of people employed in agriculture; (C) proportion of people with preparatory school education versus proportion of children

A scatterplot does not represent the individual behaviour of either of the two attributes and does not allow one to compare these individual behaviours. Instead, it represents something like a “mutual behaviour” of these

attributes, i.e. it shows how these attributes behave with respect to each other. The possible presence of a connection is indicated by the pattern perceived in the scatterplot. Thus, the scatterplots (B) and (C) in Fig. 3.21 suggest a probable connection more clearly than does scatterplot (A), which represents the mutual behaviour of the attributes “proportion of children” and “proportion of people without primary school education”. From these three scatterplots, it may be concluded that there is a probable link between the proportion of people without primary school education and percentages of people employed in agriculture and between the proportion of children and the proportion of people with preparatory school education. However, it is harder to believe that there is a connection between the proportion of children and the proportion of people without primary school education, although the summary patterns of these two attributes are very similar.

This example shows us that detecting probable connections should not be viewed as just a special case of a behaviour comparison task. On the one hand, comparison of individual behaviours is not necessarily involved in connection discovery. On the other hand, if such a comparison is involved, it appears that the patterns used to approximate individual behaviours must satisfy some specific requirements. In our opinion, these are quite sufficient reasons for introducing a special task category that deals with detecting signs of possible inherent connections and interactions. We shall call this task category “connection discovery”. This category is included in that of synoptic tasks, since it requires dealing with sets rather than individual elements.

### 3.5.2 Properties and Formalisation

So far we have successfully applied formal representations to reveal relevant distinctions between tasks and, on this basis, to define task classes. For consistency and in the hope of gaining some insights, we shall adopt the same approach in handling connection discovery tasks.

In order to arrive at a proper formalisation, let us contemplate the properties of connection discovery tasks that we already know. First, connection discovery tasks require dealing with sets rather than elements and, hence, belong to the class of synoptic tasks. Second, these tasks are about connections and, hence, involve at least two items that are supposed to be connected. Third, as we have demonstrated, connection discovery tasks do not necessarily involve consideration of individual behaviours. On the other hand, these tasks still have something to do with the notion of behaviour, which encompasses substantial features of a phenomenon.

Let us focus for a while on the notion of a phenomenon. Of course, it is not strictly defined, but we may still have some ideas concerning what can be regarded as a phenomenon. Thus, we have no special reason to assume that a phenomenon is something atomic, having no internal structure. Then, since such an internal structure exists, there is no obstacle to regarding the constituents also as phenomena. Interrelationships between these subphenomena constitute an important part of the behaviour (more precisely, the internal behaviour) of the phenomenon embracing them.

On the other hand, a phenomenon never exists in isolation, out of context. Why not regard this phenomenon plus its context, in which it coexists and interacts with other phenomena, as a more complex phenomenon? Then, the interrelations between the phenomena included in this superphenomenon belong to the (internal) behaviour of this superphenomenon.

The purpose of the above contemplation is to demonstrate that we can not only consider relations between behaviours but also treat *relations as behaviours*. We have already used the expression “mutual behaviour” when we discussed scatterplots. Our idea concerning connection discovery tasks is that they deal with such mutual, or relational, behaviours. As in other synoptic tasks, the primary goal is to investigate and understand this sort of behaviour.

We would like to distinguish relational behaviours from “stand-alone” behaviours of individual phenomena, which have been denoted using the symbol  $\beta$ . Let us use the symbol  $\rho$  for relational behaviours.

An expression containing  $\beta$  contains a single item in parentheses, and this is the function  $f(x)$ , which represents a dataset. An expression containing  $\rho$  must contain at least two items. What are these items?

We have already considered one case of a relational behaviour: a correlation between two numeric attributes, which can be detected using a scatterplot. This case can be generalised to the notion of a mutual, or relational, behaviour of two arbitrary attributes (functions) defined on the same reference set. The attributes may be far from numeric, as, for example, in a study of interrelations between a person’s nutrition and the person’s health and lifestyle. The following formal expression may be used to represent the notion of a relational behaviour:

$$\rho(f_1(x), f_2(x) \mid x \in \mathbf{R}) \quad (3.43)$$

There is no necessity to limit this expression to only two functions; it can easily be extended to contain three or more functions inside the parentheses. Another extension is to discard the assumption that the functions are necessarily defined on the same reference set. Thus, one can look for connections between the relief and the yearly variation of climate in some

territory. Here, the relief is a phenomenon defined in space (or, in other words, has space as its reference set), while the climate is defined in space and time, i.e. its reference set is a combination of two referrers. The reference sets of these two phenomena are, strictly speaking, different; however, they share a common referrer – space.

Is it necessary that the reference sets of two attributes have something in common for one to be able to explore their relational behaviour, i.e. seek possible connections? In order to investigate this, we have tried to recall or invent examples of attributes with distinct reference sets that could still be somehow related. One of the examples that came to mind is the following.

A typical task of archaeologists is to relate the spatial distribution of artefacts found in an excavation area to the historical development of that area. The first phenomenon refers to three-dimensional space (two planar dimensions plus depth), while the second phenomenon refers to time. These reference sets may seem quite different, but they cannot be treated as completely unrelated. Thus, there is a rather direct correspondence between the depth dimension and the historical time: as a rule, the deeper an artefact lies, the older it is. There is also a link between time and the planar dimensions. Thus, suppose that excavations are performed in the location of an ancient settlement. This settlement was founded at some moment in the past in a certain location, and then gradually extended over time. Therefore, the spatial position of a construction discovered by excavation, with respect to the original location of the settlement, has a certain relation to the time when it was constructed. These strong links between two reference sets make it quite natural to look for links between the respective phenomena.

A similar interplay between a spatial and a temporal reference set exists when one wishes to investigate how the variation of the width and other properties of the annual rings in the cross-section of a tree trunk is related to the variation of the climate over the years during which the tree was growing. Here, the spatial position of a ring, i.e. its distance from the centre, corresponds directly to an interval on the time axis. Therefore, it is quite easy to check whether the corresponding year was wet or dry, for instance.

There are many situations where the reference sets of two or more phenomena are simply different subsets of one larger reference set, for example, in a study of links between the concentration of people in cities and the economic situation in rural areas, or between warm and cold oceanic currents and the climate on different parts of a coast. In such cases, there are particular relations between the subsets, for example neighbourhood in space.

We could not find any examples where two reference sets were completely unrelated but where it would nevertheless be meaningful to look for connections between the respective phenomena. One example that came to mind, however, was the attempts of astrologers to relate the positions of stars and planets in the sky at the moment of a person's birth (a spatially referenced phenomenon) and the subsequent life history of that person (a temporally referenced phenomenon). However, it may be that the absence of any relation between the two reference sets discourages many people from trusting the results of such investigations.

Nevertheless, in a formal sense, the reference sets involved in connection discovery tasks may be different, as is indicated in the expression

$$\rho(f_1(x), f_2(z) \mid x \in \mathbf{R}, z \in \mathbf{Z}) \quad (3.44)$$

where  $z$  is a reference variable distinct from  $x$  and taking its values from a distinct reference set  $\mathbf{Z}$ .

So far, we have been talking about connections between different phenomena or attributes. It should not be concluded, however, that it is not possible to consider internal connections within a single phenomenon. As a phenomenon may have an internal structure, connections between the structural components may be investigated. This is, in principle, equivalent to the case of multiple phenomena, since the components may also be treated as phenomena. However, there is another possibility: to look for connections between parts of a phenomenon referring to different reference subsets. For example, people often make observations such as "if the summer is hot and dry, one may expect a cold winter". This is a connection between parts of the same phenomenon, climate, referring to different subsets of time. We have intentionally used the word "subsets" rather than "intervals" because the observation here is not about a particular summer and a particular winter. It addresses all summers, i.e. a non-contiguous subset of time, and all winters following them, i.e. another non-contiguous subset, which is specified in a relative way with respect to the first subset. Another example of exploring internal connections in a time-related phenomenon is the investigation of links between the pre-natal and post-natal development of a baby. An example with a different reference set could be found in an investigation of land cover, where an analyst may look for frequent associations between different vegetation types in adjacent locations.

Seeking connections between parts of a single phenomenon is not in principle different from the case of two or more phenomena:

- It is a synoptic task, since it aims at establishing general relations, pertinent to the phenomenon as a whole or a substantial part of it rather than to particular individual manifestations (i.e. elements of data).

- It involves at least two items to find connections between.
- It addresses the behaviour of the phenomenon but is not restricted to a comparison of its behaviours on different reference subsets.

The term “mutual behaviour” may, at first glance, seem inappropriate for denoting internal relations within a single phenomenon. Nevertheless, it sounds quite natural if these internal relations are understood as relations between certain parts of the phenomenon. It is clear that these parts may behave in this or that way with respect to each other.

Let us write down a formal expression for a mutual behaviour of two parts of a phenomenon:

$$\rho(f(x), f(x') \mid x \in R_1, x' \in R_2) \quad (3.45)$$

Here, we use the same function symbol  $f$  to indicate that parts of one and the same phenomenon are dealt with. We use different reference variables  $x$  and  $x'$  to indicate that the parts refer to distinct reference subsets, denoted by  $R_1$  and  $R_2$ . It is assumed that  $R_1$  and  $R_2$  are subsets of a certain reference set  $R$ , on which the function  $f$  is defined.  $R_1$  and  $R_2$  must be different, but it is not required that they do not overlap.

When we discussed “simple” behaviours (denoted by the symbol  $\beta$ ), we said that a data analyst represents them by patterns, i.e. compact characterisations. We tried to investigate what may serve as a pattern and introduced four types of patterns, called “association”, “differentiation”, “distribution summary”, and “arrangement”. The goal of a behaviour characterisation task is to approximate some behaviour by an appropriate pattern that satisfies the analyst’s criteria concerning generality, simplicity, precision, coverage, and other requirements (e.g. an analyst may have a special need to describe some behaviour as a trend).

Quite analogously, a mutual, or relational, behaviour (denoted by the symbol  $\rho$ ) also needs to be represented in some compact way that is appropriate to the analyst’s goals and requirements. This compact representation, which describes discovered connections, may also be considered as a kind of pattern. However, it seems that the types of patterns introduced earlier are not relevant to connection discovery tasks.

Let us investigate what a pattern describing a mutual behaviour may look like. If we take, for example, two numeric attributes, we may find that these attributes are positively or negatively correlated. Positive correlation means that low values of one attribute mostly co-occur with low values of the other attribute, while high values of the first attribute tend to occur together with high values of the other attribute. Negative correlation means co-occurrence of low values of one attribute with high values of the other attribute, and vice versa. Correlation between two attributes may be indi-

cated verbally, represented in a highly aggregated way as the correlation coefficient, specified through an equation, or portrayed graphically as a regression line.

The notion of correlation may be generalised to non-numeric attributes and understood as the co-occurrence of specific values of such attributes. For example, sunny weather in a region may usually co-occur with a strong wind, while rain co-occurs with windless weather. Sometimes correlations are specified in a negative way, i.e. by stating that some values never co-occur: for example, naturally red hair of a person and the susceptibility of this person to tanning. Statements concerning correlations between values of non-numeric attributes are often supported by statistics, for example the probability of a certain combination of value occurring, which may be related to the probability of those values occurring in combination with other values.

Hence, there is at least one possible way to describe a mutual behaviour, that is, to characterise it as a correlation. A correlation is an undirected, or symmetrical, connection: nothing is said concerning which attribute influences the other. In some cases, the direction of the connection may be unclear to an analyst, and the analyst may prefer to describe it as a correlation in such a case. In other cases, it is quite clear that neither of the attributes depends on the other. A correlation signifies that both attributes are influenced by some third attribute (or a group of attributes), probably, unknown as yet. This is the case in the example concerning the number of cars per capita, which is negatively correlated with the proportion of ill people.

However, there are cases where it is known or expected that some attribute or group of attributes influences another attribute or several attributes. In such a case, an analyst may prefer to represent the connection in some “directed” way rather than to characterise it as a correlation. For example, the analyst might describe the dependency in the form of rules

$$\text{If } f_1(x) \in \mathbf{C} \text{ then } f_2(z) \in \mathbf{D},$$

where  $f_1$  and  $f_2$  are two attributes,  $\mathbf{C}$  is a specific subset of the values of the attribute  $f_1$ , and  $\mathbf{D}$  is a specific subset of the values of the attribute  $f_2$ . A dependency could also be represented by logical implications or by equations. A very useful approximation to a dependency or a set of dependencies is a simulation model, which allows one to predict the future development of a phenomenon.

Hence, as an alternative to being described as a correlation, a mutual behaviour may be characterised as a dependency, or influence. We treat the latter two words as opposites and assume that the statement “A depends on B” is equivalent to the statement “B influences A”.

One more type of essential connection is a structural connection. As an example, consider the variation of the prices of holiday apartments over

time. This behaviour can be represented as an interplay of two components: a trend component, which accounts for the general growth of the prices over a long-term period (measured in years), and a seasonal component, which corresponds to the price variation during a year. Let us give another example, a quite famous one. Until 1530, when Copernicus published his great work asserting the rotation of the Earth around the Sun, the movement of the planets in the night sky seemed enigmatic and illogical, and could not be rationally explained. Copernicus's discovery brought the understanding that the visible movement of the planets is a composition of their own movement with the movement of the Earth.

Hence, we have found three types of pattern relevant to relational behaviours: an undirected connection (correlation), a directed connection (a dependency or influence, depending on the order chosen for the items involved in the connection), and a structural connection (when the behaviour is a composition of several essential parts).<sup>5</sup> Of course, an analyst may also find that no connection exists at all.

Usually, the discovery of a correlation pattern cannot be seen as the final result of data exploration, although it may be quite an important finding. A correlation pattern does not explain anything; on the contrary, it needs an explanation itself. What makes two or more phenomena (attributes) behave in a correlated way? Further investigation is required in order to discover any forces that influence these phenomena (an influence pattern) or detect that the phenomena are structural parts of some embracing phenomenon and act according to the principles and logic of this embracing phenomenon (a structural pattern). It is quite possible that an analyst may fail to find an explanation on the basis of existing data and domain knowledge. In this case, it becomes necessary to collect additional potentially relevant data or to look for additional potentially relevant knowledge.

Another note that we can make is that an analyst is typically not satisfied with just spotting a pattern and identifying its type; the analyst needs to specify the pattern in an appropriate level of detail, to derive a model of a phenomenon in order to be able to predict its further behaviour, for example.

Since the patterns used for approximating mutual behaviours are different from those representing individual behaviours, we shall call them “connection patterns” or “linkage patterns” rather than “behaviour patterns”. Now we can give a general definition of a connection discovery

---

<sup>5</sup> We cannot guarantee that this list is complete, and it was not our intention to provide a complete enumeration of the possible connection patterns. The major goal was to demonstrate that these patterns are quite different from those discussed in relation to behaviour characterisation tasks.

task as a task with the goal of finding a linkage pattern that approximates some mutual behaviour, either between different phenomena or between parts of the same phenomenon. Analogously to behaviour characterisation tasks, we can represent connection discovery tasks by a general formula such as

$$?l: \rho(\dots) \approx l \quad (3.46)$$

where the parentheses may contain any of the variants encoded in the expressions (3.43)–(3.45) and  $l$  is the linkage pattern that needs to be derived.

### 3.5.3 Relation to the Former Categories

How do connection discovery tasks relate to the categories of elementary and synoptic tasks described earlier? We have stressed several times that connection discovery tasks are synoptic tasks since they deal with regular linkages (such as causal, logical, or structural relations) relevant to a dataset as a whole or to substantial parts of it, rather than occasional associations between individual elements.

Let us take a particular example dataset, namely the data about stork migration. A possible connection discovery task for these data could be:

- Are there any commonalities (regular patterns) in the movements of different storks? Can any general model of stork behaviour be derived from the data? (We shall refer to this example as  $\rho_1$  from now on.)

An analyst might also be interested in relating the behaviour of the storks to other phenomena. For example:

- How does the behaviour of the storks depend on the relief and land cover of the underlying territory, on its flora and fauna, and on weather conditions, etc.? (We shall refer to this example as  $\rho_2$ .)

The example  $\rho_1$  can be reformulated as detecting significant correlations in the behaviours of different storks and finding general principle(s) explaining these correlations. Hence, this is a task of discovering connections between parts of the same phenomenon, specifically, the migration movement of storks. At the same time, it can also be classified as a “simple” behaviour characterisation task: the goal is to derive a pattern (model) that approximates the behaviour of a stork in the course of its seasonal migration.

This observation appears surprising. Why then should we discuss connection discovery tasks at all if they are just a particular flavour of behav-

iour characterisation tasks? Well, there may be different levels at which the behaviours are described. One possibility is just to recount our direct observations, for example, “The majority of storks migrate to equatorial areas of Africa, while some of them reach as far as South Africa.” Another possibility is to try to relate different aspects of the phenomenon and arrive at conclusions such as “The further north the departure point of a stork is, the further south it tends to fly” or “Storks whose destination points are further away usually move faster than those which migrate to nearer locations.” (These statements are fictitious; they have been devised for illustration purposes rather than resulting from real data analysis.)

Hence, from the formal point of view, connection discovery tasks can be subsumed under the category of behaviour characterisation tasks. We need only to extend the set of possible types of patterns (i.e. association, differentiation, arrangement, and distribution summary) with appropriate types of linkage patterns. However, from the point of view of the cognitive operations and effort involved, connection discovery differs from just describing. Thus, if we think how we would perform the task  $\rho_1$ , we would find it appropriate to perform various types of synoptic tasks, such as:

- Divide the behaviours into identifiable primitive patterns, for example “stillness”, “fast movement south”, and “wandering” (behaviour characterisation).
- Look to see if the sequences of the primitive patterns are the same or different for different storks (direct behaviour comparison).
- Find the time intervals of each primitive pattern in the movement of each stork (pattern search), and compare these intervals within individual behaviours and across behaviours (inverse behaviour comparison).
- Determine the frequencies of the various transitions between the primitive patterns, find frequently repeated sequences (relation-seeking), compare the time intervals on which these sequences are based (inverse comparison).

We do not insist that  $\rho_1$  must be done exactly in this way; we only want to demonstrate that  $\rho_1$  is expected to be a rather complex task involving many subtasks of various types. What can be said definitely is that one essential operation must be there in any case: abstraction from the results obtained by performing the subtasks towards something coherent, which could serve as an appropriate model of the behaviour of storks.

On the other hand, even a “basic” behaviour characterisation task may be rather complex and requires abstraction. The analyst needs to grasp the essential features of a behaviour, abstract from the particulars, and arrive at something sufficiently general. This cannot always be done in one stage.

Let us recall our discussion concerning the analysis of behaviours over multidimensional reference sets. In many cases, an overall behaviour has to be reconstructed from multiple aspectual behaviours. Here, there are two points quite similar to those made about connection discovery tasks: first, subtasks (of analysing aspectual behaviours) are involved; and second, significant cognitive effort is required to synthesise patterns of aspectual behaviours into a consistent pattern of the overall behaviour.

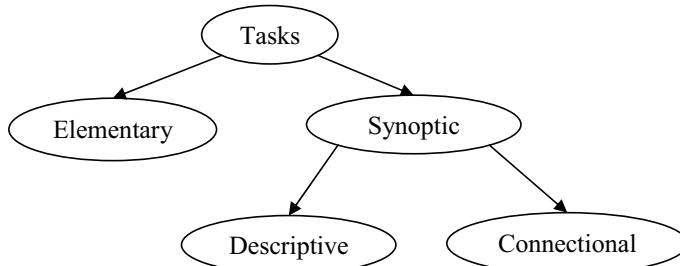
So, what should we do with connection discovery tasks? Should we treat them as a separate high-level partition, i.e. a category of the same level as elementary and synoptic tasks? Or should we subdivide synoptic tasks into “basic” and “advanced”, where the “basic” tasks are the tasks summarised in Table 3.5 and “advanced” tasks are connection discovery tasks? Or should we at all make any distinction between “ordinary” synoptic tasks and connection discovery tasks?

In our opinion, a classification cannot be right or wrong. It can be convenient or inconvenient, productive or useless. So, the question is: what is productive and convenient for our purposes?

Subsuming connection discovery tasks under the category of synoptic tasks is definitely convenient: in this case, all the subcategories defined for synoptic tasks will automatically apply to connection discovery tasks. This appears quite logical; for example, there may be a task of searching for a particular dependency pattern or of comparison of two correlation patterns. But is this solution productive?

The main purpose of our investigation of task typology is to understand what essential criteria are used or should be used in choosing or designing tools for EDA. We want to define which tools support which tasks, how to determine which tasks a particular tool is capable of supporting, and how to translate the properties of the tasks that a designer wishes to support with a new tool into the qualities and functions that this tool needs to have, and, probably, the building blocks that could be used for the construction of the tool. From this point of view, it is productive to distinguish between tasks if they require different tools to accomplish them.

As we have demonstrated, connection discovery tasks may require quite different tools from behaviour characterisation tasks. Therefore, it is more productive to consider connection discovery tasks separately from “basic” synoptic tasks, which will be called *descriptive synoptic tasks* from now on. On the other hand, since connection discovery tasks are, by their nature, synoptic tasks, we prefer to subdivide the top-level category of synoptic tasks into descriptive synoptic tasks and connectional synoptic tasks rather than to introduce a third top-level category along with elementary and synoptic tasks. So, the general classification scheme appears as shown in Fig. 3.22.



**Fig. 3.22.** General outline of our task classification framework

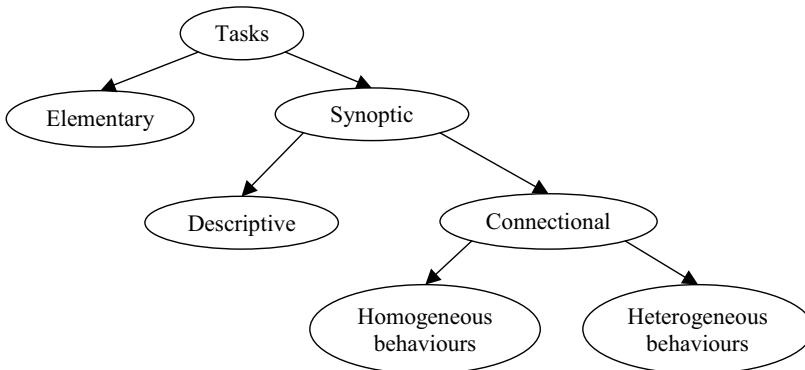
Having introduced this scheme, we would like to apply our earlier subcategorisation of synoptic tasks (i.e. into pattern identification, pattern comparison, and so on) to both varieties of synoptic tasks. Next, we would like to restrict the use of the expression “connection discovery” to the group of connectional tasks corresponding to the behaviour characterisation group of descriptive tasks. The other subcategories of connectional tasks involve search (recognition) or comparison rather than discovery.

We do not see a need for describing in detail how each subcategory defined for descriptive tasks would translate into the corresponding subcategory of connectional tasks. First, we hope that the general principle is clear, and interested readers could do this exercise by themselves. Second, we think that, although all subcategories are theoretically conceivable for connectional tasks, connection discovery *per se* is the subcategory most relevant to the practice of data analysis. This is also the most challenging subcategory: it is much more difficult to discover connection patterns than, for example, to compare such patterns once they have been discovered.

Let us now turn to the example  $\rho_2$ , which is a task that deals not only with stork movement but also with other phenomena such as relief, climate, and vegetation. How does this task fit into the overall framework?

It was noted earlier that a phenomenon never exists in isolation, out of context, and that the context together with this phenomenon could be regarded as a more complex phenomenon. In accordance with this assumption, we could treat stork movement as a part of a broader phenomenon of seasonal migration, which includes, besides the movement itself, the relief, land cover, climate, and so on. Hence, the discovery and, possibly, further investigation of external linkages are, in principle, analogous to the exploration of internal connections between parts of a phenomenon. However, there is a difference. When just the movement of storks is analysed, the analyst tries to relate homogeneous pieces, while in the exploration of the links between the movement and the climate or the relief, heterogeneous components need to be related.

Hence, it may be useful to subdivide the category of connectional tasks according to whether they deal with homogeneous parts of the same phenomenon, or with heterogeneous phenomena or heterogeneous parts of a complex phenomenon. So, the scheme presented in Fig. 3.22 could be modified, as is shown in Fig. 3.23.



**Fig. 3.23.** The modified scheme of task typology

We have already introduced formal notations for both varieties of relational behaviours, i.e. homogeneous and heterogeneous. To avoid forcing readers to search through the preceding paper for these notations, we shall repeat them here:

$$\rho(f_1(x), f_2(x) \mid x \in \mathbf{R}) \quad (3.43)$$

$$\rho(f_1(x), f_2(z) \mid x \in \mathbf{R}, z \in \mathbf{Z}) \quad (3.44)$$

$$\rho(f(x), f(x') \mid x \in \mathbf{R}_1, x' \in \mathbf{R}_2) \quad (3.45)$$

The expression (3.45) represents the mutual behaviour of homogeneous parts of the same phenomenon, while (3.43) and (3.44) stand for the mutual behaviour of heterogeneous phenomena or components. The difference between (3.43) and (3.44) is in the reference sets of the phenomena, which may be the same (3.43) or different (3.44).

The division in accordance to the task target (i.e. into the tasks of connection pattern discovery, connection pattern search, connection pattern comparison, etc.) must then be applied to each of the descendants of the node “Connectional”, i.e. to “Homogeneous behaviours” and “Heterogeneous behaviours”. We shall not elaborate the diagram in Fig. 3.23 to represent the full scheme – it would be too cumbersome. We hope, however, that it is clear to readers how we can organise the vast variety of possible tasks into a manageable system of categories and task properties.

### 3.6 Completeness of the Framework

➤ The goal of our investigation in the realm of data analysis tasks was to build an operational task typology, which could be productively and conveniently used for our further purposes. It is important, first, to relate the types of tasks to the structure of the data under analysis, and second, to ensure that the typology is complete. But how can the second requirement be fulfilled? There is no practical possibility to consider all tasks in order to organise them into groups by similarity. The solution that we found and applied was a sort of modelling: we represented data and tasks by formal expressions and then contemplated and manipulated those expressions.

Our model of a task is a structure consisting of two parts: the target and the constraints. The target specifies what information needs to be obtained, and the constraints specify what conditions this information needs to fulfil. The target and constraints can also be viewed as unknown and known information, respectively; the goal is to find the initially unknown information corresponding to the known information.

Our model of data and the underlying phenomena is a function that specifies the correspondence between the referential and characteristic components of the data. The referential components are represented as independent variables, and the characteristic components as dependent variables. We refer to this function as the *data function*.

Both models are sufficiently abstract to encompass all possible datasets and all possible tasks. The question is: did we link these models together properly? Did we lose the initial completeness? Let us review how we performed the linkage and elaborated it.

Our typology of tasks is based upon the types of the unknown and known information, which are defined in terms of the structural components of data and the relations between and within these components. In other words, the linkage between the task model and the data model was performed by filling the slots “target” and “constraints” in the task model with various information items, defined on the basis of the data model. Assuming that the model of data is complete, we need to answer the following questions:

- Did we account for all the possible types of information items that may participate in task targets and constraints?
- Did we account for all possible variants of the way of filling the slots?

Let us first try to answer the question concerning the possible types of information items, which need to be defined in terms of the structural components of data and the relations between and within these compo-

nents. The structural components of any data are references and characteristics. We state that references and characteristics can be dealt with on two levels: the level of individual elements and the level of sets. This is a rather straightforward dichotomy; no other options appear to be possible. On the level of individual elements, the data function defines a set of relations that link references (i.e. elements of the reference set) to corresponding characteristics (i.e. elements of the characteristic set). On the level of sets, subsets of the reference set are dealt with, rather than individual elements. For any subset, there is a corresponding configuration of characteristics determined by the data function. This configuration is referred to as the “behaviour”. On the level of elements, individual references and characteristics can appear in task targets or constraints. On the level of sets, (sub)sets of references, and behaviours (i.e. configurations) of characteristics appear instead of individual references and characteristics.

Either on the level of elements or on the level of sets, an analyst can focus his/her attention on a specific item (i.e. an element, a set, or a behaviour) and on what corresponds to it according to the data function. This may be called the “absolute approach”. Another possibility is the “relative approach”, where an item and what corresponds to it are considered in relation to (i.e. compared with) another item or a group of items and what corresponds to them. The division into “absolute” and “relative” seems to be an exhaustive dichotomy.

When the relative approach is taken, the analyst needs to deal with relations between homogeneous objects: between references, between characteristics, between subsets of references, or between behaviours. These relations may appear in the targets or the constraints of data analysis tasks, which may be called “relational”.

The relations that can potentially exist between objects are diverse, for the various types of objects. For elements of the reference or characteristic set, the possible relations are determined by the properties of this set. Thus, the relations “same” and “different” exist in any set, ordering relations (such as “less than” and “greater than”) exist in an ordered or at least partly ordered set, and distance relations exist in sets in which distances are defined. Between sets, there may be the relations “same” and “different”, “overlapping” and “not overlapping”, and “included” and “not included”. In an ordered set, there may be ordering relations between subsets, for example relations between time intervals: “before”, “after”, and so on. In a set with distances, there may be distance relations between subsets.

In fact, we are not concerned about the completeness of this list of relations, since we do not use the types of relations as a basis for task differentiation. Instead, we divide tasks into absolute (lookup and pattern identification) and relational (comparison, relation-seeking, and connection dis-

covery) tasks. The difference is that relational tasks involve relations between homogeneous objects (i.e. references, characteristics, sets of references, or behaviours), while absolute tasks deal only with relations between heterogeneous objects defined by the data function, i.e. relations between references and characteristics and between sets of references and behaviours. Then, relational tasks are further subdivided according to the types of object linked by the relations, i.e. whether these are references, characteristics, sets of references, or behaviours. As we have just demonstrated, this list can be considered as an exhaustive enumeration of the categories of objects dealt with in data analysis. Hence, the subdivision of relative tasks on this basis does not introduce incompleteness into the framework.

Tasks dealing with relations between behaviours are distinguished further according to the type of these relations. Specifically, we consider similarity/difference relations separately from correlations, influences, or structural relations. While similarity/difference relations are covered by behaviour comparison tasks within the category of descriptive synoptic tasks, we have introduced the category of connectional synoptic tasks to deal with the latter group of relations. We need to ensure that the coverage of the space of possible relations between behaviours is complete.

Actually, we have just separated similarity/difference relations from other possible relations between behaviours, which are often referred to as “inherent relations”, in contrast to the mostly “exterior” nature of similarities and differences.<sup>6</sup> Although we have listed some types of inherent relations, specifically correlation, influence, and structure, we shall restrain ourselves from presuming that this list is complete. In fact, it is not necessary to ensure that this list is complete, since we do not classify tasks according to the types of inherent relations. It is only important, for our purposes, to distinguish between inherent and exterior relations. The category of exterior relations consists of similarity and difference relations, and the category of inherent relations includes all other relations. Hence, the division into exterior and inherent relations can be regarded as exhaustive.

The next division is the division of connectional tasks into tasks dealing with homogeneous components and those dealing with heterogeneous components. There seem to be no problems with this division: this is an apparently exhaustive dichotomy.

Let us recapitulate our reasoning concerning the types of the information items that can participate in the targets and constraints of possible tasks. We have shown that these types of information are references, characteristics, reference subsets, behaviours, and relations (between homoge-

---

<sup>6</sup> In the next section, we provide some rationale for this division.

neous objects), and have demonstrated that this list of possible candidates for filling the slots in a task model is exhaustive. Then, we undertook an investigation into the possible categories of relations that need to be accounted for. We have demonstrated that our differentiation of relations according to the types of objects that they are supposed to relate, and our further division of relations between behaviours into exterior and inherent can be judged as complete and adequately reflected in the task typology.

Now, let us switch to the issue of slot-filling. First of all, tasks are distinguished according to what is in the target. This may be one of the following: a characteristic, a reference, a reference set, a behaviour, or a relation. When two or more items are referred to in the target, this means that the task actually consists of several subtasks and, hence, it is not necessary, in principle, to consider it as a separate type. However, certain types of compound tasks may be introduced for the sake of convenience, and we have done so for comparison and relation-seeking tasks.

Tasks are also distinguished according to what is used in the constraint part. This may be any of the five types of information items listed above. In principle, a task may have several constraints. However, in this case it can be represented as a combination of several subtasks, so that the union or intersection of results of these subtasks produces the final result of the original task. Hence, it is sufficient to consider tasks where a single item is used as the constraint.

In order to check if we have accounted for all possible combinations of types of targets and constraints, we have filled the matrix presented in Table 3.6. It appears that all meaningful combinations have been accounted for. Hence, we can conclude that our task typology is complete with respect to the possible variants of slot-filling in the task model.

The task categories listed in Table 3.6 are further subdivided according to whether they deal with a single attribute or with several different attributes, and with the same reference/reference set or with different references/reference sets. These divisions are exhaustive dichotomies and, hence, do not destroy the completeness of the framework.

Overall, we can conclude that our task typology is complete, with respect to the models of data and tasks that it is based upon. ↵

**Table 3.6.** Verification of the completeness of our task classification according to the types of information used in the target and in the constraints<sup>a</sup>

Constraint → ↓Target	Individual reference(s)	Individual character- istic(s)	Reference set	Behaviour	Relation
Individual refer- ence(s)	—	Inverse lookup	— —	— —	—
Individual character- istic(s)	Direct lookup	—	— —	—	—
Reference set	—	Inverse lookup	—	Pattern search	Relation- seeking
Behaviour	— —	—	Behaviour charac- ter- isa- tion	—	*
Relation	Inverse comparison	Direct compar- ison	Inverse compar- ison	Direct behaviour compar- ison	—

<sup>a</sup> — The data function is not involved;

— — Incompatibility between the elementary and synoptic (set) levels.

\* It is possible to formulate a task in the form “Find a behaviour related to the given behaviour in the specified way” (e.g. similar, different, or opposite).

However, this actually requires finding a reference set on which such a behaviour is based. Hence, this formulation of the task implicitly includes a reference set in its target and is therefore equivalent to a relation-seeking task.

### 3.7 Relating Behaviours: a Cognitive-Psychology Perspective

In this section, we shall try to explain why we treat similarity/difference relations between behaviours differently from other types of relations, such as correlations, influences, or structural relations.

Like all things in the world, behaviours can be compared to detect their similarities and differences. These types of relations between behaviours do not differ in principle from the “same” and “different” relations be-

tween elements and between sets. Therefore, we handle these relations in the same manner as relations between elements and between sets.

At the same time, behaviours are more complex items than elements and sets. Behaviours are configurations of characteristics, and an analyst may be interested not only in observing and describing *how* the characteristics are configured but also in trying to find an explanation of *why* they are so configured. For this purpose, the analyst may try to find the internal mechanisms that determine this configuration or to attribute the configuration to influences exerted by other behaviours. These internal mechanisms are, actually, nothing other than interactions between structural parts of the behaviour.<sup>7</sup> These internal or external interactions are quite different from the relations of similarity and difference. Therefore, we prefer to consider the purely “exterior” relations of similarity and difference separately from the more inherent relations, which account for the behaviours observed.

This division can also be viewed from another perspective, specifically, from the perspective of the major goal of a task, which may be either to just *describe* data or to *understand* data. We divide tasks into descriptive and connectional tasks, respectively. The second category is called “connectional” because it deals with essential connections between phenomena, but it might also be called “comprehensive” in the sense that the purpose of such tasks is to “comprehend”, i.e. “to understand the nature or meaning of; grasp with the mind” (Random House 1996). However, the word “comprehensive” also has other meanings, and therefore we prefer to use the term “connectional”, which explicitly indicates that these tasks are about connections.

In a brilliant book by Rudolf Arnheim called *Visual Thinking* (Arnheim 1997), we have found many ideas that seem to us extremely akin to our own thoughts. Some of these ideas are quite relevant to the topic of differentiating exterior and inherent relations. We would like to convey these ideas to readers, since they provide an additional perspective on the topic, specifically, the cognitive-psychology perspective.

Arnheim writes, “To understand an event or state of affairs scientifically means to find in it a pattern of forces that accounts for the relevant features of the system under investigation” (Arnheim 1997, p. 193) (we call such forces “inherent relations”, or “connections”). The key role in this belongs to abstraction, i.e. “the act of considering something in terms of general qualities or characteristics” (Random House 1996). Arnheim argues against the understanding of an abstraction as the sum of the properties

---

<sup>7</sup> Since a behaviour is a configuration, it necessarily has its internal structure. Non-elementary structural parts of a behaviour can, in turn, be considered as behaviours.

that a number of particular instances have in common. He has no doubt that abstraction is not based on grouping things according to their common features. He writes:

Presumably there are no two things in this world that have nothing in common, and most things have a great deal in common. Suppose now that every community of traits would induce us to group the corresponding things under a concept. Obviously, the result would be an incalculable number of groupings. Each individual thing would be explicitly assigned to as many groups as there are possible combinations of its attributes. A cat would be made to hold membership in the associations of material things, organic things, animals, mammals, felines, and so forth, all the way up to that exclusive club for which only this one cat would qualify. Not only this, but our cat would also belong among the black things, the furry things, the pets, the subjects of art and poetry, the Egyptian divinities, the customers of the meat and canning industries, the dream symbols, the consumers of oxygen, and so on forever. (Arnheim 1997, pp. 157–158)

In the context of our work, this means that an explorer cannot gain understanding of a phenomenon by merely comparing instances or parts, noting similarities and differences, and grouping similar things together.

People do not group things arbitrarily, on the basis of any commonalities noted, but do this according to their particular interests, which determine the crucial attributes to be used for groupings. Arnheim says, “...quite frequently we make groupings on the basis of one distinguishing trait alone. Flammable or non-flammable – nothing else may matter” (Arnheim 1997, p. 158). However, the identification of such crucial attributes, which need to be distilled from the multitude of features, necessarily requires abstraction:

The grouping of instances, allegedly the necessary preparation for abstraction, must be preceded by abstraction, because from where else would the criteria for selection come? Before one can generalize, one must single out characteristics that will serve to determine which things are to belong under one heading. (Arnheim 1997, p. 161)

Hence, abstraction has to take place before grouping rather than result from it: “an abstract concept, supposed to be the fruit of generalization, turns out to be its necessary prerequisite” (Arnheim 1997, p. 159). This corresponds to our idea that connectional tasks, aimed at understanding phenomena and hence indispensably involving abstraction, cannot be defined in terms of comparing behaviours and detecting similarities and differences between them. Connectional tasks are not, in general, based on comparison; rather, they are based on an “abstractive grasp of structural features”, in Arnheim’s terms. Thus, a scatterplot does not expose to us the similarities and differences between behaviours of two attributes. Instead,

our perception combines the individual dots displayed in such a plot into a unitary shape, and this shape tells us whether the attributes are related or not. Seeing a shape instead of a multitude of dots is an instance of abstraction, which is by no means based on comparing and revealing common features.

Where, then does, abstraction come from? Arnheim argues that abstraction is *inherently involved in perception*: “There is no way of getting around the fact that an abstractive grasp of structural features is the very basis of perception and the beginning of all cognition” (Arnheim 1997, p. 161). In explaining this apparent paradox, Arnheim refers to the work of Henri Bergson, who proposes that perception can be seen as an instrument of an organism, developed during phylogenetic evolution as a means of discovering the presence of what is needed for survival and being alerted to danger.

These needs, argues Bergson, refer to kinds of things, to qualities rather than to particular individuals. What attracts the herbivorous animal is herbage in general, “the colour and the odour of herbage, sensed and submitted to as forces...” The precise distinction of individual objects, he says, is “un luxe de la perception” – a luxury of perception. (Arnheim 1997, p. 160)

Hence, “high generality is a quality of perception from the very start” (Arnheim 1997, p. 166); “percepts are generalities from the outset, and it is by the gradual differentiation of those early perceptual concepts that thinking proceeds towards refinement.” (Arnheim 1997, p. 186). There is a clear parallel between this statement and Shneiderman’s Information Seeking Mantra: “Overview first, zoom and filter, and then details-on-demand” (Shneiderman 1996).

“However”, proceeds Arnheim, “the mind is just as much in need of the reverse operation. In active thinking, notably in that of the artist or the scientist, wisdom progresses constantly by moving from the more particular to the more general” (Arnheim 1997, p. 186). As an example of such generalisation, Arnheim refers to the development of the theory of the conic sections, which united such different shapes as circles, ellipses, parabolas, and hyperbolae into a single geometrical family. This generalisation could not result from revealing common traits in these figures, and the new concept does not consist of such common traits.

Something fundamentally different took place. Those basic geometrical figures had been satisfactory, self-contained entities since antiquity. Now a new perceptual entity, the sectioned cone, offered itself as a new whole, into which the formerly isolated figures could be fitted as parts. A new understanding of their structural nature was brought about by their relations to what turned out to be their neighbours in a continuous sequence of shapes and by their locations in the total

perceptual system of the cone. Generalization, then, was an act of restructuring through the discovery of a more comprehensive whole. (Arnheim 1997, pp. 186–187).

This is an example of a relation quite different from similarity/difference relations. This relation was discovered not by means of comparison but rather by means of manipulating distinct patterns and arranging them in order to make them fit together.

True generalization is the way by which the scientist perfects his concepts and the artist his images. It is an eminently unmechanical procedure, requiring not so much the zeal of the census-taker, the bookkeeper, or the sorting machine as the alertness and intelligence of a functioning mind. (Arnheim 1997, p. 187)

Hence, there are cognitive-psychological grounds for our separate treatment of similarity/difference relations, on the one hand, and inherent (causal, structural, etc.) relations, on the other hand. Not only may these two groups of relations require different tools for detecting them, but also different psychological mechanisms are involved in discovering and handling these relations.

This discussion, with massive citation of Arnheim, should not be understood as nullifying the role of comparison tasks in exploratory data analysis. We think that comparison tasks are very important, even though they may be insufficient for understanding a phenomenon. Again, we find relevant reasoning in the same book by Arnheim:

Experience indicates that it is easier to describe items in comparison with others than by themselves. This is so because the confrontation underscores the dimensions by which the items can be compared and thereby sharpens the perception of these particular qualities. However, the procedure has its dangers. It is easier to describe the United States by comparing it with China than by itself without such reference; but the comparison highlights characteristics quite different from the ones to be gotten from a comparison with, say, France, and is therefore arbitrary (Arnheim 1997, p. 63)

This means that, before comparing, an explorer needs to have a general idea of what should be compared and how this should be done. This general idea comes from Shneiderman's "overview", the beginning of any analysis, which is aimed at gaining "an abstractive grasp of structural features" (Arnheim). It is this preliminary abstraction that guides all further activities and at the same time is elaborated, rectified, and completed in the course of those activities.

### 3.8 Why Tasks?

This chapter was meant to define what the possible tasks, or questions, in exploratory data analysis are. After having done this, we would like to return to the discussion presented in the introduction of this book, specifically, whether tasks, or questions, are relevant to EDA, or whether this activity can instead be imagined as just observing without any clear purpose, rambling through the data in the hope of encountering something “interesting”, which might provide food for thought and lead to a sudden “insight”.

We must admit that the founder of exploratory data analysis, John Tukey, did not make any clear point concerning this issue. Since EDA is tightly linked to visualisation, we tried to find out what researchers in visualisation and developers of visualisation tools think about the role of tasks. Our conclusion from numerous discussions is that opinions differ.<sup>8</sup> Some of our colleagues believe that having a defined task is not (or not always) necessary in information visualisation. Others are convinced that tasks always exist, explicitly or implicitly, even when an explorer seems “just to look” at data. However, these tasks will be different from the examples given by Casner, who demonstrates that the same data need to be represented in different ways in order for different tasks to be performed effectively (Casner 1991).

Casner considers examples of tasks such as planning a journey from city A to city B with a stopover in city C (where one has an appointment at a particular time) and finding the cheapest flight or the most direct travel route. For each task, he proposes a graphical display that allows the task to be performed effectively. There is no “ideal” graphic realisation of the data suitable for all purposes. While the argument given by Casner is rather convincing, the tasks he considers seem to have little to do with exploration: namely, there is no attempt to grasp inherent characteristics of unfamiliar data and gain knowledge about the underlying phenomena.

In this chapter, we have considered a very wide range of tasks, from very specific ones such as “where was the stork A on 1 September?” to tasks as general as “what are the generic principles of stork behaviour during the seasonal migration?” While it could be argued whether the first example is pertinent to exploratory data analysis, the second question has a clear “exploratory flavour”. Our framework embraces both very specific

---

<sup>8</sup> Some outcomes from a discussion between developers of visualisation tools concerning the factors that motivate and influence tool design are presented in Andrienko et al. (2004).

questions and very general ones, with a number of intermediates between these extremes.

We count ourselves among the advocates of the task-driven nature of exploratory data analysis. Together with our allies, we argue that, usually, an explorer does not only *look at* data but also *looks for* something “interesting”. This may be, for instance, a salient pattern in a spatial distribution, a local anomaly, some indication of unusual behaviour, or an indication of a possible dependency between phenomena or processes. In this view we are, actually, unanimous with our opponents. However, and this is the difference, we understand “interestingness” as relevance to the major research question that the explorer puts to himself/herself, or, in other words, the primary task of data analysis, the motive for doing the analysis. According to Rudolf Arnheim, “The mind is always steered by purpose” (Arnheim 1997, p. 162).

This primary task may be rather general, such as the question about the principles of the migratory movement of storks. The explorer needs the available data to be represented so that he/she can overview it and detect “interesting”, i.e. potentially relevant, features. When such relevant features are detected, the analyst will typically try to compare them and investigate each of them in more detail. Here, exploratory tasks of lower generality levels come into play. This corresponds to the principle “Overview first, zoom and filter, and then details-on-demand”, known as the Information Seeking Mantra (Shneiderman 1996).

Reformulating this in the terms of our framework, exploratory data analysis is an investigation into the essential, generic properties of the data function, which defines the correspondence between references and characteristics and represents a certain phenomenon. From the beginning, the analyst has a general (synoptic) task related to the whole reference set. In most cases, this task is aimed at identifying, describing, and explaining the behaviour of the data function, and hence the underlying phenomenon. For this purpose, the explorer tries to grasp the essential features of the behaviour. In many cases, it turns out to be impossible to represent the data in such a way that the whole behaviour can be grasped at once. In such cases, the analyst needs to “cut” the behaviour into perceivable parts and investigate each of the parts. The results of these partial studies will then need to be synthesised into a coherent mental model of the behaviour as a whole. Another case where such partitioning of a behaviour is meaningful (and sometimes indispensable) is when the behaviour is uneven throughout the reference set. Then, the analyst “cuts” the behaviour into more or less uniform “pieces” and investigates each piece individually.

In any case, the exploration proceeds as an interaction between the top-down and bottom-up processes of analysis and synthesis. Here, the word

“analysis” is used in the general sense of “separating some material or entity into its constituent parts or elements” (Random House 1996). In our context, this corresponds to partitioning of the overall behaviour into portions. “Synthesis” corresponds to the construction of a model of the overall behaviour, i.e. a system of knowledge about the phenomenon, from the partial knowledge derived from the investigation of the portions that the overall behaviour was divided into. Thus, uniting the circle, ellipse, parabola, etc. into the family of conic sections is an example of the kind of synthesis we mean.

Dividing the overall behaviour into perceivable portions may be done in various ways, depending on the structure and volume of the data. Thus, the data may be multidimensional, i.e. the reference set may contain several referrers. In this case, the explorer may need to consider various aspectual behaviours, which can be viewed as “projections” of the overall behaviour. This corresponds to reducing the generality of the primary task to a set of subtasks addressing subsets of the entire reference set, where the subsets are defined by fixing values of some of the referrers, thus reducing the number of variable dimensions. Another case, which has already been mentioned, is the division of the reference set of a “patchy” behaviour. Such a division is done without reducing the dimensionality of the reference set. Yet another case is a behaviour involving several attributes. In such a case, an explorer often starts with separate investigations of the behaviour of each attribute, and then tries to derive conclusions concerning the overall behaviour.

Whatever division is applied, the analyst needs to perform tasks of characterising the partial behaviours that the overall behaviour is divided into (behaviour characterisation). In so doing, the analyst may subdivide the partial behaviours into yet smaller pieces. After a partial behaviour has been characterised, i.e. approximated by an appropriate pattern, the analyst may look to see whether the same or a similar pattern occurs in any other subsets of references (pattern search) and, if so, determine how these reference subsets are related to the subset initially considered (inverse behaviour comparison). It is also quite appropriate to directly compare partial behaviours in different reference subsets (direct behaviour comparison) and to look for particularly related patterns in particularly related subsets, for example opposite trends in contiguous time intervals (relation-seeking). All these tasks involving relations of similarity and difference between the partial behaviours are eventually aimed at approximating the overall behaviour, probably by some composite pattern. At the same time, the explorer tries to discover essential linkages between the partial behaviours, such as causal or structural connections, which could explain the overall behaviour (connection discovery).

From the very beginning, the explorer starts to build some concept of the data, which may initially be very vague. By investigating each relevant feature, the analyst verifies, amends, and refines this concept. The analyst may need to “dive” quite deep into the data, in order to ensure that the concept is sufficiently precise and valid. In this process, elementary tasks may also be actively involved. Thus, the analyst usually pays attention to outliers, i.e. individual data elements that disrupt or confuse patterns, for example a municipality with a low percentage of children inside a cluster of municipalities with high percentages of children. To investigate such cases and find an explanation for them, the explorer will need to perform various elementary tasks, such as direct and inverse lookup and comparison. Hence, although elementary tasks play a subordinate role in exploratory data analysis, an analyst needs tools that support both synoptic and elementary tasks.

We are far from thinking that any analyst consciously divides data exploration activities into different types of tasks and plans the whole process as a combination of top-down and bottom-up approaches. An explorer may be unaware of these tasks and be guided by pure intuition, by general principles such as “overview first, zoom and filter, and then details-on-demand”, or by examples. However, anyone who attempts to create a tool for exploratory data analysis must explicitly consider the tasks and deliberately design any instrument so that it can support the observation of distributions and behaviours, expose patterns, and facilitate detection of relationships. Taking into account the concurrency of exploratory tasks, it may be inappropriate to follow the approach of Casner, who advocates building a separate graphical realisation for each task. Instead, one should try to design a tool that supports a range of tasks. When a single tool appears to be insufficient, several interlinked, complementary instruments may be appropriate. In any case, we are deeply convinced that a tool designer needs to know what exploratory tasks exist and to be able to find methods of supporting them.

## 3.9 Other Approaches

There are various approaches to defining possible tasks, and numerous task taxonomies have been suggested. We are not going to criticise any of these taxonomies. As we have already said, a classification cannot be right or wrong. Any classification is right to the extent of its serving the purposes for which it was devised. We cannot restrain ourselves from citing our beloved book by Arnheim:

For example, cases can be cited in which human beings are classified by size, weight, income, skin color, number of gold teeth, or their ideas about the supernatural – no criterion of selection seems ineligible, each may be justified by the proper occasion, and what serves one purpose or direction of interest may be absurd for another (Arnheim 1997, p. 159)

Therefore, we are not going to undertake a detailed analysis of every existing taxonomy in order to reveal its weaknesses in comparison with our superb task typology. Instead, we first refer to the roots of our ideas and, second, propose a brief overview of the other approaches to task classification, motivated by purposes different from ours; as would be natural to expect, those approaches serve those purposes better than our typology would do.

As we have made clear from the very beginning, our task typology is based on the ideas expressed by Bertin. We have already discussed those ideas in much detail, and there is no need to describe them again. Our conception was also greatly influenced by the work of Klir (1985), although this work is not concerned with task classification but rather with developing a general framework for exploring and describing various phenomena on the basis of the treatment of any phenomenon as a system. Thus, our division of data components into references and characteristics and our formal view of a dataset as a function that matches references with characteristics come from Klir's work. Moreover, Klir also uses the notion of a behaviour (of a system), and it is rather close, although not identical, to our notion of a behaviour. Klir defines a behaviour as “a simple characterisation of the overall support-invariant constraint among variables”. Let us recall that Klir's “support” corresponds to the reference set in our terminology. “Constraint among variables” can be understood as a statement concerning relations between components of the data, which may be different attributes or values of attributes corresponding to different references. This statement needs to be support-invariant, i.e., in our terms, be true throughout the whole reference set.

Strictly speaking, this notion of a behaviour corresponds to our notion of a pattern approximating the overall behaviour of the data function on the entire reference set. We prefer to use the term “pattern” rather than “behaviour” because we understand the behaviour as something objective, existing independently of the views of the analyst. The goal of the analyst is to understand the behaviour and approximate it by an appropriate pattern, which can be seen as a “support-invariant constraint among variables”.

Besides these apparent links, Klir's work has also influenced us in another way. It served for us as an example of a productive abstraction, of reasoning about data analysis on a high level of generality. Basically, we

followed Klir's approach in our study, but our study had a different direction and therefore led to different results. A small deviation from the "model" is that we did not want to make our scheme as formal and mathematical as Klir's framework. To our taste, his framework is too formal and therefore hard to perceive and understand. We hope that, despite the use of some formal notation, our scheme is still understandable and does not require a solid mathematical background.

A key feature that differentiates our typology from many others is that it defines task types in terms of structural components of the data under analysis. However, some other typologies take the same approach. Thus, Peuquet (1994) considers spatio-temporal data as consisting of three components, namely, space (*where*), time (*when*), and objects (*what*). This view is well known as the "triad model" of spatio-temporal data. In accordance to the three components, Peuquet defines three basic types of possible questions about such data:

- *when + where → what*. Describe the objects or set of objects that are present at a given location or set of locations at a given time or set of times.
- *when + what → where*. Describe the location or set of locations occupied by a given object or set of objects at a given time or set of times.
- *where + what → when*. Describe the times or set of times when a given object or set of objects occupied a given location or set of locations.

This classification evidently parallels the notion of "*question types*" introduced by Bertin, and hence is quite close to our ideas. However, unlike Peuquet, we do not restrict our framework to only spatio-temporal data. Although such data are our primary interest, we did not feel a need to narrow the scope of data types to be considered in order to obtain useful results. We are sure that the generality of our scheme does not diminish its practicality (nevertheless, we shall have to check this in the following chapters).

MacEachren (1995) and Kraak et al. (1997) classify the possible questions concerning spatio-temporal data into seven query types, addressing the existence of an entity (if?), its location in time (when?), its duration (how long?), its temporal texture (how often?), its rate of change (how fast?), the sequence of entities (what order?), and synchronization (do entities occur together?). These types can be viewed as an elaboration of a more general task "describe the times or set of times when a given object or set of objects occupied a given location or set of locations" (*where + what → when*) in the classification suggested by Peuquet.

As we mentioned earlier, Bertin does not explicitly consider tasks involving comparison and, more generally, relations. Blok (2000) uses a distinction between exploratory tasks of identification and comparison as one of two orthogonal dimensions for differentiating questions that may arise in monitoring spatio-temporal changes. “Comparison” is treated in a broader sense than just discovering similarities and differences. It also includes detecting relationships between processes, in particular, cause–effect relationships. The second dimension considered by Blok is the length of the time series to be analysed. Thus, questions about trends (identification) or cause–effect relationships (comparison) can only be answered when sufficiently long time series are available. In our opinion, this dimension roughly parallels the notion of reading levels.

Besides typologies based on the structure of the data, many other task typologies have been suggested in the areas of visualisation and human–computer interaction. In order to understand their differences better, we have made an attempt to classify these typologies. We have considered them from the perspective of a generalised view of the process of data analysis adapted from Qian et al. (1997). Initially, an analyst has some *information need*. This need can be described by stating what is known and what needs to be found. This corresponds to our model of a task as consisting of constraints and a target. In order to find the information needed, the analyst plans a sequence of *operations* to be applied to the data. Finally, he/she tries to perform these operations using the available *tools*.

The different approaches to defining possible tasks refer to the different stages of this data analysis process. Thus, the approaches of Bertin and Peuquet can be classified as defining tasks in terms of information needs, i.e. they refer to the first stage. Among the typologies related more to the intermediate stage, some may be characterised as “user-centred”, i.e. they define possible tasks in terms of cognitive operations performed by a user, for example “locate”, “identify”, and “distinguish”. (Wehrend and Lewis 1990, Roth and Mattis 1990, Casner 1991, Robertson 1991, Jung 1995, Knapp 1995, Gahegan and O’Brien 1997, Zhou and Feiner 1998). Other researchers (Qian et al 1997) define tasks in terms of operations on sets: “union”, “intersection”, “selection”, etc. There are typologies that define tasks mostly as abstractions of tools and functions of existing GIS and can therefore be regarded as referring to the final stage of data analysis, i.e. choosing and applying tools. As an example, we can mention the typology described in Yuan and Albrecht (1995) and Albrecht et al. (1997), which introduces such tasks as “interpolation”, “buffer”, and “overlay”.

Since the user-centred task typologies are so numerous, and many of them have been suggested for purposes close to ours, let us discuss them in

more detail. First of all, let us consider the purposes that these task typologies were created for. Some authors created their typologies in order to provide guidelines for users to choose appropriate visualisation techniques. For example, Wehrend and Lewis (1990) define a set of tasks (the authors use the term “operations”) to be used for classifying visualisation techniques. The idea of Wehrend and Lewis is that a user who has a certain problem to be solved by means of visualisation breaks up this problem into subproblems, describes these subproblems in terms of the objects to be represented and the operations to be supported by the representation, selects applicable visualisation techniques according to an operation-based technique classification, and combines these representations into a composite representation for the original problem. Some other researchers use their typologies in knowledge-based software systems capable of automated design of visual representations according to task specifications provided by users.

These purposes are, to a certain extent, rather close to ours: we also aim at evaluating various tools (not only visualisation techniques) from the perspective of their capability to support various tasks pertinent to the process of exploratory data analysis. However, our ideas concerning the use of our task typology and the results of our tool evaluation are quite different from the idea of a user specifying his/her information needs in terms of a set of basic operations, followed by either manual or automatic design of the most appropriate data representation.

As we have discussed in the previous section, tasks in EDA cannot be specified precisely in advance. Some researchers even think that there are no tasks at all, and we actually agree with them in that an explorer does not consciously plans and fulfils any sequences of operations of the kind

### **Locate <?x, ?locator>**

(Zhou and Feiner 1998). Exploration starts with a very general task such as “What is the behaviour of this phenomenon?”, and all subsequent tasks on different abstraction levels emerge dynamically, depending on what the analyst finds (or does not find) in the previous steps and what attracts his/her attention. Hence, it cannot be expected that the analyst, before starting exploratory data analysis, will decompose his/her problem into subproblems and translate those subproblems into operations on data: the problem is too ill-defined to allow such decomposition. Consequently, one cannot build an “action plan” that could be used for the manual or automated design of an “ideal” data representation for the particular problem. Moreover, no “ideal” representation can exist even in principle: the initial problem is refined and modified in the course of analysis, and hence even an initially good representation may soon become unsuitable. Again, it is

hard to expect that the analyst will create a new “action plan” and redesign the representation each time.

A more realistic scenario is that the explorer is given a sufficiently powerful and flexible tool or set of tools (not just a certain data representation) that supports a wide range of tasks. Hence, it is the responsibility of tool designers to build such tools and provide them to data explorers. To do this, the designers need to know the range of tasks that need to be supported, and it was our goal to make such tasks explicit.

Hence, the purposes of our typology differ from the purposes of the operations-oriented typologies in the following respect: our typology is meant for initially vague and very general problems, which are dynamically refined and redefined in the process of analysis, whereas the operations-oriented typologies are intended for more specific problems, which can be clearly formulated and translated into collections of required operations before starting the analysis.

It may be argued that some operations-oriented typologies contain more general tasks than “locate” and “distinguish”. Thus, Gahegan and O’Brien (1997) describe a knowledge-based system for designing data visualisation for tasks called “exploration”, “search”, and “comparison”.

Although these names sound very abstract, in practice the word “exploration” is used merely in the sense that multiple data channels are displayed (data channels correspond to attributes or referrers in our terminology). Search implies that a single channel is visualised, and comparison means displaying two or more channels. Hence, the tasks actually define only the number of data components to be visualised simultaneously. Of course, such underspecified tasks cannot serve as a sufficient basis for choosing appropriate visualisation techniques. To cope with this difficulty, Gahegan and O’Brien assume that the user provides additional information concerning the relative importance of the data channels. Then, more important channels are represented by means of visual primitives with higher impact (i.e. they are more readily perceived by people).

The task typology suggested by Shneiderman (1996) can also be classified as operations-oriented, although it is quite different from the group of typologies just discussed. Actually, Shneiderman’s tasks look more like requirements to be fulfilled by designers of exploratory tools or like a specification of functions that should be present in such tools:

- *overview*: Gain an overview of the entire collection;
- *zoom*: Zoom in on items of interest;
- *filter*: Filter out uninteresting items;
- *details-on-demand*: Select an item or group and get details when needed;

- *relate*: View relationships among items;
- *history*: Keep a history of actions; in particular, this should support an undo function;
- *extract*: Allow extraction of subcollections.

Besides describing tasks, Shneiderman refers to examples of how these functions are implemented in existing software systems.

While Shneiderman's instructions are certainly valid and very useful, they do not provide an understanding of the possible information needs of a data explorer. In our opinion, only such an understanding (what might the purpose of the overview be, what are the potential items of interests, what relationships between what items might be relevant, etc.) can allow a tool designer to fulfil the requirements in an appropriate way, so that the explorer can really obtain what he/she needs.

The possible information needs of an explorer are the primary concern of our task typology, which hence refers mostly to the initial stage of data analysis with respect to the scheme of the data analysis process suggested by Qian et al. (1997).

It is also relevant to mention the classification of tasks adopted in data mining (see Fayyad et al. (1996), and Miller and Han (2001)):

- *Segmentation*: Partitioning data into meaningful groupings or classes. This includes two major subtasks:
  - *Clustering*: Determining a finite set of implicit classes that describe the data.
  - *Classification*: Finding rules to assign data items to pre-existing classes.
- *Dependency analysis*: Finding rules to predict the value of an attribute on the basis of the values of other attributes.
- *Deviation and outlier analysis*: Searching for data items that exhibit unexpected deviations or differences from some norm.
- *Trend detection*: Fitting lines and curves to data in order to summarise the database.
- *Generalisation and characterisation*: Obtaining a compact description of the database, for example as a relatively small set of logical statements that condense the information in the database.

For each of these tasks, there is a corresponding group of data-mining methods.

With respect to our typology, the tasks of data mining fit into such categories of synoptic tasks as behaviour characterisation, pattern search, and connection discovery. We count data mining-methods among the tools that

can be useful for accomplishing such tasks and will discuss these methods in their proper place.

## Summary

In this chapter, we have presented our task typology. It is based on two formal models:

- A data model, which represents a dataset as a function (in the mathematical sense) that defines the correspondence between the references and characteristics.
- A task model, which represents a task as a combination of a target and constraints, or unknown and known information. The goal is to find the initially unknown information corresponding to the known information. The target and constraints are viewed as slots to be filled with various types of information related to the structure of the data.

The major partitions that we used for classifying tasks are following:

- Tasks are divided into elementary and synoptic tasks according to the level of generality. Elementary tasks deal with individual elements of data, i.e. individual references and characteristics. Even if several elements are dealt with simultaneously, each of them is handled individually. Synoptic tasks deal with the dataset as a whole and its subsets, considered in their entirety. The principal notion on this level is the notion of a behaviour, i.e. a certain configuration of characteristics corresponding to a set of references. In contrast to the notion of a set, a behaviour implies the presence of a certain structure, for example its elements may be linked into sequences or by neighbourhood relationships. Synoptic tasks deal with identifying and understanding behaviours. Synoptic tasks play the primary role in exploratory data analysis, and elementary tasks are subordinate and are mostly used as subtasks of more general tasks.
- Synoptic tasks are subdivided according to their purpose into descriptive and explanatory, or connectional, tasks. The purpose of descriptive tasks is to identify and describe behaviours, as well as compare behaviours to identify their similarities and differences. The purpose of connectional tasks is to investigate behaviours and discover essential interrelations (or connections; hence the name of the task category) between behaviours or between structural components of the same behaviour. An analyst is interested in the types of connections pertinent to the nature of the behaviours, such as cause–effect relations or principles of internal or-

ganisation. We treat such connections as mutual behaviours of phenomena or of components of the same phenomenon.

- Tasks are distinguished according to the types of information referred to in the target and specified in the constraints. In elementary tasks, the target may refer to characteristics, references, or relations between characteristics or between references. In synoptic tasks, the target may refer to a behaviour (in particular, a mutual behaviour), a reference set, or a relation between behaviours or between reference sets.

The main purpose of our investigation into task typology was to understand what essential criteria are used or should be used in choosing or designing tools for exploratory data analysis. On the basis of our task typology, we intend to analyse the existing techniques for EDA to find out what tools can support what tasks. We wish then to generalise the results of our analysis into some fundamental principles, which would allow one to do the following:

1. Having a particular tool, determine what tasks it is capable of supporting.
2. Having a plan to support some tasks with a new tool yet to be designed, translate the properties of the tasks into a tool specification, i.e. the set of functions and characteristics that this tool needs to possess (plus, possibly, the building blocks from which to construct the tool).
3. Having some data to analyse and tasks to accomplish, find out which of the tools available are reasonable to apply, and in what combination.

In the next chapter, we are going to explain what we mean by “data analysis tools”, what categories of tools exist, and how each of these categories is used in exploratory data analysis.

## References

- (Albrecht et al. 1997) Albrecht, J., Jung, S., Mann, S.: VGIS: a GIS shell for the conceptual design of environmental models. In: *Innovations in GIS 4: Selected Papers from the 4th National Conference on GIS Research UK, GIS-RUK*, ed. by Kemp, Z. (Taylor & Francis, London 1997) pp. 154–165
- (Andrienko et al. 2003) Andrienko, N., Andrienko, G., Gatalsky, P.: Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing* **14**(6), 503–541 (2003)
- (Andrienko et al. 2004) Andrienko, G., Andrienko, N., Dykes, J., Gahegan, M., Mountain, D., Noy, P., Roberts, J., Rodgers, P. Theus, M.: Creating instruments for ideation: software approaches to geovisualization. In: *Exploring*

- (Geovisualization, ed. by Dykes, J., MacEachren, A., Kraak, M.-J. (Elsevier, Oxford 2005) pp. 103–125
- (Arnheim 1997) Arnheim, R.: *Visual Thinking* (University of California Press, Berkeley 1969, renewed 1997)
- (Bertin 1967/1983) Bertin, J.: *Semiology of Graphics. Diagrams, Networks, Maps* (University of Wisconsin Press, Madison 1983). Translated from Bertin, J.: *Sémiologie graphique* (Gauthier-Villars, Paris 1967)
- (Blok 2000) Blok, C.: Monitoring change: characteristics of dynamic geo-spatial phenomena for visual exploration. In: *Spatial Cognition II*, ed. by Freksa, C., Brauer, W., Habel, C., Wender, K.F., Lecture Notes in Artificial Intelligence, Vol. 1849 (Springer, Berlin, Heidelberg 2000) pp. 16–30
- (Casner 1991) Casner, S.M.: A task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics* **10**, 111–151 (1991)
- (Encarta 2004) Perception (psychology). In: *Microsoft® Encarta® Online Encyclopedia*, <http://encarta.msn.com> (Microsoft, Redmond, WA 2004). Accessed 28 Mar 2005
- (Fayyad et al. 1996) Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In: *Advances in Knowledge Discovery and Data Mining*, ed. by Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (AAAI Press/MIT Press, Menlo Park 1996) pp. 1–36
- (Gahegan and O'Brien 1997) Gahegan, M., O'Brien, D.: A strategy and architecture for the visualisation of complex geographical datasets. *International Journal of Pattern Recognition and Artificial Intelligence*, **11**(2), 239–261 (1997)
- (Heylighen 1997) Heylighen, F.: Occam's razor. In: *Principia Cybernetica Web*, ed. by Heylighen, F., Joslyn, C., Turchin, V. (*Principia Cybernetica*, Brussels 1997), <http://pespmc1.vub.ac.be/REFERPCP.html>. Accessed 28 Mar 2005
- (Jung 1995) Jung, V.: Knowledge-based visualization design for geographic information systems. In *Proceedings of the 3rd ACM International Workshop on Advances in GIS*, Baltimore, 1995, ed. by Bergougnoux, P., Makki, K., Pissinou, N. (ACM Press, New York 1995) pp. 101–108
- (Klir 1985) Klir, G.J.: *Architecture of Systems Problem Solving* (Plenum, New York 1985)
- (Knapp 1995) Knapp, L.: A task analysis approach to the visualization of geographic data. In: *Cognitive Aspects of Human–Computer Interaction for Geographic Information Systems*, ed. by Nyerges, T.L., Mark, D.M., Laurini, R., Egenhofer, M.J. (Kluwer, Dordrecht 1995) pp. 355–371
- (Koussoulakou and Kraak 1992) Koussoulakou, A., Kraak, M.J.: Spatio-temporal maps and cartographic communication. *The Cartographic Journal* **29**, 101–108 (1992)
- (Kraak et al. 1997) Kraak, M.-J., Edsall, R., MacEachren, A.M.: Cartographic animation and legends for temporal maps: exploration and/or interaction. In *Proceedings of the 18th International Cartographic Conference*, Vol. 1 (1997) pp. 253–261

- (Miller and Han 2001) Miller, H.J., Han, J.: Geographic data mining and knowledge discovery: an overview. In: *Geographic Data Mining and Knowledge Discovery*, ed. by Miller, H.J., Han, J. (Taylor & Francis, London 2001) pp. 3–32
- (MacEachren 1995) MacEachren, A.M.: *How Maps Work: Representation, Visualization, and Design* (Guilford, New York 1995)
- (Qian et al. 1997) Qian, L., Wachowicz, M., Peuquet, D., MacEachren, A.: Delining operations for visualization and analysis of space–time data in GIS. In: *Proceedings of the GIS/LIS '97*, Cincinnati, (1997) pp. 872–877
- (Peuquet 1994) Peuquet, D.J.: It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers* **84**(3), 441–461 (1994)
- (Random House) Random House Webster's *Unabridged Electronic Dictionary* (Random House, Broadway, NY 1996)
- (Robertson 1991) Robertson, P.K.: A methodology for choosing data representations. *IEEE Computer Graphics and Applications* **11**(3), 56–67 (1991)
- (Roth and Mattis 1990) Roth, S.M., Mattis, J.: Data characterization for intelligent graphics presentation. In: *Proceedings of SIGCHI'90: Human Factors in Computing Systems*, Seattle, 1990 (ACM Press, New York 1990) pp. 193–200
- (Shneiderman 1996) Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages* (IEEE Computer Society Press, Piscataway 1996) pp. 336–343
- (Tukey 1977) Tukey, J.W.: *Exploratory Data Analysis* (Addison-Wesley, Reading, MA 1977)
- (Wehrend and Lewis 1990) Wehrend, S., Lewis, C.: A problem-oriented classification of visualization techniques. In: *Proceedings of the First IEEE Conference on Visualization: Visualization 90* (IEEE, Los Alamitos 1990) pp. 139–146
- (Yuan and Albrecht 1995) Yuan, M., Albrecht, J.: Structural analysis of geographic information and GIS operations from a user's perspective. In: *Spatial Information Theory: a Theoretical Basis for GIS: International Conference COSIT'95, Proceedings*, ed. by Frank, A.U., Kuhn, W., Lecture Notes in Computer Science, Vol. 988, (Springer, Berlin, Heidelberg 1995) pp. 107–122
- (Zhou and Feiner 1998) Zhou, M., Feiner, S.: Visual task characterization for automated visual discourse synthesis. In: *Proceedings of the SIGCHI conference on Human factors in computing systems* (ACM Press, New York 1998) pp. 392–399

## 4 Tools

### Abstract

In this chapter, we make an inventory of the tools suitable for supporting exploratory data analysis. Our major point is that the primary tool for analysis is the human imaginative mind, and that all other tools are supplementary. Only the human mind actually does the analysis; the other tools supply it with the necessary material, appropriately prepared and presented. The most appropriate form for the presentation of such material is visual, since the mind, as most scientists tend to agree, operates predominantly with images.

The techniques and software tools usable in exploratory data analysis are currently very numerous, and new tools continue to appear. It would be completely unfeasible to survey all of them. Therefore, we have tried instead to set out the major tool categories and describe the key functions and properties of each category. The resulting classification looks as follows:

- *Visualisation.* The primary function of this tool category is representation of data in a visual form, i.e. creating various pictures from data: graphs, plots, diagrams, maps, etc. For this purpose, elements of data are translated into graphical features, such as positions within a display, colours, sizes, or shapes. It is important, however, that these graphical features coalesce into a single image rather than being perceived separately.

We divide the visual expressive means into display dimensions and visual, or retinal, variables. Display dimensions provide a set of positions within a display at which graphical elements, or marks, can be placed. Retinal variables represent various properties of the marks: shape, size, colour, texture, orientation, etc. In addition to the visual dimensions of a display, such as width, height, or depth, we consider also the display time, which can be used, for example, in animated presentations.

- *Display manipulation.* This class consists of interactive tools that support dynamic modification of the appearance of visual displays. The

general purpose of such modification is to enhance the image produced: to make it clearer and easier to perceive, to accentuate the distinctive features of the data represented, to focus on a particular item or subset of interest, etc. The manipulation is done through modifying the formula or algorithm used for the translation of data elements into visual features (we call this formula or algorithm the “visual encoding function”).

- *Data manipulation*, i.e. derivation of new references and characteristics from existing ones. There are two major purposes in doing this: to simplify the data and make it easier to analyse, and, conversely, to enrich the data and consider various aspects of it. Thus, data aggregation reduces the amount of data and hence simplifies the analysis. Data interpolation, in contrast, produces additional data.
- *Querying*, i.e. the automated search for answers to user-specified questions. Most typically, this is to search for references with specified characteristics or to search for the characteristics of specified references. Dynamic query tools, which allow the user to easily modify query conditions and quickly provide the required answer, are especially important for EDA.
- *Computation*. In this category, we briefly consider the computational methods of statistics and data mining. Unlike the computations involved in data manipulation, which prepare data for further analysis, for example by transforming the data into a more suitable form, the function of computational tools is a kind of data distillation, or extraction of the essential features of data. Some examples of the outputs produced by computational tools are statistical characteristics of a dataset as a whole, indicators of relatedness between attributes, and models that predict some characteristics on the basis of other characteristics, in particular, future developments on the basis of the current state and of the history.

In exploratory data analysis, it is usually not enough to use a single tool. Various tools need to be combined. We consider two basic modes of tool combination, sequential and concurrent, and discuss the various mechanisms used for tool combination. Visualisation is an essential component of any tool ensemble. Initial data visualisation is used in order to understand what tools should be used for further work. Results produced by any non-visual tool need to be visualised so that the analyst can see and interpret them.

Throughout this chapter, we provide many examples of various tools. Even when discussing non-visual tools such as data manipulation or computational methods, we use visualisation intensively to illustrate the examples. Readers can easily note that we have taken every opportunity to stress the great role of visualisation in exploratory data analysis. At the begin-

ning of the chapter, we make an attempt to substantiate the importance of visualisation.

## 4.1 A Few Introductory Notes

In this chapter, we shall try to talk about tools for exploratory data analysis on the same level of generality as we adopted in the discussion of tasks. This means that we shall not describe and analyse particular software products and prototypes, particular techniques for data visualisation, particular methods of statistical analysis, or particular algorithms for data mining. We are going to deal with broad categories, such as data visualisation in general or computational data analysis methods in general. Of course, we shall refer to particular techniques and methods as examples, as we did before for tasks.

Let us define what we consider as a tool for EDA. From the definitions of the word “tool” given by our near and dear friend, the Random House Webster’s Dictionary (Random House 1996), the most appropriate is “anything used as a means of accomplishing a task or purpose”. In principle, we have no reason to restrict our discussion to only computer-based tools. Although it is now hard to imagine any data analysis being done without using a computer, there may be situations in which paper and pencil serve better. Computers often restrict the imagination by offering predefined techniques and inducing standard approaches, while a white sheet of paper puts no limits on creative thinking. Besides, computers are not yet sufficiently convenient and suitable for making arbitrary drawings and notes anytime, anywhere, in an attempt to capture half-formed ideas or to externalise non-verbal conceptions.

From this note, it should be clear that we consider the foremost tool for exploratory data analysis to be the mind of the analyst. All other tools are supplementary. Their role is to facilitate and advance the work of the mind but not to substitute for it. We do not believe in machines that, after being supplied with data, could automatically produce the required knowledge or the solution of a problem. Any output of any tool is just material for the explorer’s mind, a subject for thinking.

Unfortunately, we do not feel ourselves capable of discussing the work of the principal tool for data analysis, i.e. the human mind. Therefore, we focus on supplementary tools. Of these supplementary tools, we deem data visualisation to be the most important.

## 4.2 The Value of Visualisation

First of all, we must explain why we believe visualisation to be so important. But before that, we need to define what we mean by visualisation. If we refer to a dictionary, we can find the following meanings of the word “visualise”:

- (intransitive) to recall or form mental images or pictures;
- (transitive) to form a mental image of;
- (transitive) to make perceptible to the mind or imagination

(Random House 1996). Of these meanings, the most appropriate is the third one. We see this short phrase as a mission statement, as the definition of the primary purpose of visualisation tools: they are required *to make data and the corresponding phenomena perceptible to the mind or imagination* of the explorer. That is all. No more and no less.

Despite its simplicity, this mission statement provides a full justification for the primary importance of visualisation as a supplementary (to the human mind) tool for exploratory data analysis: in order to be able to think about data, the mind needs to perceive the data. No thinking is possible without prior perception. And it is clear that the perception must be, on the one hand, correct with respect to the data, and on the other hand, opportune for reasoning. This imposes very high requirements upon data visualisation tools.

In the research areas related to exploratory data analysis, that is, information visualisation, geographic visualisation, and human–computer interaction, the word “visualisation” is mostly associated with graphical representation of data, i.e. encoding elements of data by graphical primitives such as positions on a plane, sizes, colours, textures, or shapes of graphical symbols. Does this contradict the general concept of visualisation or significantly reduce its scope? We believe that neither the former nor the latter is true. The human mind has rather limited capabilities for perceiving data represented in a non-graphical form, for example as a table of numbers. Imagine a table of the daily values of the prices of two stocks collected over a month, on the one hand, and the same data represented as two lines on a time graph. Which representation is more “perceptible to the mind or imagination”? We do not expect that anyone will need much time or significant mental effort to find an answer to this question. It is quite clear which representation can tell us better about the trend in the price of each stock and allow us to compare those trends. It is a common truth that “a picture is worth a thousand words” or, in our case, that one picture is worth much more than a collection of numbers.

---

Hence, it is quite natural that researchers focus on graphics as a primary means of making data perceptible to the mind or imagination. Some researchers are actively exploring other means such as acoustic or touchable representations, but the mainstream relies on human vision as the major channel of perception.

We would like to relate this primacy of graphics and images to ideas from cognitive psychology. As might be expected, we shall refer again to the book by Rudolf Arnheim *Visual Thinking* (Arnheim 1997).

Arnheim argues that not only perception provides material for thinking but perception and thinking are inseparable: perception involves thinking and thinking involves perception. On the one hand, perception is not mere recording of stimulus material, but organisation of this material into concepts:

Perception consists in fitting the stimulus material with templates of relatively simple shape, which I call visual concepts or visual categories. The simplicity of these visual concepts is relative, in that a complex stimulus pattern viewed by refined vision may produce a rather intricate shape, which is the simplest attainable under the circumstances. What matters is that an object at which someone is looking can be said to be truly perceived only to the extent to which it is fitted to some organized shape. (Arnheim 1997, p. 27)

It is most natural to attribute this process of forming visual concepts to the activity of the brain, i.e. to see perception as thinking:

In order to account for the complexity and flexibility of shape perception, it seems preferable to assume that the decisive operations are accomplished by field processes in the brain, which organises the stimulus material on its arrival according to the simplest pattern compatible with it. (Arnheim 1997, p. 28)

On the other hand, it is these visual concepts that serve as material and tools for thinking. When these concepts are not obtained directly from the environment, they are retrieved from memory:

Thinking ... can deal with objects and events only if they are available to the mind in some fashion. In direct perception, they can be seen, sometimes even handled. Otherwise they are represented indirectly by what is remembered and known about them. (Arnheim 1997, p. 98)

These internal representations are generally called mental images. Arnheim thinks that this usage of the word “image” is not accidental, because mental representations of things have a visual rather than any other (for example, verbal) nature. In substantiation, Arnheim states that any representation on any level of abstraction has to meet one condition: it must be structurally similar (isomorphic) to the pertinent features of the situation for which the thinking is valid. If everything what we see is indeed trans-

formed in the mind into a completely different representation, this representation should be equivalent to a visual representation; otherwise, the requirement of isomorphism may be violated. Thus, verbal language is not a suitable means for representing sensory experiences, although it can co-operate successfully with imagery. Two principal features of verbal language make it unsuitable as a primary vehicle for thought. First, this medium is purely linear and therefore very weak in representing any spatial information. Second, the elements of this medium, i.e. words and phrases, are arbitrary. They represent things and concepts only by virtue of a convention, i.e. an agreement among people how the signs that the language consists of must be interpreted. Therefore, in many situations, there may be no appropriate signs to express what one sees or thinks, just because such signs have not been yet created and adopted by society. For a visual image, there is no such limitation.

Hence, there is no reason to regard mental images as different in principle from the results of direct perception. It may be objected that mental images, of whatever nature, are necessarily more general than perceptual images. However, Arnheim convincingly demonstrates that generality and abstraction are essential qualities of perception. We shall not reproduce here all of the argument provided by Arnheim in support of the thesis of the unity of perceptual and mental images, or of percepts and concepts. Here is just one quotation:

Strictly speaking, no percept ever refers to a unique, individual shape but rather to the kind of pattern of which the percept consists. ... Even the image of one particular person is a view of a particular pattern of qualities, of that kind of person. There is, therefore, no difference in principle between percept and concept, quite in keeping with the biological function of perception. In order to be useful, perception must instruct about kinds of things; otherwise organisms could not profit from experience. (Arnheim 1997, p. 28).

In a summary of his argument, Arnheim says:

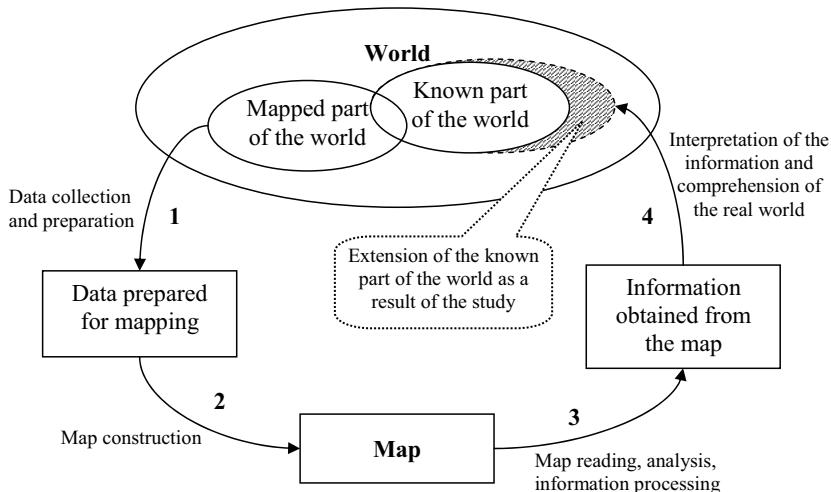
I have tried to show that perception consists in the grasping of relevant generic features of the object. Inversely, thinking, in order to have something to think about, must be based on images of the world in which we live. The thought elements in perception and the perceptual elements in thought are complementary. They make human cognition a unitary process, which leads without break from the elementary acquisition of sensory information to the most generic theoretical ideas. The essential trait of this unitary cognitive process is that at every level it involves abstraction. (Arnheim 1997, p. 153)

On this basis, it should be quite clear how important data visualisation is, since it provides material for perception and thereby enables “the grasping of relevant generic features of the object”. Thinking operates with mental

images, which are of the same nature as or at least equivalent to images formed by means of vision. Therefore, it is quite natural and appropriate that data visualisation is meant first of all for human eyes, by providing graphical, cartographical, or pictorial representations of data.

We also feel it apposite to refer to another author, Konstantin Salichtchev, the leader of the former Soviet school of cartography. His textbook on cartography (Salichtchev 1982) greatly influenced us at the beginning of our research career and, in fact, determined our orientation towards data visualisation and exploratory data analysis. Salichtchev regarded maps first of all as instruments for exploration of the real world. According to Salichtchev, a map is an abstracted, generalised, and simplified image of the world, that is, a *model* of the world that reflects the aspects, properties, and processes of reality that are relevant to the purposes of a particular investigation. Salichtchev defined the concept of the *cartographic research method*, which consists in applying maps for the scientific description, analysis, and comprehension of phenomena. The essence of the method is that one explores maps as models of reality instead of exploring reality itself (this opportunity is especially valuable when direct exploration of reality is difficult or impossible, as, for example, in studying long-term global processes). Hence, maps play a dual role: as the instrument and as the subject of the investigation.

Salichtchev suggested a schematic representation of the cartographic research method. We reproduce Salichtchev's diagram (translated from Russian) in Fig. 4.1.



**Fig. 4.1.** A schematic representation of the cartographic research method (after Salichtchev (1982))

Salichtchev stresses that steps 2, 3, and 4 of the method involve not only excluding unnecessary information, but also acquiring new knowledge by information processing and by inductive and deductive reasoning. Map construction results in a qualitatively new spatial image of the real world; an analysis of this image using the cartographic research method creates essentially new information about the phenomena represented on the map. An interpretation of this information on the basis of the available knowledge and experience leads to further extension of what is known concerning these phenomena. This means that map construction and map use generate new information in addition to the original information involved in map creation.

The cartographic research method comprises several groups of techniques, such as map-based measurements and computations, statistical analysis techniques, mathematical modelling, and information theory approaches. However, the most important is *visual analysis*, which is based on the capacity of maps to represent spatial shapes, relations, and structures in a directly perceivable form.

We strongly believe that everything said by Salichtchev concerning maps applies equally to all other forms of visual representation of things and phenomena, when such representations are created for the purposes of exploration and analysis rather than illustration or decoration. Instead of spatial information, graphs or diagrams can reflect other kinds of essential features and relationships. Their role is to make these features and relations easily perceptible, as maps do with respect to spatial information. Hence, the cartographic research method is a special case of a more general concept, which could be called the “visualisation-based research method”. However, there is no need for a new term: this content is included in the conventional notion of exploratory data analysis. The scheme in Fig. 4.1 could be made to present the essence of this notion perfectly just by replacing the words “map” and “mapping” by “visualisation”.

It should not be concluded from this ode to visualisation that any diligent translation of data into graphical elements automatically enables “the grasping of relevant generic features of the object”. Unfortunately, representations that are really fertile and conducive to thinking occur much less often than useless, unproductive ones. This is not surprising, since there is no recipe for making productive visualisations. Designers and users of visualisation tools mostly rely on “best practice”, a few general principles, and some grains of experience gained through trial and error.

One of the goals that we pursue in this book is to share our knowledge of the general principles and the grains of our experience, as well as refer to the “best practice” known to us. This refers not only to visualisation but also to other kinds of tools, and we plan to pursue our goal in connection

with our task typology. But before doing this, let us make an inventory of the tools and overview their functions and properties.

## 4.3 Visualisation in a Nutshell

### 4.3.1 Bertin's Theory and Its Extensions

Bertin, whose notions of reading levels and question types provided us with the starting point for building our task typology, is generally known as the founder of the theory of graphical representation of data. This seminal theory, which is expounded in the extensive treatise *Semiology of Graphics. Diagrams, Networks, Maps* (Bertin 1967/1983), can be briefly summarised as follows.

Graphical representation is the encoding of components of data by means of *visual variables*. There are two functionally different classes of visual variables, *planar* and *retinal*. The planar variables are the two spatial dimensions of the plane. The retinal variables include *size*, *value* (brightness), *colour* (hue), *texture*, *orientation*, and *shape*. The plane is the mainstay of all graphical representations. Data items are represented by means of marks positioned on a plane. There are three types of marks, called, in accordance with their form, *point*, *line*, and *area* marks. Bertin calls these types “implantations”. The retinal variables may be applied to marks, and define their visual properties.

The visual variables have different *levels of organization* in accordance with their perceptual properties: associative, selective, ordered, and quantitative. A variable is *associative* when it permits the immediate grouping of all marks differentiated by this variable. For example, shape is associative because squares, triangles, and circles of the same size and colour are seen as similar signs. A variable is *selective* when it enables us to immediately isolate all marks belonging to the same category of this variable. Thus, a “family” of red marks is well differentiated visually from a “family” of blue or green marks; hence, colour is a selective variable. A variable is *ordered* when the visual classing (ranking) of its categories is immediate and universal. For example, grey is perceived as intermediate between white and black, and a medium size as intermediate between a small and a large size. A variable is *quantitative* when the visual distance between two categories can be immediately expressed by a numerical ratio. For example, one length can be perceived as three times another length, and one area can be perceived as one-quarter of another area.

The perceptual properties of the visual variables are summarised in Table 4.1.

**Table 4.1.** Perceptual properties of visual variables (according to Bertin)

	Associative	Selective	Ordered	Quantitative
Planar dimensions	yes	yes	yes	yes
Size		yes	yes	yes
Brightness (value)		yes	yes	
Texture	yes	yes	yes	
Colour	yes	yes		
Orientation	yes	yes		
Shape	yes			

The general principle of data presentation is that “the visual variables must have a level of organisation at least equal to that of the components which they represent”. Bertin refers here to the type of data scale that a visual variable can portray, i.e. nominal, ordinal, or numeric (ratio). Associative and selective variables correspond to the nominal scale, ordered variables to the ordinal scale, and quantitative variables to the numeric scale. If, for example, the goal is to represent values of a numeric attribute graphically so that a viewer can extract ratios from the visualisation, for example to see immediately that one value is twice another, one must choose a visual variable with quantitative organisation, that is, either one of the planar variables or a size. The general principle formulated by Bertin parallels the already mentioned statement by Arnheim that any representation “must be structurally similar (isomorphic) to the pertinent features of the situation for which the thinking shall be valid” (Arnheim 1997, p. 227).

Another requirement concerns the “length” of the visual variables, i.e. the number of categories or steps that can be distinguished visually: for example, distinguishably different colours or brightness levels. A visual variable must have a length equal to or greater than the component that it represents. If the length of the variable is insufficient, the observer will perceive some of the different data categories as being identical.

While compliance with these requirements concerning the level of organisation and the length of the variables is a necessary prerequisite for building good graphics, it is still not sufficient. According to Bertin’s *im-*

*age theory*, a visualisation is good if it permits immediate extraction of the necessary information, i.e. finding the answer to the observer's question at a single glance, with no need to move one's eyes or to shift one's attention and involve memory. Bertin uses the term "image" to refer to "the meaningful visual form, perceptible in the minimum instant of vision". An optimal visualisation contains a *single image* providing the answer to the observer's question. Visualisations with more images are inefficient because they require integration across images. For example, chart maps require inspection of each chart and comparison between charts, which takes much time.

The most efficient constructions are those in which any question, whatever its type and level, can be answered in a single instant of perception, that is, *in a single image*. Such constructions are restricted to using a maximum of two planar variables and one retinal variable. When the information necessitates more than three variables, one cannot construct a figure that could provide an immediate response to all types of questions. The image will not accommodate the representation of a meaningful fourth variable. Consequently, it is necessary to construct multiple images in order to provide answers to all questions. Any designer who uses only a single construction is limited to answering only one preferred type of question.

Bertin says that graphical representation can perform three functions:

- recording information (inventory drawings);
- communicating information;
- processing information.

Of these three functions, the last is the closest to exploratory data analysis. In order to be fit for this function, a visualisation, on the one hand, must be comprehensive, and on the other hand, must be reduced to the smallest number of memorizable images. Comprehensiveness means avoiding any prior reduction of the information (e.g. by classification), using the "complete information, which alone provides all the givens for pertinent correlations and choices... But also it matters that all types of comparisons and classings are possible and easy. The most useful questions will obviously involve the overall level of reading, where their answer will be found in a limited number of comparable images" (Bertin 1967/1983, p. 164). This reference to the overall reading level fully corresponds to our ideas concerning the primacy of synoptic tasks in exploratory data analysis.

In the next paragraph, Bertin points to the role of manipulation in data exploration:

With this function, the graphic is an experimental instrument leading to the construction of collections of comparable images with which the researcher “plays”. We class and order these images in different ways, grouping similar ones, constructing ordered images to discover the synthetic schema which is at once the simplest and most meaningful.

The latter phrase, actually, points to the ultimate goal of EDA: “to discover the synthetic schema [i.e. the underlying organisational pattern or structure; this corresponds to connection discovery tasks in our terminology] which is at once the simplest and most meaningful”. This also corresponds to the views in gestalt psychology concerning the innate pursuit of simplification involved in all cognitive activities, as well as the scientific principle of parsimony, or Occam’s razor.

Although the work of Bertin greatly influenced further theoretical and practical endeavours in the area of data visualisation, many researchers criticised Bertin’s theory for being based solely on his introspection, with no attempt being made to provide any empirical support or any links to perceptual or psychological research.

An attempt to ground Bertin’s theory in the first principles of perception has been undertaken recently by Mark Green (1998). Green believes that Bertin’s concept of an “image” is related to psychological experiments on detecting a “target” element in a picture by an observer, where the target element appears among other elements but differs from them by one or more attribute, such as colour or orientation. In some situations, the observer seems able to detect the target effortlessly, as if he/she were processing the entire visual field in a single automatic, parallel operation. This is often termed “pre-attentive search” because there is no need to focus attention on specific objects in the image: the target simply seems to “pop out”. In other cases, the observer seems to search for the target by purposefully moving his/her attention through the visual field and serially scanning each object in it. The dichotomy of pre-attentive and attentive perception is analogous to Bertin’s distinction between immediate and sign-by-sign perception.

In trying to understand the conditions that produce pre-attentive, effortless image processing, researchers in psychology have detected that search is pre-attentive and parallel if (1) the target and distracters differ in a single feature such as colour or orientation, and (2) the difference in the value corresponding to this feature is great enough. If the target is defined by some combination of features (e.g. red and horizontal), then the search becomes slow and requires effort. There are several theories that aim to explain this phenomenon. The most popular theory is that different parts of the brain (“feature modules”) are responsible for representing different features. Thus, the brain has a colour module, an orientation module, etc.

When it is necessary to look for a combination of two or more features, the brain has to involve two or more feature modules, somehow organise their work, and integrate the results. This requires much more time and effort than in a case where a single module can perform all the work.

Regardless of the exact theory, the general belief is that pre-attentive feature search can only occur when one is examining the contents of a single-feature representation. Serial-conjunction search, on the other hand, requires the observer to integrate features of a single object by reference to their common spatial coordinates. The integration is actually performed by focussing attention on the particular spatial location.

The empirical findings and the theories suggested to explain them run parallel to many aspects of Bertin's image theory, in particular, Bertin's statement that a graphic can be perceived as a single image, i.e. immediately, if it contains a maximum of two planar variables and one retinal variable. Although the search paradigm is not exactly a visualisation task, there are close similarities.

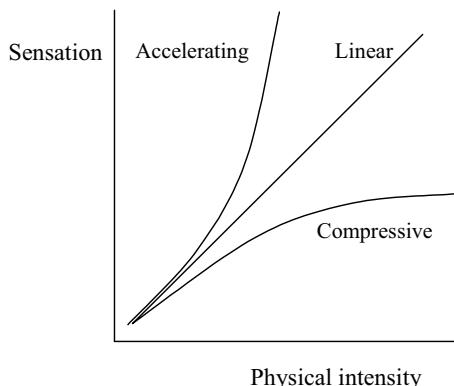
Hence, Green concludes, vision research provides first-principles explanations of many aspects of image theory. The three-component limit is due to the way image features are represented in the nervous system and the difficulty of conjunction search. Planar and retinal variables are different because spatial location ties all other attributes together.

Another part of Green's discussion concerns the levels of perceptual organisation of the visual variables. He believes that the distinctions between the variables can be explained on the basis of psychophysical scaling studies, which investigate the relationships between the real intensity of various physical stimuli (such as light intensity) and its perceived magnitude (such as apparent brightness). Three basic types of dependencies are possible (see Fig. 4.2):

- *linear*, where the sensation grows in direct proportion to the physical intensity;
- *compressive*, where the sensation grows slower than the physical intensity;
- *accelerating*, where the sensation grows faster than the physical intensity.

When the dependency is linear, a doubling of physical intensity produces a doubling of sensation, so ratios are maintained. This property is required for Bertin's quantitative perception. Unfortunately, linear dependencies occur very rarely. The most common dependency is compressive. Thus, a doubling of light intensity is perceived as a far smaller increase in apparent brightness. Moreover, a doubling of the intensity from 10 to 20 and a doubling from 40 to 80 are perceived as different relative changes.

Physical variables producing such non-linear scales would not permit a quantitative organisation level.



**Fig. 4.2.** Three types of dependency between the physical intensity of a visual stimulus and the apparent sensation (after Green (1998))

Psychophysical experiments have yielded significant evidence that the only variable that reliably demonstrates a linear dependency between the physical intensity and its sensation is spatial extent, for example judgements of line length. This finding is in good agreement with Bertin's assertion that the only variables with quantitative organisation are planar dimensions and size. All these variables involve judgements of spatial extent. Bertin's statement that brightness is ordered but not quantitative can be explained by the fact that brightness is a compressive function and therefore does not permit direct perception of physical ratios. However, the brightness function is monotonic, so ordered organisation is supported.

Furthermore, Green believes that Bertin's theory can not only be substantiated by psychological research but also be extended on the basis of this research. One of the extensions is that the variable "shape" can be selective (while Bertin claims that shape is only associative). Actually, this depends on the choice of shapes. Thus, Bertin considers solid shapes such as filled triangles, squares, and circles, which are highly similar, while, for example, it has been experimentally proved that "X" and "O" shapes are very well differentiated visually.

Another extension concerns colour (hue). It is true that hue is in general a nominal variable: red, green, blue, etc. do not form an ordered scale. However, over small ranges, hue can be ordered. For example, it is possible to construct an ordered scale of yellow starting from the unique yellow hue (without any trace of other colours) and extending to either (but not

both) the unique red or the unique green hue. Observers can then readily order hues along the yellow-red or yellow-green continuum.

Bertin considered two properties of colour, hue and brightness, but did not take into account saturation, i.e. the amount of white (or grey) mixed with a spectral hue, i.e. the purest hue. Saturation is an ordered variable: viewers can readily order lights by increasing amount of saturation. Concerning brightness (as well as other ordered variables with compressive functions), Green deems it possible to turn it into a quantitative visual variable by means of appropriate scaling. Thus, to represent a twofold increase of some quantity, the brightness must be multiplied by a factor of more than two. Most computer monitors have built-in rescaling of the brightness level, which compensates for the non-linearity of the visual brightness. Although the perception of various levels of brightness from a computer monitor varies with viewing conditions, brightness nevertheless can be used on a computer screen (but not on a printed page) to approximate quantitative data.

Besides the discussion of the “traditional” visual variables, Green suggests some additional variables that can be used on computer screens:

- *motion*, which can be split into two subvariables, *velocity* and *direction*;
- *flicker*, with two subvariables, *frequency* and *phase*;
- *binocular disparity*, i.e. giving the left and right eyes slightly different views of the same visualisation, which supports easy perception of relative depth.

In our opinion, it is unclear whether these variables have any advantages over the traditional variables, and the application of the new variables in data visualisation will certainly require people to get accustomed to them before they can be used effectively in data analysis.

Table 4.2 presents the list of visual variables and their properties updated by Green.

Green also points to some deficiencies of Bertin’s image theory. Thus, Bertin does not acknowledge the possibility of interactions between retinal variables, for example the influence of colour variation upon the perception of texture. Some combinations of variables are perceptually inseparable, for example hue and saturation: observers cannot ignore hue when trying to attend to saturation, and vice versa. There are also some perceptual effects that contradict Bertin’s assertions concerning the properties of the visual variables. For example, it is easier for an observer to select tilted lines from among vertical ones than the other way around. Therefore, whether a variable is associative or selective depends in part on the exact values chosen for a given visual variable. For a more detailed discussion of

**Table 4.2.** Perceptual properties of visual variables, updated (after Green (1998))

	<b>Associative</b>	<b>Selective</b>	<b>Ordered</b>	<b>Quantitative</b>
Planar dimensions	yes	yes	yes	yes
Size		yes	yes	yes
Brightness (value)		yes	yes	yes, if scaled
Texture	yes	yes	yes	
Colour (hue)	yes	yes	yes (limited)	
Orientation	yes	yes		
Shape	yes	yes		
Motion: velocity		yes	yes	yes, if scaled
Motion: direction		yes		
Flicker: frequency		yes	yes	yes, if scaled
Flicker: phase		yes		
Disparity		yes	yes	

these and other relevant issues, we refer readers to the work of Alan MacEachren (MacEachren 1995, pp. 82–92).

While, in general, experiments support Bertin's assertion that an image (i.e. a visualisation perceived immediately in its entirety) can contain no more than three components, i.e. two planar variables and one retinal variable, Green mentions some empirical findings that suggest that, under certain circumstances, images with four or more components are possible. In particular, depth can be an additional image component. The use of depth cues, such as perspective, shading, transparency, motion parallax, or binocular disparity, might admit the third dimension as a third spatial variable, allowing four-component visualisations. Again, MacEachren's book can serve as a source of additional information concerning depth cues and different approaches to representing the third spatial dimension on a plane (MacEachren 1995, pp. 136–147).

While Green analyses and extends Bertin's theory from a psychologist's point of view, MacEachren discusses Bertin's ideas from a cartographer's perspective. MacEachren cites various elaborations, amendments, and additions to Bertin's theory of visual variables suggested by a number of researchers, including MacEachren himself (MacEachren 1995, pp. 272–276). Here is a brief summary:

- Texture needs to be divided into several variables:
  - arrangement of texture elements;
  - density of the elements (the only aspect considered by Bertin);
  - size of the elements;
  - shape of the elements;
  - orientation of the elements.
- Colour saturation is explicitly included among the visual variables.
- MacEachren introduces a new visual variable, *clarity*, consisting, like colour, of three subvariables:
  - *crispness* (or *fuzziness*) of object edges;
  - *resolution* (spatial precision);
  - *transparency*.

MacEachren deems these subvariables very useful for, in particular, representing uncertainty in data or information, but warns that no more than two or three values of these variables should be used at once.

MacEachren presents a highly elaborated table of perceptual properties of visual variables, which is reproduced in Table 4.3 here. The last column in MacEachren's table refers to the notion of visual levels: a viewer of a graphic depiction can group sets of objects into common wholes that are seen as occupying different visual (or conceptual) planes. This notion is related to Bertin's selective and associative properties of visual variables. Hence, the two columns on the right correspond to Bertin's concepts of "selectivity" and "associativity", respectively.

Besides the visual variables, which can be used in static images either on a computer screen or on paper, MacEachren considers tactile, acoustic, and dynamic variables. We shall not reproduce here the discussion of all three groups of variables, but shall focus on the dynamic variables, which include, according to MacEachren:

- *Display date*: The time at which some display change is initiated. Display date can be used to represent chronological date, but also for other purposes; for example, for highlighting particular places.
- *Duration*: The length of time between two identifiable states. Duration can be applied to individual frames in an animation or to sequences of frames. In a repeating cycle, duration can be applied to the period of the cycle, i.e. the interval between repetitions. The simplest application is the binary cycling of on/off states used in "blinking".
- *Order*: The sequence of frames or scenes. MacEachren refers to some successful examples of using presentation order as a variable matched to the numerical order of some quantity other than chronological time.

**Table 4.3.** The extended typology of visual variables and their perceptual properties (after MacEachren (1995), p. 279)<sup>a</sup>

	Numerical	Ordinal	Nominal	Visual isolation	Visual levels
Location	++	++	++	++	
Size	++	++	++	++	++
Crispness		++ <sup>b</sup>		++	++ <sup>b</sup>
Resolution		++ <sup>b</sup>		++	++ <sup>b</sup>
Transparency		++ <sup>b</sup>	+	++	++ <sup>b</sup>
Colour value	+	++		++	++
Colour saturation	+	++		+	++ <sup>c</sup>
Colour hue	+ <sup>f</sup>	+ <sup>d</sup>	++	++	+ <sup>d</sup>
Texture	+	+	++ <sup>e</sup>	++	++
Orientation	+ <sup>f</sup>	+ <sup>f</sup>	++	++	
Arrangement			+	+	g
Shape			++		

<sup>a</sup> ++, good; +, marginally effective; no mark, poor.

<sup>b</sup> The clarity variables of crispness, transparency, and resolution can be used for no more than two or three categories. These variables are untested, but are assumed to be most useful for representation of uncertainty. They may prove to be most practical in an interactive setting in which an analyst is able to toggle them on and off when needed.

<sup>c</sup> Purer, more saturated colours appear to be in the foreground, while dull, unsaturated colours fade into the background.

<sup>d</sup> Hues must be carefully selected for an order of hierarchy to be apparent (e.g. the part-spectral sequence from yellow through orange to red). Hues interact with one another in somewhat unpredictable ways, so it is often difficult to determine which hues will dominate others.

<sup>e</sup> Pattern texture is good for only two, or perhaps three, identifiable categories.

<sup>f</sup> Orientation provides limited ability to communicate numerical or ordered information – glyphs based on a clock face, and geologic strike and dip symbols are successful examples.

<sup>g</sup> Pattern arrangement is best as a redundant variable to make a visual difference between categories more obvious.

- *Rate of change*: The difference in the magnitude of the change per unit time for each of a sequence of frames or scenes.
- *Frequency*: The number of identifiable states per unit time.

- *Synchronisation* (phase correspondence): The temporal correspondence of two or more time series.

As to the properties of the dynamic variables, MacEachren notes that “display date” is clearly a nominal variable. It can be used to show that a feature is or is not in a location at a particular point in time. Dates in relation to one another are, of course, ordinal. “Duration” is measured on a ratio scale; patterns produced by changing the duration can suggest nominal distinctions, such as “fuzzy”, “jittery”, or “pulsating” temporal patterns. “Order” is clearly ordered but offers no way to signify numerical differences. “Rate of change” shows a numerically measurable difference from scene to scene and is therefore suited to ordered and numerical depiction. Theoretically, “synchronisation” is capable of depicting ratio-level differences: it is possible to measure the degree to which the phases of two time series match. In practice, it seems that synchronisation (or the lack of it) produces two nominal categories of “in phase” and “out of phase”.

Visualisation-related research includes not only comprehensive studies aimed at building general theories or frameworks for visualisation, but also investigations into particular aspects or issues in visualisation. It is not our goal to do a systematic survey of all literature related to the theory and methodology of visualisation. However, we would like to mention several works that we learned from in the initial stages of our career in visualisation or were influenced by in the later stages:

- Salichtchev (1982): Cartographic visualisation for modelling of real-world phenomena; techniques of cartographic visualisation, their properties, applicability, and the opportunities for analysis that they provide.
- Tufte (1983, 1990), and Kosslyn (1994): Practical guidelines for designing good graphics; and numerous examples of good and bad graphics and analysis of their positive or negative features.
- Cleveland and McGill (1984, 1986): Experiments on graphical perception, which address mainly the accuracy with which values are read from graphics; and suggestions for redesigning several popular types of graphics to make the perception of information from them more accurate.
- Mackinlay (1986), Roth and Mattis (1990), Casner (1991), and Senay and Ignatius (1994): The use of Bertin’s theory for automated or computer-aided design of data visualisation; formalisation of the principles of visualisation; rules for combining visual primitives; and elaboration of the characteristics of data to be taken into account in graphics design.
- Lyutyy (1986): An investigation of a cartographic sign system – “map language”; definition of two sublanguages representing spatial and the-

matic contents, respectively (i.e. positions, shapes, neighbourhood, etc., and identities, qualities, and various non-spatial properties, respectively); and the structure of a graphical sign system, in particular, division into a set of positions and a set of graphical morphs, signs, and sign combinations that can be placed in these positions.

- Brewer (1994): A deep investigation into the properties of colour; analysis of various colour scales; practical guidelines for using colour in visualising data; and the principles of building colour scales.
- Wilkinson (1999): Principles of graphical representation expressed according to the paradigm of object-oriented design; formalisation of the process of moving from raw data to a graphic, which includes data selection, transformation, choosing a coordinate system, the type of graph (e.g. point, line, bar, schema, or contour), so-called “aesthetic attributes” (which correspond to Bertin’s “retinal variables”), and the introduction of “guides” (e.g. axes and legends); and inventories of graph types, aesthetic attributes, scales and their possible transformations, coordinate systems, and other components of the “grammar of graphics”.

It should be stressed that these and other researchers in visualisation and related areas have not refuted Bertin’s theory. They have elaborated, refined, extended, and deepened it, but the main principles of the theory remain valid.

We have always used Bertin’s theory and its later extensions in our practice. Nevertheless, we found it convenient for our purposes to rearrange the elements of the theory and re-express some of its principles in a particular way. In particular, we have applied Lyutyy’s idea of dividing graphical primitives into positions and signs and have incorporated parts of Wilkinson’s “grammar of graphics” concerning coordinate systems, plane transformations, and “facets”, i.e. arrangements consisting of multiple uniform graphics. We stress that what is described in the following sections should not be understood as an alternative theory of data visualisation or as an invalidation of any of the principles formulated by Bertin.

### 4.3.2 Dimensions and Variables of Visualisation

We would like to divide the visualisation primitives capable of representing components of a dataset into two groups, which will be referred to as *dimensions* and *retinal variables*, or simply *variables*. Dimensions provide sets of *positions*, which can be used for placing marks (point, line, or area symbols) or fragments of graphics. Retinal variables define the appearance of marks, i.e. their shape, size, colour, etc. In other words, dimensions dif-

fer from retinal variables in that they are “external” with respect to marks and serve as “containers” for marks, while retinal variables define the “internal”, individual properties of marks.

Dimensions include:

- The two spatial dimensions of a plane, which may be a sheet of paper, a computer screen, or any other two-dimensional medium. These dimensions may be used separately, i.e. each dimension is used for a different purpose, or in combination, as a unified two-dimensional space.
- The third spatial dimension, which may be available in some media such as immersive environments. The third spatial dimension can also be simulated on a plane by providing perspective views of a three-dimensional object or scene. Like the other two spatial dimensions, the third dimension can be used on its own or in combination with the other two dimensions to form a unified three-dimensional space.
- The temporal dimension, which is available in computer-based representations capable of changing over time (unlike paper-based representations). This dimension is often referred to as display time.
- Various arrangements of the display space (explained in detail below).

In fact, our list of dimensions extends Bertin’s concept of planar variables, which have quite different properties from the remaining (retinal) visual variables. Bertin did not consider the third spatial dimension and the display time as visual variables, since he focused on static planar graphics. Green (1998) mentioned that the third spatial dimension (referred to as depth) has properties close to those of the two planar dimensions. In particular, using depth cues in addition to the planar dimensions combined with a single retinal variable does not prevent one seeing a visualisation as an “image” (i.e. the pre-attentive perception of the whole picture), despite the three-component limitation for an image formulated by Bertin and mostly confirmed in experiments.

We include the display time among the dimensions because, analogously to the spatial dimensions, it provides a set of positions that can be used for placing marks or fragments of graphics. This approach to dealing with time is different from the approaches of other researchers. Thus, Green (1998) and MacEachren (1995) do not consider the display time directly but incorporate it into special variables, such as motion and flicker (Green) or display date and duration (MacEachren).

Like the other dimensions, arrangements provide places for marks or fragments of graphics. Unlike other dimensions, they are not completely independent but are built on top of other dimensions; specifically, they reorganise the display space (one-, two-, or three-dimensional) and change

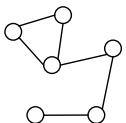
its properties. This may be the reason why Bertin and other researchers did not explicitly include arrangements among the visualisation primitives. However, this does not mean that these researchers did not consider arrangements at all. Thus, Bertin introduced the notion of an “imposition” to denote a particular way of utilising the two planar dimensions. He considered the following impositions:

- arrangement (which means that marks may be dispersed over the entire plane);
- rectilinear construction;
- circular construction;
- orthogonal coordinates;
- polar coordinates.

Furthermore, Bertin divides graphical representations into four groups according to the correspondences that they can represent:

- diagrams, which reflect correspondences between values of one component and values of another component (but not between different values of the same component);
- networks, which represent correspondences between different values of the same component;
- maps, which reflect correspondences between values of the same component arranged according to a geographic order;

**Table 4.4.** The correspondence between the groups of graphical representations and the types of impositions (after Bertin (1967/1983))

	Arrangement	Rectilinear	Circular	Orthogonal	Polar
<b>Diagrams</b>					
<b>Networks</b>					
<b>Maps</b>					
<b>Symbols</b>					

- symbols, which do not contain any internal correspondences but refer to something exterior and are recognisable owing to acquired habits or conventions.

Each type of graphical representation can be used with certain types of impositions, as is specified in Table 4.4.

A different approach to dealing with utilisation of the plane is taken by Wilkinson (1999). He groups together, on the one hand, various coordinate systems (Cartesian, polar, triangular, parallel, etc.), plane transformations (e.g. rotation, stretching, or warping), and projections from multidimensional spaces onto the plane, and on the other hand, methods of arranging multiple similar graphs (called “facets”), which include tables, trees, polar arrays, and mosaics.

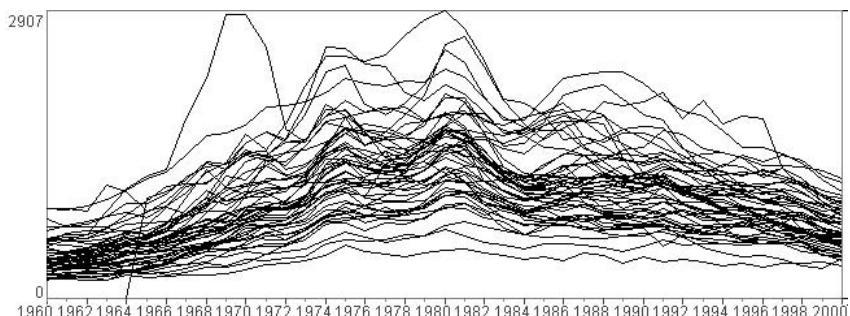
Our approach is based on the idea of differentiating positions from signs (according to Lyutyy (1986)). We distinguish dimensions, which provide positions, from variables, which define the appearance of signs. Coordinate systems, plane transformations, and the methods of arranging multiple graphs are all about positions, and hence should be included among the dimensions. We use the general term “arrangements” to denote different ways of organising or transforming the plane or, possibly, the other display dimensions, i.e. three-dimensional space and time.

We are not going to give a comprehensive overview of all known arrangements. However, we find it useful for a general understanding to consider some examples of arrangements. Below, we discuss a few widely used arrangements, which we also often use in our practice.

One of the simplest known arrangements is the juxtaposition of several uniform displays. This technique is often referred to as “small multiples”, a term introduced by Tufte (1983). An example of “small multiples” is the series of maps shown in Fig. 3.16, where each map shows the distribution of burglary rates over the territory of the USA in a particular year. In “small multiples”, each of the multiple displays (“facets”, in Wilkinson’s terms) has its own spatial dimensions, which are used as containers for placing marks. Simultaneously, the spatial dimensions of the overall composite display are used to place the facets. The arrangement of the facets can represent some component(s) of the dataset. Thus, the arrangement of the maps in Fig. 3.16 represents the temporal component of the USA crime dataset. Actually, it would be sufficient (and even more appropriate) to use for this purpose only one spatial dimension of the overall display, since time has a one-dimensional organisation (as a linearly ordered set). The second dimension is involved merely for better utilisation of the space provided by the medium, i.e. a computer screen or a sheet of paper.

Another arrangement is used in Fig. 3.13, where the dynamics of burglary rates in each state are represented by a time graph positioned within or near the boundaries of the state. We call this arrangement “space embedding”. Here, each time graph has its own two-dimensional space. The horizontal dimension represents time and the vertical dimension is used for representing the values of the burglary rate at different time moments. The space of the time graph is embedded in the space of the overall display (i.e. the map), which represents the geographical component of the data.

Multiple uniform displays may be not only juxtaposed within the space of the overall display or embedded in it but also overlaid within this space. Thus, Fig. 4.3 demonstrates a display resulting from overlaying 51 time graphs (like the one shown in Fig. 3.12) within the same coordinate space. Each time graph represents the dynamics of the burglary rate in a particular state of the USA. Hence, the overlay arrangement represents one of the referrers of the dataset, specifically, the set of states. Here, the spatial nature of this referrer is ignored, and it is treated as a reference set without ordering or distances. In a similar way, one might represent, for example, the variation of the performance of different schoolchildren over a school year. The overlay arrangement would represent, in this case, the set of schoolchildren, which is a population-type referrer.



**Fig. 4.3.** Overlaid time graphs representing the dynamics of burglary rates in all states of the USA over the time period from 1960 to 2000

The overlay arrangement is used intensely on maps. Almost any map represents multiple geographically related phenomena: sea coasts and boundaries of countries, rivers and lakes, mountains and forests, cities and roads, etc. In a GIS, a map is treated as a stack of “layers”, where each layer represents a particular phenomenon. The content and appearance of a map are defined by choosing the layers to be overlaid within the two-dimensional display space and specifying their visual properties, such as colour, size of symbols, and degree of transparency. The order of the lay-

---

ers plays an important role: layers put on top may suppress the visibility of layers placed underneath.

The representation of a data component by a collection of nodes connected by links, that is, as a graph, tree, network, or flowchart, is, in our terms, also a kind of arrangement. Such representations are useful, for example, in exploring genealogies or analysing the performance of algorithms and programs. In both examples, there are specific relations between some elements or subsets of the respective data component, i.e. the set of family members or the set of operations. These relations are potentially relevant for understanding the data. The standard spatial dimensions of a plane cannot adequately represent these relations. This deficiency is compensated by transforming the “normal” two-dimensional space (a continuous set with distances) into an artificial space that has quite different properties: it is discrete, it has no distances, and the relative positions of its elements have, in general, no meaning. Only particular relations, which are explicitly represented by lines or arrows linking corresponding elements, are meaningful.

Hence, this variant of arrangement consists in transforming the display space into another space with different properties, which conform better to the properties of the set to be represented. In this connection, we should recall again the principle of isomorphism: any representation must be structurally similar (isomorphic) to the pertinent features of the situation for which the thinking is valid (Arnheim 1997).

The use of nodes and connectors instead of a continuous space is not the only variant of transformation of the display space. Thus, sectioning of the spatial dimensions of a display may reflect a discontinuity of the data components mapped onto those dimensions. An example is a table-like display (where the cells do not necessarily contain texts; they may contain geometric shapes or be filled with various colours). The division of the display space into rows and columns indicates the absence of continuity in the data components represented by the horizontal and vertical dimensions of the table. For example, the rows of such a table may correspond to a set of schoolchildren, the columns may correspond to a set of school subjects, and the content of the cells may represent (graphically or as text) the marks obtained at the end of a school semester. It may be noted that the display space transformed in such a way is still not isomorphic to the components that it represents: it has “artefacts” such as ordering and distances, which are absent in the data components. However, the conventional use of table displays does not take these artefacts into account. Moreover, ordering of table rows or columns is often used to represent arbitrary additional information. For example, the rows of the table containing the marks of the pu-

pils may be sorted according to the age of the pupils or according to how far from the school they live.

In some types of graphics, the display space is transformed by introducing a different (i.e. non-rectangular) coordinate system, for example polar coordinates. Instead of the horizontal and vertical spatial dimensions (which, theoretically, have no beginning and no end), a space with polar coordinates has a particular point and a particular direction such that all positions are defined in terms of the distance from this point and the divergence from this direction. The distance component has a beginning, a “true zero”, which is the origin of the coordinates. The divergence component, or angle, has an interesting property: its beginning coincides with its end. Therefore, polar coordinates may be useful for representing cyclic data such as cardiograms: the angle corresponds to the temporal position within a cycle, and the distance represents the value of the attribute. Data from several cycles can be overlaid or put at different distances from the coordinate origin. Such a representation may expose discrepancies between the cycles. The use of polar coordinates is also convenient for data related to spatial directions, such as the intensity or frequency of winds in a wind rose.

In a polar arrangement, the distance between two arbitrary positions on the plane plays no role; only the difference between the distances of these positions from the origin of the coordinates is important.

Another arrangement, similar to polar coordinates is a diverging arrangement of several one-dimensional spaces, i.e. axes extending in different directions from a common point, as in a star diagram. In this arrangement, the angle is not used, the directions of the axes are arbitrary, and their relative positions do not imply any meaning. In a diverging arrangement, it is supposed that each axis defines its individual one-dimensional space, and positions on different axes cannot be directly compared.

Of the other possibilities of transforming the display space, we would like to mention cartograms, in which the geographical space is transformed so that the sizes of countries or other units of division of the territory do not correspond any more to the actual sizes of those geographical objects but represent some other characteristics, such as the population of the respective unit. In cartograms, the purpose of the transformation is to replace irrelevant properties and relations of the data component which needs to be represented (e.g. a set of countries or districts) by relevant properties. Thus, for demographic studies, the geographical extents of countries or districts are not as important as how much population lives in those areas (Dorling 1992).

In general, it may be said that any arrangement somehow structures the originally continuous, homogeneous and undifferentiated display space.

On this basis, we can give rather consistent names to the types of arrangement that we have discussed:

- Space partitioning (juxtaposition or “small multiples”), for example a series of maps for different time moments, or multiple parallel or diverging axes (one-dimensional spaces).
- Space embedding, for example time/attribute spaces embedded in geographical space.
- Space sharing (overlay), for example multiple time graphs in a common coordinate space.
- Space transformation, for example transformation of a continuous plane into node-link structures, or continuous plane into the discontinuous space of a table display, or of geographical space into a specific problem-oriented space in a cartogram.

Of these arrangements, the first three types supply additional dimensions, while the last type changes the properties of the standard display space without adding any new dimensions.

In general, dimensions (i.e. display space, display time, and arrangements) provide a set of positions (or, in other words, a framework) for placing graphical elements, or marks. Bertin distinguished three types of marks: *points* (zero-dimensional marks), *lines* (one-dimensional marks), and *areas* (two-dimensional marks). Since we include three-dimensional representation in our considerations, we extend this list with *volumes*, i.e. three-dimensional marks. The visual properties of marks are defined in terms of values of retinal variables. We do not define our own list of retinal variables, but refer instead to the standard set of variables introduced by Bertin and extended by other researchers; see Tables 4.1 to 4.3.

We shall use the term “graphical features” to denote instantiations (values) of graphical variables, such as particular sizes (as instantiations of the variable “size”) and particular colours (as values of the variable “colour”).

### 4.3.3 Basic Principles of Visualisation

Now, we would like to propose our elaboration of the basic principles of visualisation formulated by Bertin. This is done in accordance with our division of data components into referrers and attributes and the division of the visual primitives into dimensions and variables, introduced in Sect. 4.3.2. We should note that it is not always possible to fulfil all of the principles; sometimes compromises are unavoidable. Therefore, the principles should not be treated as strict requirements but rather as guidelines.

1. Referrers of the dataset should be represented by display dimensions, unless a different method of representation allows one to reduce the violation of the other principles. Some examples will be provided later.
2. The properties of the dimensions used to represent referrers should be consistent with the properties of the value domains of the referrers. This concerns the general properties of ordering, continuity, and presence of distances as well as particular properties and relations pertinent to the nature of the referrers, for example kinship relations between members of a family.
3. The reference set of the data should, whenever possible, be represented so as to be seen as a whole. For each referrer, there should be a dimension representing it, and for each value of the referrer, there should be a corresponding position in this dimension. Preferably, it should be possible to see all positions simultaneously. This means that the temporal dimension is not highly recommended for use, since its positions can only be viewed in a sequence, one position at each time moment.
4. Attributes of the dataset may be represented by retinal variables or by dimensions that are not used for representing referrers. For example, the horizontal spatial dimension in a time graph is used to represent a temporal referrer, while the vertical dimension represents the values of an attribute.
5. The properties of the variables or dimensions used for representing attributes must be consistent with the properties of the value domains of those attributes, i.e. ordering, distances, the presence of a “true zero”, and the cardinality of the domain, i.e. the number of elements.
6. The representation should permit the unambiguous ascertaining of which reference corresponds to each mark present in a display. Marks corresponding to different elements of the reference set must be sufficiently differentiated, for example by position or colour. When some referrer is represented by a space-sharing arrangement (overlay), marks corresponding to the same value of the referrer must be visually linked for better differentiation from marks corresponding to other values of the referrer. For example, in the overlaid time graphs shown in Fig. 4.3, points corresponding to the same state are linked by lines. On a map, objects within a map layer are linked by common visual properties such as colours or the shapes of symbols used for representing those objects.
7. Values of different attributes corresponding to the same value of a referrer must be represented by several graphical features combined within a single mark or by a visually linked set of marks. For example, the values of one numeric and one qualitative attribute may be encoded by the size and colour of a geometrical figure. When several marks are involved, some dimensions need to be used for placing them in a common frame-

work and for positioning the resulting composite marks in the framework of the entire visualisation. For example, the values of several numeric attributes referring to the same spatial location may be represented on a map by a bar chart, i.e. a visually linked sequence of bars. The horizontal spatial dimension is used for placing the bars in a bar chart. The space-embedding arrangement is used for placing bars on a map.

8. A redundant use of variables and dimensions is recommended when this allows better conformity with the principles or better display legibility. The latter means, in particular, visual differentiation, visual linking, and accuracy of perception, i.e. the capability of a viewer to determine correctly the corresponding references and attribute values for any of the marks and graphical features present in a display. Thus, retinal variables may be used to improve differentiation between positions of a dimension. For example, in a bar chart representing volumes of sales of different products, the horizontal dimension is used to represent the set of products. Additionally, the bars may be coloured distinctly for better differentiation.

Among these principles, there are two principles concerning the correspondence between the properties of visual primitives and the properties of data components represented by those primitives. In fact, we have formulated the same requirement separately for referrers (principle 2) and for attributes (principle 5). Let us elaborate on this requirement by specifying, in Tables 4.5 and 4.6, the properties of the dimensions and variables and indicating what types of data these visual primitives are most appropriate for. In these tables, we actually reformulate the definition of the perceptual properties of the visual variables provided by Bertin and other researchers.

We have not included in Table 4.6 the clarity variables of crispness, transparency, and resolution introduced by MacEachren. MacEachren admits that these variables are untested and can be used for no more than two or three categories. While these variables may be quite useful for the representation of uncertainty, their appropriateness for arbitrary attributes is unclear.

Among the visualisation principles, there are also requirements concerning visual differentiation and visual linking: marks corresponding to different references must be sufficiently differentiated, while marks or graphical features corresponding to the same reference must be visually linked. The techniques that can be used for differentiation and linking of marks are enumerated in Table 4.7.

**Table 4.5.** Properties of display dimensions and the types of data components that may be represented by them

Dimension	Properties	Most appropriate data types
A single horizontal or vertical dimension	Linear ordering; distances; continuity	Time (referrer or attribute) Numeric attribute
Two-dimensional plane	No ordering; distances; continuity	Two-dimensional space (in particular, geographic) Two numeric attributes
Three-dimensional display space	No ordering; distances; continuity	Three-dimensional space Two-dimensional space plus time (space–time continuum)
Display time	Linear ordering; distances; continuity	Time as referrer
Space partitioning: 1 dimension (sequence)	Linear ordering; distances; discreteness	Time as referrer
Space partitioning: 2 dimensions (grid)	Partial ordering; distances; discreteness	Two referrers, e.g. time and population
Space partitioning: diverging or parallel axes	No ordering; no distances; discreteness	Population referrer with a few elements Multiple numeric attributes
Space embedding	Properties of the container dimension	Depends on the properties of the container dimension
Space sharing	No ordering; no distances; discreteness	Population referrer
Space transformation: sectioning	Removes continuity	Population referrer Other non-continuous referrers
Space transformation: nodes and connectors	Removes continuity; removes distances; exhibits arbitrary relations	A referrer consisting of pairs of items linked by specific relations
Space transforming: polar coordinates	Removes or inhibits distances; introduces cyclicity	Cyclic time as referrer Spatial directions Proportions in a whole
Space transforming: attribute-based distortion	Removes distances; exhibits arbitrary quantities	Quantitative attribute referring to a division of two-dimensional space

**Table 4.6.** Properties of the retinal variables and the types of data components that may be represented by them

Retinal variable	Properties	Appropriate data types <sup>a</sup>
Size	Linear ordering; distances; continuity; true zero (zero size)	Ratio-scale attributes <i>Interval-scale attributes</i> <i>Ordinal-scale attributes</i>
Colour/hue	No ordering; no distances; discreteness	Nominal-scale attributes
Colour/saturation	Linear ordering; distances not readily perceived; continuity; true zero (no colour)	Ordinal-scale attributes <i>Temporal attributes</i> <i>Numeric attributes (ratio or interval scale)</i> <sup>b</sup>
Colour/brightness	Linear ordering; distances not readily perceived; continuity; true zero (black)	Ordinal-scale attributes <i>Temporal attributes</i> <i>Numeric attributes (ratio or interval scale)</i> <sup>b</sup>
Texture/arrangement	No ordering; no distances; discreteness	Nominal-scale attributes with a few values
Texture/density	Linear ordering; distances not readily perceived; discreteness; true zero (hollow texture)	Ordinal-scale attributes <i>Numeric attributes (ratio or interval scale)</i> <sup>b</sup>
Texture/size	Linear ordering; distances; continuity; true zero (hollow texture)	Ratio-scale attributes <sup>b</sup> <i>Interval-scale attributes</i> <i>Ordinal-scale attributes</i>
Shape and texture/shape	No ordering; no distances; discreteness	Nominal-scale attributes
Orientation and texture/orientation	No ordering; distances; continuity	Spatial directions Various directions (e.g. links in a node-connector arrangement)

<sup>a</sup> Data types that can be represented by the corresponding visual variables although their properties are not fully consistent with the properties of these variables are given in italics.

<sup>b</sup> Saturation, brightness, texture/density, and texture/size can represent absolute quantities only if applied to marks of equal sizes. For example, it is not recommended to use any of these variables on a map for filling the shapes of countries or districts according to their absolute population number. However, it is quite valid to use these variables for portraying relative numbers, such as population density or number of cars per capita.

**Table 4.7.** Techniques for visual linking and visual differentiation of marks and graphical features

Linking	Differentiation
Adjoining	Separation by space
Connecting (e.g. by line segments)	Separation by boundaries (sectioning)
Same colouring	Different colouring
Same shape	Different shape
Combining features in the same mark (e.g. size and colour)	Using distinct marks

In our practical work, we have dealt mostly with geographically related data and their visualisation on maps. The general principles of visualisation certainly apply to cartographic visualisation as well. Are there any specific principles of cartographic visualisation?

According to Bertin, the main difference between maps and other graphics is that the planar variables are used in maps for representing geographical space, and hence attributes can be portrayed only by retinal variables. For us, this is an embodiment of one of the general principles rather than something specific to maps: the display space (two- or three-dimensional) should be used for representing geographical space because the properties of the former are, out of all dimensions and variables, the most consistent with the properties of the latter. Since two planar dimensions and probably the third spatial dimension are already in use, all other data components have to be represented by other dimensions or by retinal variables.

So, what is specific to maps? Is it the methods of handling the curvature of the Earth when one is representing geographical space by a flat display space, i.e. projection techniques? Although geographical projections are, indeed, pertinent to maps, these are just particular methods of encoding elements of data by elements of graphics. Such encoding takes place for any component of data, not only for geographical space. The formulae or rules for the encoding differ from case to case, but these are different implementations of the same general principle. Thus, in representing a numeric attribute by size, one may choose a linear or a logarithmic scale and arbitrarily define the maximum size to be used for the maximum attribute value. Similarly, one may choose the Mercator or the Gauss–Krueger projection for representing geographical space on a plane.

In our reasoning concerning the specifics of cartographic visualisation, we came to the following conclusion. Geographical space, unlike the display space used to represent it, is extremely heterogeneous: oceans and

seas are very different from continents and islands, mountains are very different from valleys, forests differ from deserts, coasts differ from inland regions, cities differ from rural areas, countries differ from each other, etc. Any geographically related phenomenon is necessarily affected by this heterogeneity. Hence, it is very important for the analysis of geographically related data that not only the metric properties of geographical space are reflected in a visualisation but also the heterogeneity of this space. For this purpose, the representation of the phenomenon or phenomena under study must go together with portraying the geographical features that exhibit the heterogeneity of the geographical space, such as coasts, rivers, relief, state boundaries, cities, and roads. However, the relevance of each particular kind of feature may differ from situation to situation. Thus, state boundaries or roads may be relevant to demographic or economic analyses but not to global climate studies.

There are various types of maps, differing according to the intended use. Some maps are designed for showing the locations of various geographical objects and are used for orientation and navigation. Other maps portray the spatial distributions of some phenomena and/or the variation of their characteristics (attributes) over the space. These latter maps, which are called thematic maps, are often intended for the exploration of those phenomena (thematic maps can also be used for demonstrating already available knowledge concerning phenomena, i.e. the results of an exploration done earlier). Since this book is about exploratory data analysis, only thematic maps are considered in it. In thematic maps, any geographical features represented in addition to the phenomena under analysis are intended not for orientation or navigation but for reflecting the heterogeneity of geographical space, which may account for the distribution or variation observed.

The representation of the heterogeneity of geographical space follows the general principle of isomorphism, i.e. that any representation must be structurally similar (isomorphic) to the pertinent features of the situation for which the thinking is valid. This is quite analogous, for example, to the necessity of representing kinship relationships between family members in a study of the mechanisms of the transmission of genetic characteristic, in particular, proneness to certain diseases. In such research, it is not appropriate to treat the individuals just as elements of a homogeneous population, as is usually done in statistical studies.

Let us give some more examples to demonstrate that heterogeneity is pertinent not only to geographical space. In exploring the dependence of the physical and chemical properties of a substance on temperature, it is important to distinguish the temperature ranges in which the substance is solid, liquid, and gaseous. In an analysis of trends in stock prices, the temporal referrer is also not homogeneous: vacation times and holidays are

quite different from other periods, and the beginning of a financial year plays a different role from the end of the year. In climate or vegetation studies, winter differs radically from summer, and spring from autumn. It is appropriate to reflect the diversity of time periods in visualisation, although, unlike in cartography, there are no established methods for doing this.

Hence, we may conclude that cartographic visualisation is done fully in accordance with the general principles of visualisation. Therefore, we can apply our reasoning to visualisation in general and assume that it will be valid for cartographic visualisation as well.

Let us now analyse a few example displays from the perspective of their compliance with the principles of graphical representation.

#### 4.3.4 Example Visualisations

When we formulated the principle that referrers should be represented by display dimensions, we made the reservation that this principle can be violated in order to fulfil the other principles, and promised to provide appropriate examples. Let us now analyse an example of the representation of a temporal referrer by a retinal variable and, at the same time, an attribute by a dimension.

The map in Fig. 4.4 represents the routes of four storks from Europe to Africa. As we discussed in Chap. 2, this dataset has two referrers, time and the population of storks, and one attribute, geographical location. According to principle 1, the referrers should be represented by display dimensions, while the attribute, according to principle 4, may be represented by either a free dimension or a retinal variable. However, according to principle 5, the dimension or variable used to represent the attribute must have properties consistent with the properties of the attribute's value domain. The value domain of the attribute "location" is geographical space. Of all available dimensions and variables, the two-dimensional display space (the plane) has properties most consistent with those of geographical space. Moreover, the plane is suitable for representing the heterogeneity of the geographical space, since relevant geographical features can be depicted on it. Hence, the plane is the most suitable match for the value set of the attribute "location".

Of the remaining dimensions, those appropriate for representing time are the third spatial dimension, the display time, and one-dimensional display space partitioning.

Visualisation of movement using the third spatial dimension to represent time is known as the "space–time cube" technique. This technique is de-

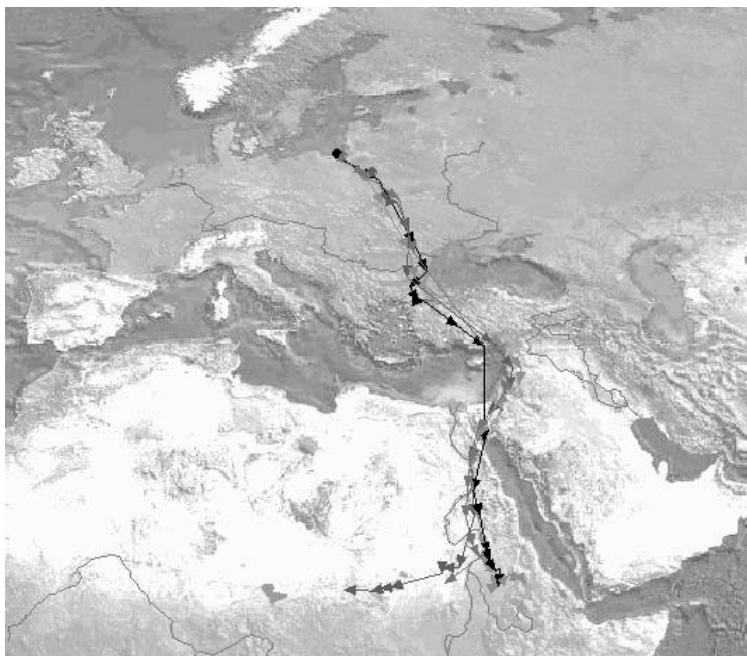
scribed, for example, by MacEachren (1995, pp. 252, 254). However, the third spatial dimension is available to its full extent only in special visualisation environments. On a computer screen or on paper, the third dimension can be merely simulated using depth cues, which have limited representational capabilities compared with the planar dimensions. In a space–time cube display on a plane, depth cues are used to represent one of the dimensions of geographical space. This representation is not isomorphic; in particular, distances cannot be correctly perceived, and directions are substantially distorted.

The use of the display time does not allow one to see the reference set as a whole, i.e. it contradicts principle 3. Partitioning of the display space (i.e. the use of the “small multiples” technique) also has its drawbacks. First, individual maps in a “small multiples” display have to be rather small, which reduces their legibility. Second, the space-partitioning arrangement provides very limited opportunities for visual linking of marks corresponding to the same value of a referrer. In the particular case of the storks, the positions of the same stork must be visually linked, according to principle 6. In a “small multiples” display, this visual linking can be provided only by using the same mark colour or the same mark shape throughout all the maps. This is more like a hint about relatedness than direct linking; it requires a significant cognitive effort by the viewer of the display to actually link the marks indicating the positions of the same stork. Third, a “small multiples” display cannot in general be perceived as a single image. Hence, a single-map display is preferable, at least for some tasks (we shall discuss in the next chapter what tasks can be better served by “small multiples” than by a single-map display).

Hence, none of the unused display dimensions is perfectly suited for representing the temporal referrer of our stork dataset, and some of the principles have to be compromised. One of the possible compromise solutions is to disregard principle 1 and to use marks and one or more retinal variables to portray time, for example as is done in the map in Fig. 4.4. We are far from regarding this map as an ideal visualisation, but it has its advantages (as well as its drawbacks).

In Fig. 4.4, time is represented by sequences of connected linear marks, one sequence for each stork. Each linear mark has a particular orientation, indicated by an arrow. The orientation portrays, on the one hand, the spatial direction of the movement, and on the other hand, the order of the time moments: the end of a mark corresponds to the moment following the moment correspondent to the beginning of the mark. The other referrer of the dataset, i.e. the group of storks, is represented by a space-sharing (overlay) arrangement. Visual differentiation of the movements of distinct storks is achieved by using individual shades for the marks corresponding

to each stork. Visual linking of marks corresponding to the same stork is achieved, additionally to the identical shading, by connecting the positions of that stork by line segments.



**Fig. 4.4.** A map depicting the routes of four storks. The sequence of time moments is portrayed by linear marks, with arrows indicating the direction of movement. Hence, a retinal variable, specifically orientation, is used to represent the temporal referrer of the dataset

An obvious drawback of this visualisation is that it does not allow absolute identification of time moments: without using additional tools, one cannot find out the date when each of the locations was visited. Only relative judgement is enabled: it is possible to ascertain which of two positions of the same stork was reached earlier. Moreover, this visualisation alone does not allow one to figure out the relative positions of different storks on the same date. It is hard to determine whether all four storks moved to the south synchronously, started their movement at different times, or moved with different speeds.

An obvious advantage is that the sequence of positions of each stork is seen as a single figure (trajectory), owing to an effective visual linking. Such a trajectory can be perceived in an instance of sight, in Bertin's words. Moreover, in the same instance of sight, it is possible to see all four

trajectories simultaneously and note their similarities and differences. Hence, a viewer can gain much information concerning the movement of the storks just from a first glance at the map.

Furthermore, if it is known that the measurements of the positions of the storks were made at regular time intervals, for example every day (this is, unfortunately, not completely true in our case), the representation of the trajectories provides important information about the speed of movement. Specifically, long line segments indicate fast movement (a long distance covered in a unit of time), while short segments correspond to slow movement.

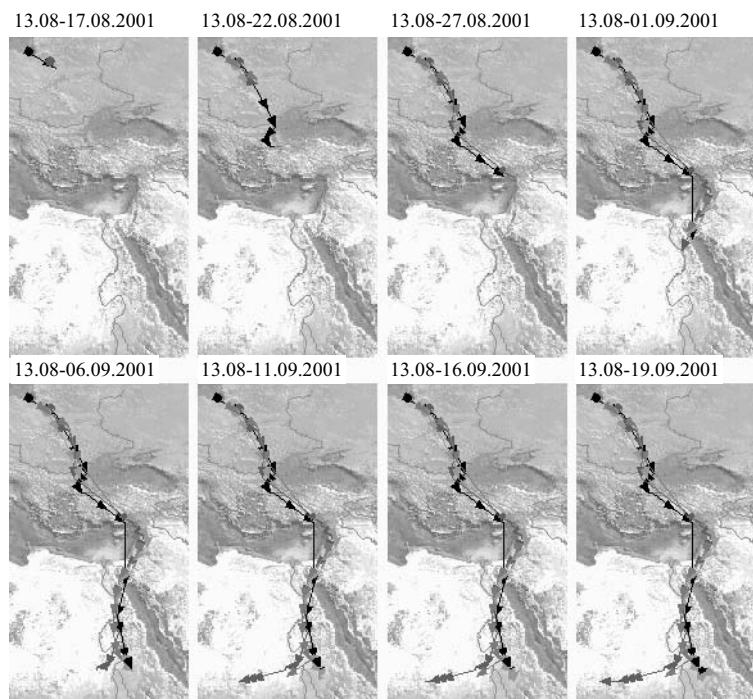
According to principle 8, we may try to use redundant variables or dimensions to improve the display. The main problem is that the display does not allow one to see what happened on a particular date and, as a consequence, determine which movements occurred simultaneously. Hence, the temporal component of the data needs to be represented in a more direct way than just by partial ordering of positions by means of lines with arrows (the ordering is partial because the order between positions of different storks is not specified).

As we have discussed, the display time, and space-partitioning arrangements are suitable candidates for representing a temporal referrer. Let us try to use one of them as a redundant dimension. In fact, both dimensions may be used in the same way. The illustrations that we have included in the book look like “small multiples”, although they are actually sequences of screenshots taken from an animated representation.

The visualisation shown in Fig. 4.4 can be combined with the use of an additional dimension (i.e. either the display time or a “small multiples” arrangement) in the following way. Each position in the chosen dimension corresponds to some time moment and contains a map representing the trajectories followed by the storks from the beginning of their movement until this time moment. When the display time is used, it is possible to have a position for each date during the whole period of the movement of the storks. For “small multiples”, one needs to consider the available display space. The dataset used for this example covers the period from 13 August to 19 September, i.e. there are 38 dates. A “small multiples” visualisation consisting of 38 maps would not be effective: the maps would be too small and hardly perceptible.

In Fig. 4.5, we have included a map for every fifth date in the stork movement period, which results in eight maps. To save space on the page, we have not reproduced the entire maps but have reduced them to small fragments, which are still sufficient for seeing the trajectories.

What are our gains from the redundant use of the additional dimension?



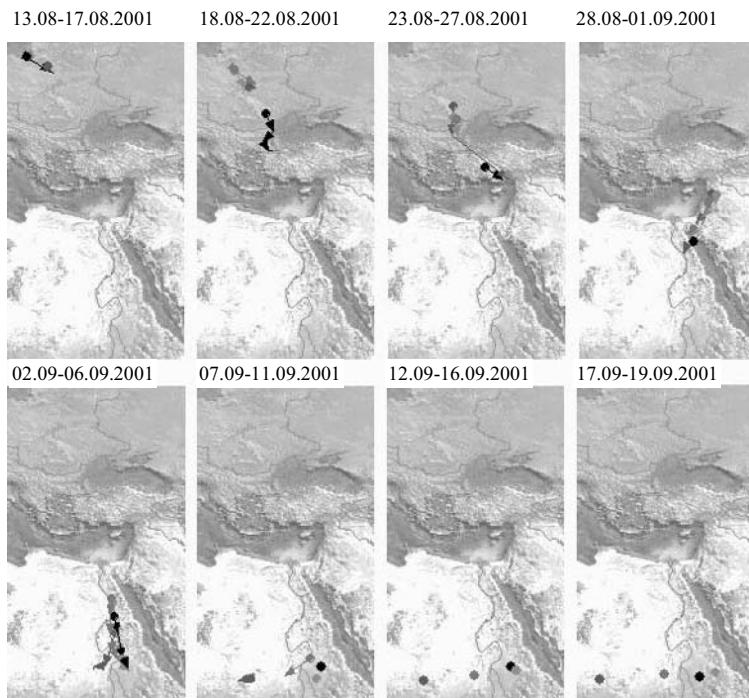
**Fig. 4.5.** A “small multiples” display representing stork trajectories over periods of 5, 10, 15, 20, 25, 30, 35, and 38 days, respectively, from the beginning of the movement

- It is now easier to relate the positions of the storks to absolute time moments. The “small multiples” display in Fig. 4.5 allows one to do this with 5-day accuracy. Complete accuracy may be achieved by means of including a map for every date in the visualisation. This is quite possible when the display time is used rather than the “small multiples” arrangement.
- The relative positions of different storks on the same date can be easily figured out.
- Not only can the order between the positions of any particular stork be ascertained, but also the order between the positions of different storks. Therefore, it is now possible to determine whether or not all four storks moved to the south synchronously. Thus, in our particular case, it may be seen that the storks started their movement at different times. However, since the birds initially moved with different speeds, their movements eventually became synchronised: the fourth and fifth frames demonstrate a coherent movement of all birds to the south. This corresponds to the period from 28 August to 6 September.

All these benefits result from a better opportunity to determine the temporal references corresponding to the attribute values (i.e. stork positions) represented in the display. In other words, these are consequences of principle 6 being more properly fulfilled.

Are there any losses? Well, the whole visualisation cannot now be perceived as a single image. However, each individual map retains this property, as well as all other advantages of the map shown in Fig. 4.4 (actually, this map is present in the new visualisation: it is the last map in the sequence). Hence, we have managed to improve the visualisation by means of a redundant use of an additional dimension.

The additional dimension could also be introduced in other ways. Fig. 4.6 demonstrates an alternative “small multiples” visualisation of the same data. The difference from the previous visualisation is that each individual map represents the movements during a 5-day time interval rather than the trajectories from the beginning of the whole period.



**Fig. 4.6.** A “small multiples” display representing stork movements during 5-day intervals

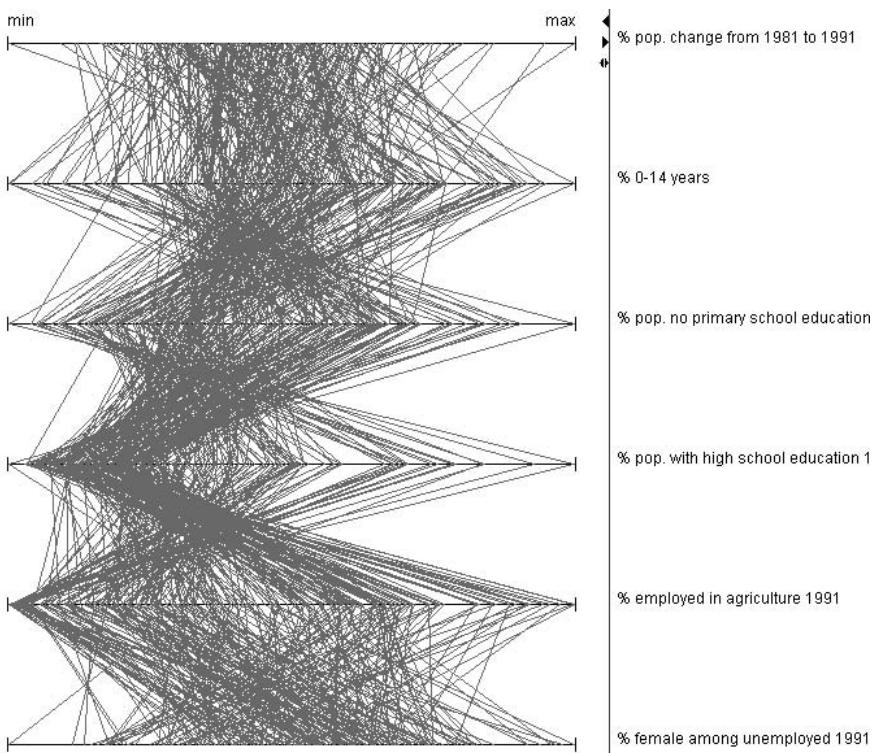
This visualisation is, perhaps, even better suitable for the investigation of movement synchronisation than that in Fig. 4.5. The periods of fast

movement and the periods of staying in the same place are also better identifiable. On the other hand, the visualisation in Fig. 4.6 supports, in Bertin's terms, only the intermediate and not the overall reading level: a viewer can see movements during subintervals of the whole period but not the entire trajectories. In our terms, the visualisation is not supportive of the task of describing the overall behaviour of each individual stork and that of all storks together (but it is more appropriate for some other tasks).

Let us now discuss a quite different visualisation of another dataset. Fig. 4.7 demonstrates a “parallel coordinates” display of the demographic data related to the districts of Portugal. As we described in Chap. 2, this dataset contains two referrers, one spatial (a set of districts in Portugal) and one temporal (a set of two years, 1981 and 1991), and a number of attributes characterising the population of the districts. For the display in Fig. 4.7, only data from the year 1991 have been selected; hence, the temporalreferrer does not need to be represented. In fact, for building the display, we have disregarded the spatial nature of the other referrer and treated it as a population referrer. Therefore, this display alone is, of course, not an adequate representation of the dataset; it can be used only in combination with other tools. However, for the purposes of this discussion, let us assume that the set of districts is a referrer of population type (since we have always been dealing with spatially referenced data up to this point, we have no suitable non-spatial dataset at our disposal).

The visualisation in Fig. 4.7 has been designed to represent some characteristics of the districts of Portugal in terms of values of six attributes. The names of the attributes can be seen on the right of the display; their meanings are explained below.

- “% pop. change from 1981 to 1991”: The relative change of the population that occurred in each district, i.e. the difference between the population numbers in 1981 and 1991 expressed as a percentage of the population in 1981.
- “% 0–14 years”: The proportion of people aged from 0 to 14 years in the population of each district in 1991.
- “% pop. no primary school education”: The proportion of people who did not receive even a primary school education in the population of each district in 1991.
- “% pop. with high school education”: The proportion of people who have received a high school education.
- “% employed in agriculture”: The proportion of the population of the district employed in agriculture.
- “% female among unemployed”: The proportion of women among unemployed people.



**Fig. 4.7.** A “parallel coordinates” display of demographic data related to districts of Portugal. The values of several attributes are represented by positions on parallel axes. Positions corresponding to the same district are connected by lines

According to principle 7, the values of several attributes corresponding to the same reference should be represented so as to be visually linked. There are two ways of doing this: the attribute values may be portrayed by the visual properties of the same mark (e.g. size and brightness) or by several marks linked by adjoining, connection, or another appropriate technique (see Table 4.7). In the latter case, it is necessary to use some dimensions for arranging the marks.

In a parallel-coordinates display, the values of each attribute are represented as positions on an axis, i.e. using one of the planar dimensions. This is consistent with the properties of the value sets of the attributes: all of them are numeric. In our variant of parallel coordinates, the horizontal dimension is used for this purpose, but it is equally possible to use the vertical dimension.

In order to include several attributes in the visualisation, a space-partitioning arrangement is used: the remaining planar dimension (in our

case, the vertical dimension) is partitioned into six segments, and in each segment, the axis of one of the attributes is placed. As a result, the display contains six parallel horizontal axes (in another realisation, they might be vertical).

In order to fulfil the requirement that attribute values corresponding to the same reference must be visually linked, positions on adjacent axes corresponding to the same district are connected by line segments. As a result, the characteristics of each district are represented by a polygonal line. The set of all lines represents the reference set, i.e. the set of districts. The lines are overlaid within the same display space; hence, a space-sharing arrangement is involved. Since we have decided to treat the set of districts as a population-type referrer, a space-sharing arrangement is an appropriate match for it.

Although most of the visualisation principles have been fulfilled, the parallel-coordinates display still does not comply with principle 6: it does not permit unambiguous ascertaining of which reference corresponds to each mark. Thus, a viewer cannot determine which district is represented by each polygonal line. This means that the lines are not sufficiently differentiated. Can we, according to principle 8, involve an additional variable or dimension in order to cope with this problem?

In choosing such a variable or dimension, we must take into account the number of references that need to be differentiated: the variable or dimension must have a sufficient number of different values (i.e. these values need to be perceived as different). The number of districts in Portugal is about 300. None of the retinal variables has so many different values. Among the dimensions, the two-dimensional display space and display time have sufficient capacity. However, the display space is already fully in use, and the display time does not support seeing the entire reference set as required by principle 3. Besides, the display time is a linearly ordered dimension and therefore is not quite suited for representing a population-type referrer.

Hence, the problem cannot be solved effectively by purely visual methods but requires the visualisation to be combined with other tools, for example, querying tools, which will be discussed later.

Let us now try to visualise the same data as in Fig. 4.7 in a different way, so that the spatial (more specifically, geographical) nature of the referrer is appropriately accounted for. As we have already discussed, the most suitable medium for representing geographical space is a map. So, let us apply cartographic visualisation.

Figure 4.8 demonstrates a part of a map that we have built (in order to reduce the overlap of marks, we have zoomed in on the central and southern parts of the territory).



**Fig. 4.8.** The same attributes as in Fig. 4.7, represented on a map of Portugal by bar charts

Here, the two-dimensional display space is used to portray the geographical referent, i.e. the set of districts. The values of the six attributes are represented by bar-shaped marks. In accordance with principle 7, the values of different attributes corresponding to the same reference are visually linked by arranging the respective set of marks in an appropriate way. Specifically, the horizontal spatial dimension is used for putting the marks (i.e. bars) in a common framework. For better suitability for this purpose, this dimension has been transformed by means of sectioning, which neutralises the continuity of the dimension (see Table 4.4). Hence, each attribute is given its particular section in a one-dimensional space. The bars corresponding to the attributes are visually linked by means of adjoining. This produces composite marks – bar charts. The bars for different attributes are shaded differently for better differentiation.

Each bar chart uses the vertical spatial dimension to represent the attribute values. The space of each chart, which is outlined by a frame, is embedded in the space of the map. Within a chart, the attribute values are portrayed by positions in the vertical dimension: the lower end of the vertical axis corresponds to the minimum attribute value, and the upper end to the maximum value. The formula for the conversion from attribute values to vertical positions is specified individually for each attribute; hence, different bars of the same height do not represent the same numeric value.

Unfortunately, the charts also allow another interpretation: a viewer may believe that attribute values are represented by the sizes of the bars rather than by the vertical positions. This interpretation is wrong, since the sizes of the bars are not proportional to the attribute values. Thus, a bar of zero size does not represent a value of zero but rather the minimum value of the respective attribute out of the values available in the dataset. Another possible mistake is to regard the bars as comparable, i.e. having the same correspondence between attribute values and positions or bar heights. In order to preclude false interpretations, the meaning of the symbols and the rules for the encoding of values must be appropriately explained, for example in the map legend.

Unlike the parallel-coordinates display, the cartographic representation supports quite well the identification of the reference corresponding to each mark: the marks are positioned on the map within or near the boundaries of the corresponding districts.

Still, as with almost any visualisation, there are some problems that cannot be solved by merely visual means. In this case, the most critical problem is symbol overlap. It can be partly solved by increasing the size of the map and decreasing the sizes of the symbols, but these opportunities are quite limited regarding the available display space and the viewer's capabilities for discerning symbols. Hence, as in the previous case, additional tools are needed.

Neither the parallel-coordinates display nor the map with bar charts permits efficient, pre-attentive perception of the entire visualisation. This is not surprising, since the number of data components (one spatial referrer plus six attributes) exceeds the maximum number of visual components that can form a single image, i.e. three, according to Bertin (two planar and one retinal), or four, according to Green (three spatial and one retinal).

By discussing these examples of visualisations, we wanted to demonstrate how visualisation design is guided by the basic principles of visualisation. It is not always possible to satisfy all of the requirements, and some of the principles have to be compromised. Besides compliance with the visualisation principles, data displays must satisfy requirements of other kinds, such as legibility and the preclusion of false interpretations.

Despite the high importance of visualisation for EDA, it is usually not sufficient to use only visualisation. We have just shown a few cases where the deficiencies of data displays could not be compensated by purely visual means. In such cases, visualisation needs to be combined with other tools. Let us now go on to consider other groups of tools for EDA.

## 4.4 Display Manipulation

As we have already mentioned, Bertin considers graphics intended for information processing (i.e. exploratory analysis) as something to “play” with rather than passively view: “We class and order these images in different ways, grouping similar ones, constructing ordered images to discover the synthetic schema which is at once the simplest and the most meaningful” (Bertin 1967/1983, p. 164).

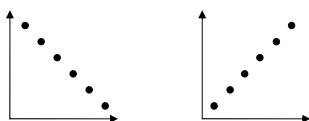
In general, Bertin sees information processing as a process of logical simplification, which can be performed verbally, mathematically, or graphically. Graphical information processing operates by simplification of an image. Bertin refers to Georges Th. Guilbaud, who characterises a simple visual form by two qualities: connectivity, which means not having gaps, that is, being homogeneous, or avoiding meaningless intersections in a network; and convexity, which means being delimited by convex angles and thus forming a uniform area, inside of which any straight line will cross the figure only once. Any visual simplification must tend towards these characteristics. This can be achieved in two ways: (a) by ordering a qualitative component; and (b) by eliminating certain correspondences in ordered components.

### 4.4.1 Ordering

In our terms, ordering may be defined as changing the positions of marks or parts of a graphic within a dimension, for example, changing the positions of bars in a bar chart or the positions of axes in a parallel-coordinates display. According to Bertin, arbitrary ordering of parts of a graphic is possible when there is no natural ordering between the values of the data component that they represent, i.e. the component is qualitative. We can add that it is also possible to reorder parts of graphics corresponding to different attributes (this is also a case of the absence of a natural ordering).

Bertin stressed that simplification of an image by ordering preserves all the information originally contained in it. Bertin described two reordering techniques: diagonalisation of diagrams and transformation of networks.

Diagonalisation of a diagram means bringing it as close as possible to one of the forms shown in Fig. 4.9. This may be done, for example, by permuting rows and columns of a table. Bertin described some instruments that could be used for performing such operations in the pre-computing era.

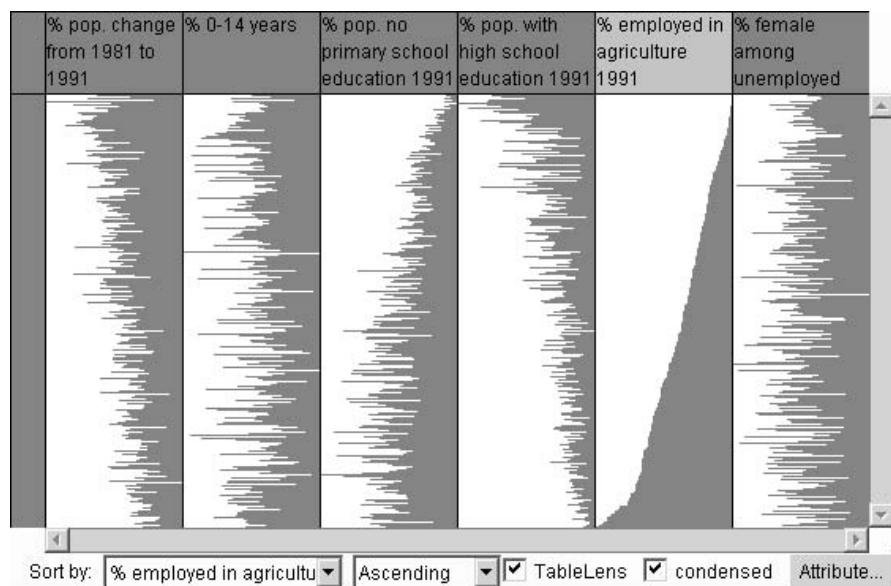


**Fig. 4.9.** A “perfect” correspondence between two components reveals itself in a diagonal form

Nowadays, permutations in a table are performed by means of interaction with a computer. Besides manual reordering, the user can apply automatic sorting procedures. For example, Fig. 4.10 demonstrates a table display of six demographic attributes related to the districts of Portugal (these are the same attributes that we used for our example visualisations in the previous section). As we have mentioned earlier, a table may contain not only figures or texts but also graphical elements. In Fig. 4.10, attribute values are represented in table cells by horizontal bars stretching from right to left. In this way, the horizontal display dimension is used. The rightmost position in a column corresponds to the minimum value of the respective attribute, and the leftmost position to the maximum value.

Figure 4.10 shows the result of applying an automatic procedure of ordering table rows according to the values of the attribute “% employed in agriculture 1991”. It may be seen that the arrangement of the bars in the corresponding column (second from right) forms a nearly triangular shape, i.e. it is close to the diagonal form recommended by Bertin. Of course, it is impossible to achieve the same effect in all columns simultaneously. Nevertheless, ordering according to one of the attributes may be very useful for detecting correlations between attributes. Thus, one can learn from Fig. 4.10 that as the proportion of people working in agriculture increases, the proportion of people without primary school education also tends to increase, while the proportion of people having high school education mostly decreases.

In relation to the illustration in Fig. 4.10, we would like to note that a proper ordering or, more generally, arrangement of display items is very important in such “condensed” views, in which large amounts of data are fitted into a limited display space at the cost of decreasing the sizes of the marks representing the elements of the data.



**Fig. 4.10.** A table display of six demographic attributes related to the districts of Portugal. The values of the attributes are represented by horizontal bars in the table cells: zero length corresponds to the minimum attribute value available in the column, and the maximum value is represented by a bar using the whole cell width. The rows of the table have been automatically ordered according to the values of the attribute “% employed in agriculture 1991”

In our example table in Fig. 4.10, the height of the rows is so much reduced that it is impossible to determine what attribute values each row contains and what district it corresponds to. However, the ordering allows us to make useful observations concerning the entire dataset, specifically, to detect correlations between attributes.

In an extreme case of information condensation, one data item may be represented by just a single pixel on a display, which certainly makes it indistinguishable. However, an appropriate arrangement of the visual items, such as the grouping of pixels with identical or similar colours, allows the viewer to obtain useful synoptic information about the properties of the entire dataset (Keim and Kriegel 1994).

Transformation of networks is the second application of ordering considered by Bertin. The goal is to arrive at a construction that has the fewest meaningless intersections, while preserving any groupings, oppositions, or potential orders contained in the data component represented in network form. When the information is not too complex, Bertin recommends a circular construction as the most facilitative for discovering the optimal ordering of the nodes. When the information is highly complex, a permutable

matrix affords a means of proceeding to an initial simplification prior to construction of the network. The rows and columns of such a matrix correspond to the nodes of the graph, and marks in the cells indicate the presence of links between the respective nodes.

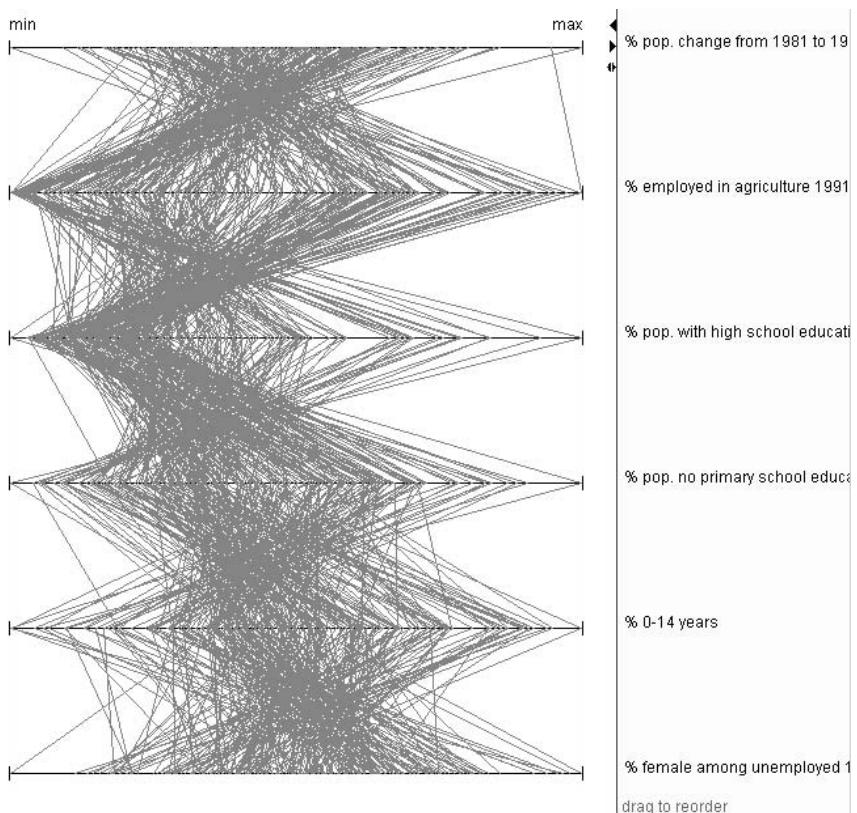
Currently, there are numerous software tools for building and displaying graphs. These tools often include computational procedures for optimising graph layout. Also, they allow the user to reorder graph nodes interactively, for example by dragging them to desired positions.

As one more example of reordering, let us consider the reordering of axes in a parallel-coordinates display (see Fig. 4.7). The purpose of such reordering is to reveal correlations between attributes. The presence of a correlation between two attributes may be detected if their axes are adjacent. When the lines between these two axes are mostly parallel, this indicates a positive correlation: low values of one of the attributes correspond to low values of the other attribute, and, similarly, high values correspond to high values. A negative correlation makes most of the lines stretch from one end of one of the axes to the opposite end of the other axis, which results in an X-like figure between the axes. Thus, in Fig. 4.7, one can observe X-shapes between the axes of the attribute pairs “% 0–14 years” and “% pop. no primary school education”, between “% pop. no primary school education” and “% pop. with high school education”, and between “% pop. with high school education” and “% employed in agriculture”. These shapes indicate that these pairs of attributes are negatively correlated.

Since a parallel-coordinates display exposes correlations only between attributes whose axes are adjacent, it is reasonable to make the axes reorderable so that more correlations can be discovered. Therefore, most software implementations of the parallel-coordinates display allow the user to change the order of the axes interactively, and some of them provide automatic procedures that optimise the order of the axes to reveal the maximum number of correlations.

Let us now take a look at Fig. 4.11, which shows the same display as in Fig. 4.7 but with an alternative ordering of the axes. The axes have been interactively reordered so as to demonstrate the maximum number of correlations: a clear X-shape can be seen between each pair of neighbouring axes (hence, the correlations are negative). At the same time, the display exposes several cases that do not obey the general rule. They manifest themselves as lines having an atypical inclination, such as the line on the right between the upper two axes: this line is almost vertical, while all neighbouring lines are diagonal. Such cases (called outliers) require special attention: an explorer needs to find an explanation for their unusualness.

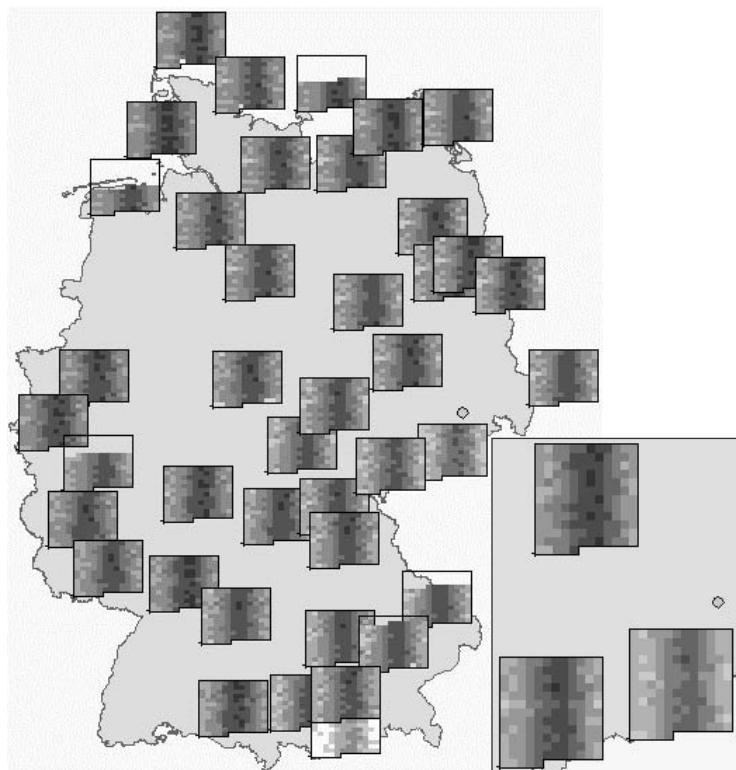
It is hard to state definitely whether the transformation of the parallel-coordinates display demonstrated in Fig. 4.11 resulted in a simplification of the image. Nevertheless, it was quite useful, since it allowed us to discover some correlations that had not been seen before, such as the negative correlation between the attributes “% pop. change from 1981 to 1991” and “% employed in agriculture 1991”. Hence, striving for image simplification is not the only meaningful reason for trying to transform a display. Although simplification often increases understanding, this does not mean that better understanding can never be achieved without simplification.



**Fig. 4.11.** The axes in the parallel-coordinates display of Fig. 4.7 have been reordered here so as to reveal the strongest correlations between the attributes

The technique that we would like to demonstrate with the next example may also be seen as a kind of ordering in the sense that marks are placed in a convenient order. However, instead of the relative positions of individual marks within a single display dimension being changed, the marks are organised into equal-sized groups, which are placed next to one another in an

additional display dimension. In other words, a one-dimensional arrangement of marks is turned into a two-dimensional arrangement. Such an arrangement is shown in Fig. 4.12, where the signs on the map represent the monthly average temperatures measured at different weather observation stations in Germany during the time period from January 1991 to May 2003.

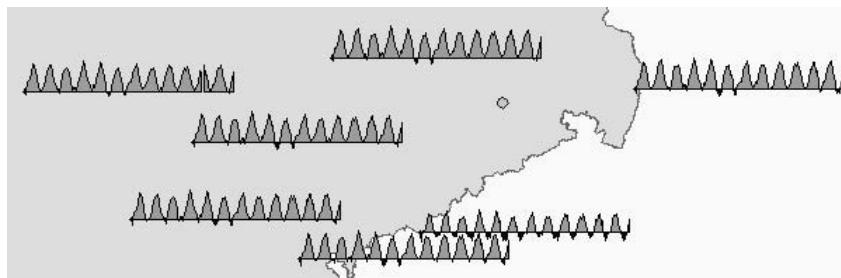


**Fig. 4.12.** The signs represent the monthly average temperatures for the time period from January 1991 to May 2003. For each month, there is a small square mark, which is shaded according to the temperature value in the month, within each sign: the higher the temperature, the darker the shade. The squares are organised in rows, with 12 marks per row. Hence, each row represents a year. The rows are put one below the other; hence, each column represents a certain month of the year

For each weather observation station, there is a sign (let us call it a “mosaic”) consisting of a number of small square marks, or “tiles”. Each tile corresponds to a certain month within the observation period. The tile is shaded according to the temperature value in that month: the higher the

temperature, the darker the shade (i.e. the retinal variable “brightness” is used). The tiles within a mosaic are organised in rows, with 12 tiles per row. Hence, each row in a mosaic represents a year, and each column represents a certain month of the year. In the lower right corner of the figure, a few mosaics are enlarged for better visibility.

In this example, we have used two dimensions of the display space (here we mean the space of each mosaic sign, which is embedded in the space of the map) to represent a temporal referrer, although time is a linearly ordered set, which could be represented by a single dimension. Thus, instead of the two-dimensional mosaics, the data could be portrayed by one-dimensional sequences of tiles or by time graphs, as on the map fragment in Fig. 4.13. However, time has a dual nature: it is both linear and cyclic. The cyclic aspect of time is not always relevant, but in this case it plays a very important role and must be taken into account in the exploration of the data. Therefore, a two-dimensional arrangement of marks corresponding to time moments is much more appropriate in this and similar cases than is representing a temporal component by a single dimension. In such a two-dimensional arrangement, one dimension is used to represent the appropriate circle, e.g. 12 months of a year, 7 days of a week, or 24 hours of a day, and the other dimension is utilised to juxtapose representations of consecutive cycles.



**Fig. 4.13.** The monthly temperatures (i.e. the same data as in Fig. 4.12) are represented by time graphs embedded in a map

If we compare the visualisations of the same data in Figs 4.12 and 4.13, we shall find the one in Fig. 4.12 both simpler and more informative than the other. We can detect locations with generally lower or higher temperatures than in other places: the corresponding mosaics look lighter or darker as a whole. Mosaics where the central part is much darker than the margins indicate locations with a high contrast between summer and winter temperatures. Locations where the highest or lowest temperatures were observed are also easily detectable. Comparison of rows in a particular mo-

saic sign allows an analyst to detect warmer and colder years (or springs, summers, Januaries, etc.).

A two-dimensional arrangement is useful not only for visualising periodic data with a previously known period length; it can also be used as an instrument for detecting periodicity in data and determining the length of the period. For this purpose, an analyst interactively changes the number of marks to be put in a row and observes the resulting display until some regular pattern emerges (if the data are indeed periodic, something like vertical stripes can be expected when the row length corresponds to the length of the period). Investigations of this kind have been done, for example, by the Russian mathematician Zenkin, who discovered a periodicity in the distribution of various properties of the set of natural numbers. One of his findings is described in Zenkin (1990).

#### 4.4.2 Eliminating Excessive Detail

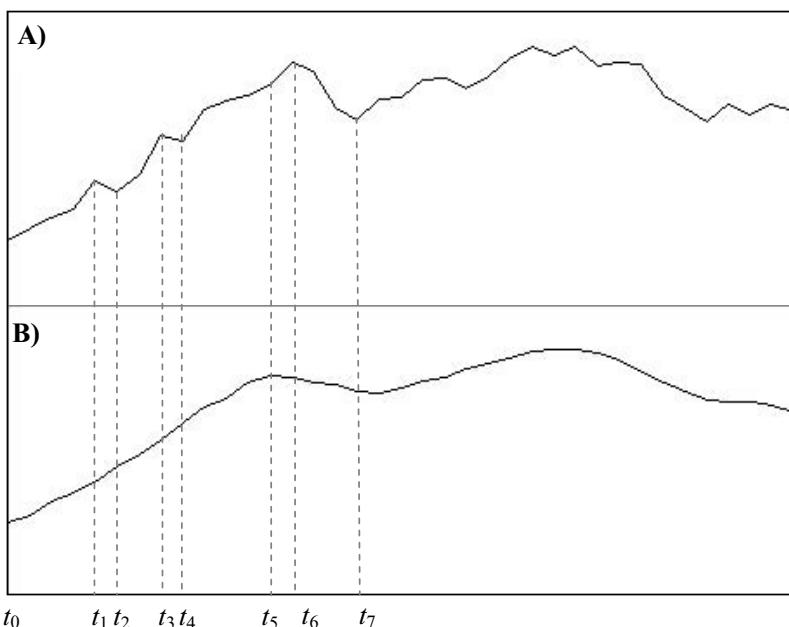
The second of the two approaches to image simplification considered by Bertin consists in eliminating part of the information from the display. According to Bertin, an ordered data component does not permit reordering of the graphical elements representing it; hence, simplification can only be achieved by eliminating a number of details. As examples, Bertin mentions the smoothing of curves in diagrams, and regionalisation and generalisation in maps.

The technique of smoothing is demonstrated in Fig. 4.14. The upper curve is a time graph representing the variation of the burglary rate in the state of New Mexico in the USA over the time period from 1960 to 2000. After smoothing has been applied, the curve takes the form shown in the lower part of Fig. 4.14. We shall not explain here how the smoothing algorithm works (this information can be found in numerous handbooks on statistics). The general idea is to eliminate fluctuations and thereby expose the general trend of the changes.

If we compare the upper curve in Fig. 4.14 with the lower one, we can notice that the upper curve has small peaks at time moments  $t_1$  and  $t_3$  and troughs at moments  $t_2$  and  $t_4$ , which are absent in the lower curve. Instead, the lower curve exhibits a steady growth of the burglary rate during the time interval from  $t_0$  to  $t_5$ , which includes the interval from  $t_1$  to  $t_4$ . Hence, the lower curve allows us to disregard the minor oscillations in the burglary rate that occurred at the moments  $t_1$ ,  $t_2$ ,  $t_3$ , and  $t_4$ , and see more clearly the general increasing trend.

However, not all sudden rises or drops can be regarded as minor and unimportant. Thus, the peak at the moment  $t_6$  followed by a rather steep drop

looks like something more serious than just a random fluctuation. Perhaps something happened at  $t_6$  that caused the crime rate to go down. An explorer might be interested in investigating this case; in particular, in looking for additional information. However, the distinctive feature consisting of a peak at  $t_6$  followed by a drop during the interval  $(t_6, t_7)$  does not appear in the smoothed curve B. It seems that that curve is oversmoothed: it hides not only small fluctuations but also more prominent features, which may potentially be important for understanding the phenomenon under analysis. Such distinctive features require special attention from the explorer, but may easily be skipped as a result of excessive smoothing and simplification.

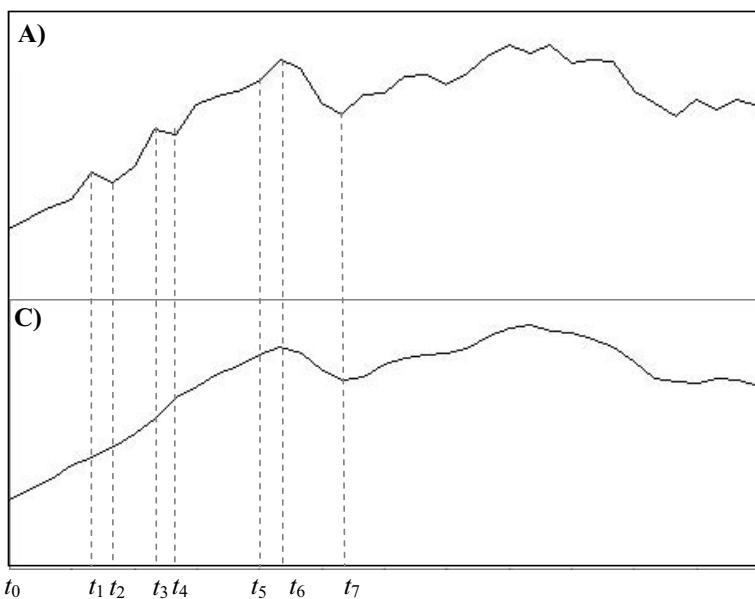


**Fig. 4.14.** Curve A represents the variation of the burglary rate in the state of New Mexico. Curve B results from smoothing the curve A using the “moving average” technique (averaging over five consecutive time moments)

Figure 4.15 represents the result of smoothing the same curve (i.e. the variation of the burglary rate in New Mexico) with different parameters. Specifically, curve B in Fig. 4.14 was obtained by averaging over five consecutive years, while curve C in Fig. 4.15 was obtained by averaging over three consecutive years. It may be seen that curve C preserves the characteristic features of curve A at the moments  $t_5$  and  $t_6$  better, while removing

the minor fluctuations at  $t_1$ ,  $t_2$ ,  $t_3$ , and  $t_4$  and after  $t_7$ . The removal of the fluctuations makes the overall shape of curve C much simpler than that of the original curve A although not so simple as the shape of curve B. However, curve B eliminates too much information, including potentially significant features. Therefore, curve C is more appropriate for exploring the burglary rate data than curve B is.

Hence, smoothing must be done in such a way that, on the one hand, an appropriate degree of simplification is achieved, but on the other hand, distinctive features are preserved (and maybe even sharpened, to attract the attention of the explorer). In fact, this requirement concerns not only curve smoothing but also all other simplification techniques involving information loss, for example generalisation on maps.



**Fig. 4.15.** The curve of the variation of the burglary rate in New Mexico has been transformed here using different smoothing parameters (specifically, averaging over three consecutive time moments). The prominent peak at  $t_6$  and the trough at  $t_7$  have been preserved in curve C, while the minor fluctuations at  $t_1$ ,  $t_2$ ,  $t_3$ , and  $t_4$  have been removed

Map generalisation is a complex process, which involves many aspects: simplification of lines and contours, elimination of small objects, merging similar small objects into larger shapes (e.g. separate houses into blocks of houses), replacement of contours by symbols, and so on. These procedures may be characterised as graphical methods of generalisation and simplifi-

cation. Although there are software tools for automated map generalisation, they are intended mostly for cartographers rather than data analysts, and are included in systems for professional map production rather than for exploratory analysis.

However, graphical generalisation is not the only possible means of simplification of a map. The easiest way to make a map (on a computer screen) simpler is to remove irrelevant groups of objects or phenomena. Objects or phenomena of the same kind are usually organised into “map layers” (according to a GIS terminology). The opportunity to switch the drawing of any layer on and off is a standard function available in any mapping software.

Simplification may also be achieved by means of generalising quantitative and qualitative characteristics. This approach is often referred to as classification. It is often used in thematic maps, but actually has a more universal applicability; therefore, let us consider it in more detail.

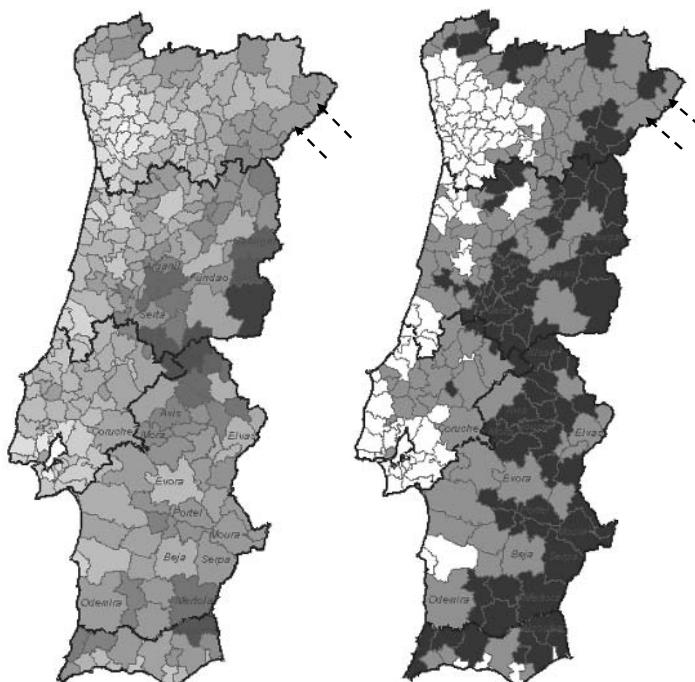
#### 4.4.3 Classification

Let us start with an example. Figure 4.16 shows two maps, on which the proportions of elderly people (aged 65 years or more) in the districts of Portugal are represented using the retinal variable “brightness”: lighter shades correspond to smaller values, and darker shades to higher values.

The difference between the maps is in the way in which the values of the attribute “% 65 or more years” have been encoded by values of the retinal variable “brightness”. On the left, the values from the minimum (6.7) to the maximum (35.2) are matched with a gradual scale of increasing darkness, as illustrated in Fig. 4.17. On the right, the value range of the attribute (i.e. from 6.7 to 35.2) has been divided into three subintervals, or classes, with class breaks at 15 and 20. All values within each of the intervals are represented identically; hence, only three different shades are used: the lightest shade (white) for the interval (6.7, 15), a medium shade for the interval (15, 20), and the darkest shade for the interval (20, 35.2). According to cartographic terminology, the map on the left is called an unclassified choropleth map, and the map on the right a classified choropleth map.

The classified choropleth map looks much simpler than the unclassified one, which contains very many different shades. The simplification has been achieved at the cost of losing a considerable amount of information: values in identically shaded districts may differ rather much, but the differences cannot be perceived. At the same time, some minor differences are exaggerated because quite close values may happen to fall into differ-

ent classes. Thus, the values in two districts in the north-east (indicated by dashed arrows) are quite close to the values in the neighbouring districts. This can be seen well from the unclassified map: these two districts are shown in nearly the same shade as their neighbours. On the classified map, however, they are shaded differently from their neighbours. This is because the proportions of elderly people in these two districts (19.87 and 19.90) happen to be slightly below the class break at 20%, while their nearest neighbours, with 20.14% and 21.13%, fit into the next class.



**Fig. 4.16.** Both maps here represent the attribute “% 65 or more years” by means of brightness, with lighter shades corresponding to smaller values and darker shades to higher values. On the left, the values from the minimum (6.7%) to the maximum (35.2%) have been matched with a gradual scale of increasing darkness. On the right, the value range from 6.7 to 35.2 has been divided into three subintervals, or classes, with class breaks at 15 and 20. All values within each of the intervals are represented identically, i.e. only three different shades are used



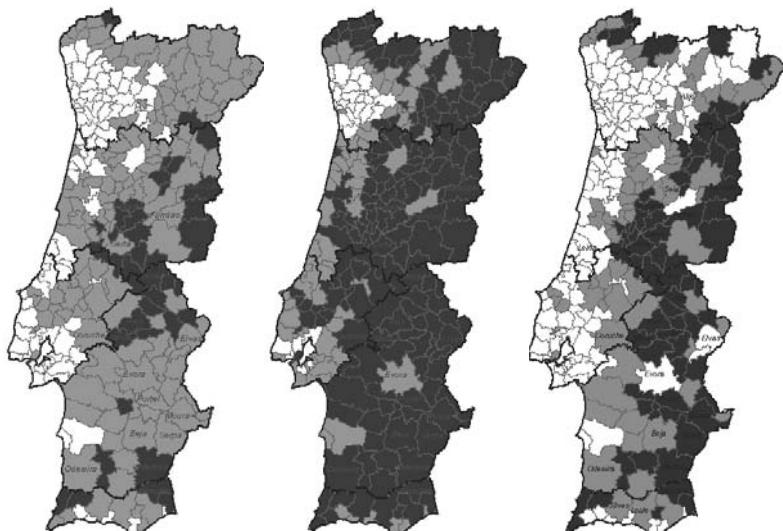
**Fig. 4.17.** The principle of encoding attribute values from the minimum to the maximum by gradually increasing degrees of darkness

These fallacies refer mostly to the elementary level of analysis; in contrast, a classified map may be quite advantageous for the overall analysis level, i.e. for synoptic tasks. Thus, the classified map in Fig. 4.16 shows some distinctive features of the spatial distribution of the attribute values more clearly than does the unclassified map. Specifically, we can easily identify low proportions of elderly people near the western coast, with two big clusters in the north and in the middle (around the two biggest cities in the country, Porto and Lisbon, respectively), and high values in the inland areas. The classified map helps an explorer to disregard superfluous detail and to grasp the general character of the distribution pattern. As we know, such grasping plays a very important role in exploratory data analysis.

This does not mean, however, that classified choropleth maps are always superior to unclassified ones as tools for accomplishing synoptic tasks, in particular for finding a pattern (i.e. a parsimonious description) approximating the overall behaviour of a phenomenon. In the previous chapter, we discussed various types of patterns. One of them is a trend pattern; for example, in spatially related data, an increase in the values from the north to the south or from the centre of the territory to its periphery. Another type is an association pattern. To derive such a pattern, one divides the territory into a possibly smaller number of coherent regions with low variations of attribute values within those regions. This procedure is often referred to as “regionalisation”. Unclassified maps are better suited for detecting trends, because they do not hide differences. Classification discards differences between values within a class interval and gives the corresponding objects a similar appearance on the map. When these objects are geographical neighbours, they tend to be visually associated into clusters. This property makes classified maps very suitable for regionalisation. It depends on the data and not on the preferences of the analyst which of the two methods of simplification is possible or more effective in any particular case. Therefore, the explorer needs both to look at an unclassified choropleth map and to apply a flexible classification tool, in order to find the most appropriate pattern for data with a previously unknown character of their spatial distribution.

It was not accidental when we said that an explorer needs a flexible classification tool rather than simply a classified choropleth map. The reason is that a single static classified map cannot support regionalisation appropriately. It is well known in cartography that different selections of the number of classes and class breaks may radically change the spatial pattern perceived from a map (see, for example, MacEachren (1994) and Slocum (1999)). Thus, let us consider Fig. 4.18, which shows three more variants, in addition to that shown in Fig. 4.16, of the classification of the districts of Portugal according to the proportion of elderly people. In all three maps,

the number of classes is the same, specifically three. On the left, the method known as statistically optimal classification has been applied, which produces class breaks at 15.24 and 22.69. The general idea of this method is to minimise, for the specified number of classes, the variation within the classes and to maximise the differences between them. A more detailed description can be found in Jenks (1977) or Andrienko et al. (2001). In the centre, the districts have been divided into groups with approximately equal total populations: the first class, with proportions of elderly people below 10.47%, contains 34.6% of the whole population of the country; the second class, with proportions from 10.47 to 15.88%, contains 32.9% of the population; and the third class, with proportions 15.88% and over, contains the remaining 32.5% of the population. On the right, the districts have been classified so that the total areas occupied by the classes are approximately equal (more precisely, 33.0, 33.2, and 33.8% of the area of the whole country). The corresponding breaks are at 16.5 and 20.4%.



**Fig. 4.18.** Different variants of the classification of the districts of Portugal according to the values of the attribute “% 65 or more years”. Left: statistically optimal classification with class breaks at 15.24 and 22.69%. Centre: division into classes with approximately equal populations (34.6, 32.9, and 32.5% of the total population of Portugal); the values of the class breaks are 10.47 and 15.88%. Right: division into classes with approximately equal total areas (33.0, 33.2, and 33.8% of the whole area of the country); the class breaks are at 16.5 and 20.4%

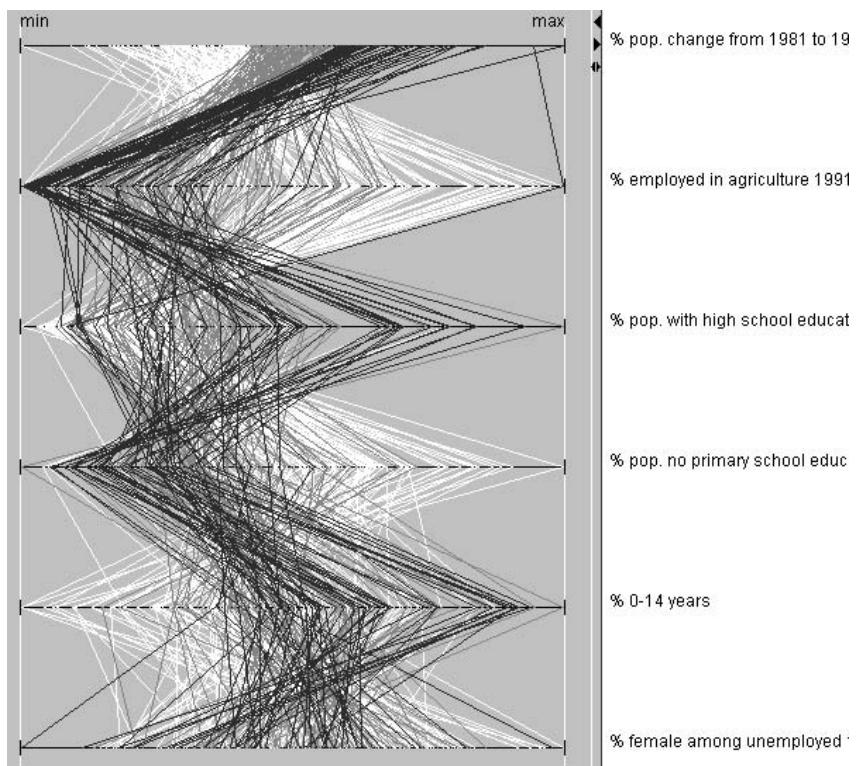
Each map suggests a rather salient but different pattern. Thus, the map on the left exposes clusters of low and high values as spots on a back-

ground of medium values. It may be noticed, by the way, that the two districts in the north-east which were marked in Fig. 4.16 are no longer visually separated from their neighbours by different colouring. Hence, the optimal-classification algorithm has been quite good in handling close values. The map in the centre stresses that the smallest proportions of elderly people occur in small but densely populated areas around the biggest cities but not in those cities themselves (after applying appropriate zooming, it may be seen that the municipalities of Porto and Lisbon are shaded differently from the surrounding districts). It also shows that two-thirds of the population of the country live in districts with less than 15.88% of elderly people (these are the districts of the first two classes), and that these districts are mostly situated along the western coast from the north of the country to the centre. In the map on the right, the areas with low, medium, and high proportions of elderly people appear almost as strips stretching in a north-south direction.

Among these patterns, there are no “right” or “wrong” ones. Each pattern is quite meaningful, and each map contributes to the understanding of the distribution of elderly people over the territory of Portugal. In general, there is no universal recipe for how to obtain an “ideal” classification with understandable class breaks, on the one hand, and interpretable coherent regions, on the other hand. Therefore, when we say that classification may be used as an instrument of data analysis, we mean not a classified map by itself but an interactive tool that allows the analyst to change the classes and to observe immediately the effect on the map. Implementations of such interactive tools are described, for example, in Egbert and Slocum (1992) and Andrienko and Andrienko (1999).

As we have mentioned, the applicability of classification is not limited to maps. Thus, it is quite possible, for example, to classify lines on a parallel-coordinates display such as the one shown in Fig. 4.11. Figure 4.19 demonstrates the result of classifying the lines in this display according to the values of the attribute “% pop. change from 1981 to 1991” (i.e. the relative change of the population, expressed as a percentage of the population in 1981). The value range of the attribute is from -31.3% to 31.11%. We have introduced two class breaks, at -5% and 5%, to divide the districts into those with a population decrease (the change ranging from -31.3% to -5%), a relatively stable population (from -5% to 5%), and a population increase (where the change is 5% or more). As on a classified choropleth map, different shades have been assigned to the classes: the lightest shade (white) to the class with a population decrease, a medium shade to the class with stable population, and the darkest shade to the class

with a population increase. These shades have been used for the lines in the parallel-coordinates display.

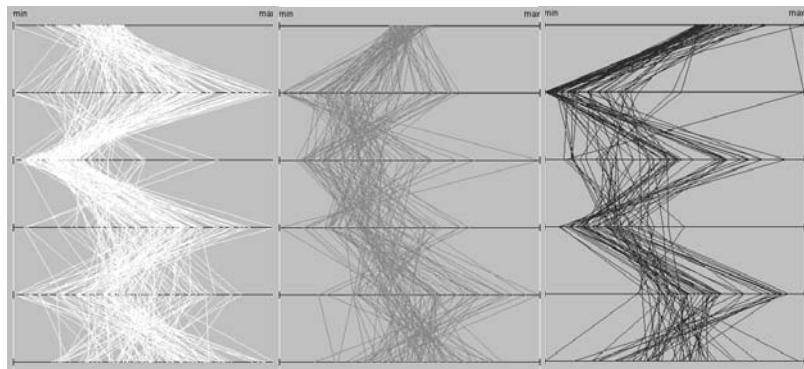


**Fig. 4.19.** Classification of lines in a parallel-coordinates display, which shows characteristics of the districts of Portugal. The lines, which represent the districts, are coloured according to the values of the attribute “% pop. change from 1981 to 1991” in the respective districts. The value range of the attribute (from −31.3 to 31.11) has been divided into three classes by class breaks at −5 and 5

The purpose of applying classification to the map was to facilitate the revealing of distinctive features of the spatial distribution of the attribute values and the finding of an appropriate pattern encompassing those features. The purpose of applying classification to the parallel-coordinates display was to facilitate the revealing of significant correlations between attributes and the finding of an appropriate pattern (for the behaviour of the attributes) encompassing those correlations. Thus, if we compare the shaded parallel-coordinates display in Fig. 4.19 with the original display in Fig. 4.11, we shall notice that the shading indeed helps us to see better how the attributes are related. In particular, we see that the population increased

in districts with low employment in agriculture, a low proportion of people without education, and medium to high proportions of children. The population decreased in districts with a low proportion of people who had high school education and a relatively high proportion of non-educated people; many of these districts have high percentages of people working in agriculture. Moreover, in almost all districts with high employment in agriculture the population decreased, except for a single district with a rather large increase of the population (such atypical cases usually require special consideration and, very often, the involvement of additional information in order to be properly explained).

The explorer can observe the behaviour of each class more conveniently if he/she has an opportunity to view only the lines of one or two selected classes while the remaining classes are hidden. Thus, Fig. 4.20 demonstrates three screenshots of the parallel-coordinates display. In each screenshot, one of the classes is visible and two others are hidden. It might be even more convenient for analysis if the explorer could easily transform the original display into several displays showing each class separately. Juxtaposition of these displays, as is shown in Fig. 4.20, facilitates comparison of the classes.



**Fig. 4.20.** The explorer may have the possibility to look at the lines for each class separately, and to compare the classes represented in multiple juxtaposed displays

With this example of a parallel-coordinates display, we have demonstrated that classification may be possible and useful not only in maps. In fact, we believe that classification has nearly universal applicability. Of course, different colours or shades may be used for distinguishing classes only in visualisations that have not yet used colours for representing other information. However, colouring is not the only possible means for differentiating classes; other visual variables may be used as well.

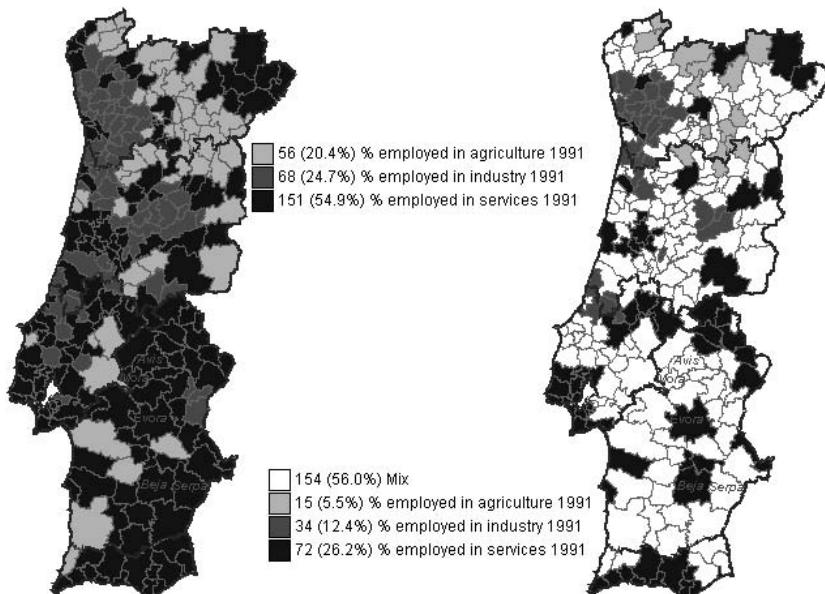
Up to now, we have only considered classification (of elements of a reference set) according to the values of a single numeric attribute. However, the notion of classification is much more general. It is possible to classify references on the basis of attributes of any type, taken separately or in combination. Let us consider a few examples of different classifications.

In Fig. 4.21, we can see two maps using a classification of the districts of Portugal according to the dominant sector of employment in 1991: agriculture, industry, or services. On the left, the districts are coloured depending on which of the three attributes “% employed in agriculture 1991”, “% employed in industry 1991”, or “% employed in services 1991” has the highest value. The map allows us to detect a big industrial area in the north-west and another, which is located close to it, slightly to the south-east. We also see a large area in the central east and south-east of the country with a prevailing employment in services. Agricultural employment dominates mostly in the central north.

However, our observations cannot be deemed completely valid, because of the approach taken to defining the dominant occupation. Suppose, for example, that the distribution of the working population between the three sectors in some district was 33.3, 33.3, and 33.4%. In this case, the sector where 33.4% population worked would be chosen as the dominant one. The district would be coloured on the map in the colour corresponding to this sector and look identical to a district in which 80% of people work in that sector. This feature is hardly desirable and useful for analysis.

To get rid of this undesirable feature, the analyst may wish to change the definition of dominance. One possible way is to introduce a threshold so that an attribute may be regarded as dominant only when its value is above this threshold. The analyst may specify as the threshold some number within the value range of the attribute or a percentage of the sum of the values of all attributes. In our case, the two methods are equivalent: the values of all three attributes are percentages and make 100% in total. For example, the map on the right of Fig. 4.21 is the result of applying a dominance threshold of 50% to the original map shown on the left. Districts in which none of the attributes reaches above the threshold are classified as “mix”. These districts are shown in white.

The map on the right suggests quite a different pattern of the spatial distribution than does the map on the left. Instead of large areas with a prevailing employment in services, we now see only relatively small clusters in the central west (around Lisbon), along the southern coast (in the Algarve province), and in the central inland area. The “industrial” and “agricultural” areas have also shrunk considerably, and the bulk of the territory of Portugal has been classified as “mix”. This means that less than 50% of the population of these districts works in any of the three sectors.



**Fig. 4.21.** The districts of Portugal are classified here according to the dominant sector of employment: agriculture, industry, or services. On the left, the districts are coloured depending on which of the three attributes “% employed in agriculture 1991”, “% employed in industry 1991”, or “% employed in services 1991” has the highest value. On the right, a constraint has been imposed: an attribute is regarded as dominant if its value constitutes 50% or more of the sum of the values of all attributes. Districts in which none of the attributes satisfies the constraint are classified as “mix”. These districts are coloured in white

It is clear that the introduction of a dominance threshold has allowed us to obtain more accurate information from the map than was possible before. However, there is a problem with the specification of the dominance threshold as a proportion of the sum of the values of all attributes. If, for example, the distribution of the working population between the three sectors was 0, 49.9, and 50.1% (which is quite possible, for example, in cities, where nobody works in agriculture), the third sector would be chosen as dominant even though it had only 0.2% more working people than the second sector. In order to be sure that the sector nominated as dominant indeed involves substantially more people than the others, the explorer may wish to define the dominance threshold as a minimum absolute or relative difference between the attribute with the maximum value and the attribute with the next highest value.

Figure 4.22 allows us to compare the results of two approaches to defining the dominance threshold. The map on the left has been constructed,

like that in Fig. 4.21, with a threshold of 50% of the sum of the values. The map in the centre results from applying a 20% threshold to the difference between the maximum value and the next highest value. Although there is a clear similarity between the maps, many of the shaded districts in the first map (i.e. those classified as having a prevalent sector) are white in the second map (i.e. they are evaluated as having no substantial prevalence of any of the sectors). At the same time, a few white districts in the first map have become coloured in the second map. For instance, the cluster of districts in Algarve with dominant employment in services has expanded slightly on the second map in comparison with the first map.

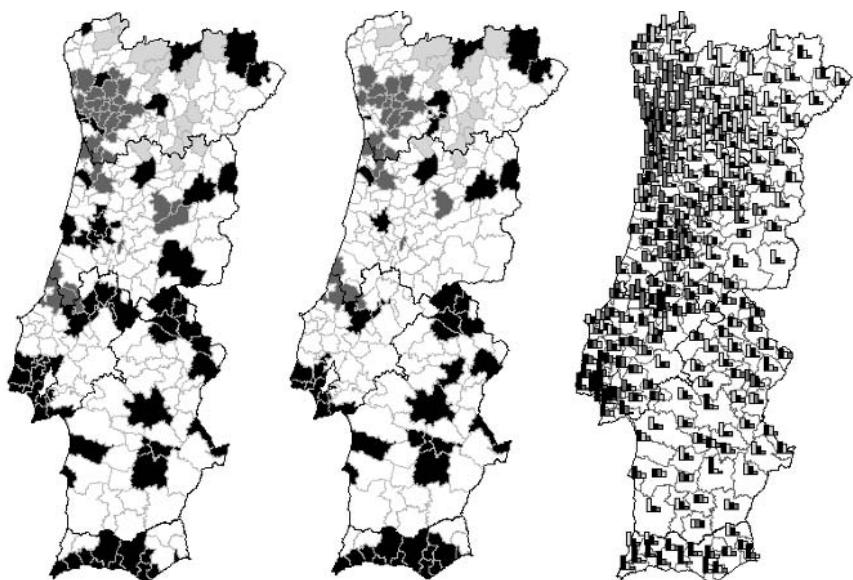
The two approaches to determining the dominant attribute can easily be combined. The analyst may specify two thresholds simultaneously: a minimum percentage of a sum and the minimum distance to the next value.

To understand the value of classification according to the dominant attribute, let us compare any of the classification maps with the map on the right of Fig. 4.22, on which the same data are represented by bar charts positioned inside the contours of the districts. For each sector, there is a bar with a height proportional to the percentage of the population working in this sector. The shades of the bars are the same as the shades used in the classification maps. The bars are arranged in order of decreasing height; hence, the first bar from the left corresponds to the sector of dominant employment in the respective district.

Unlike the classification maps, the map with the bar charts does not involve any information loss: it shows all available values of all three attributes. The map is very informative concerning the employment structure in each particular district. One can not only see which sector is dominant but also estimate the proportion of people working in it and compare this with the proportions for the other two sectors. However, the map with the bar charts does not promote the revealing of the general characteristic features of the spatial variation of the employment structure over the territory of Portugal. In Bertin's terms, this map does not support the overall reading level. In our terms, it does not suggest suitable patterns for the spatial behaviour of a combination of three attributes. The classification maps, in contrast, involve substantial information loss but are quite good for perceiving the whole spatial distribution "in the minimum instant of vision" and thereby defining the general patterns of the behaviour. This is their primary value for exploratory data analysis.

What method of defining dominance and what threshold value are appropriate in a particular case of data analysis depend on the data and the goals of the analyst. In most cases, it is advisable to try different variants. It is important to stress that the analyst needs a flexible, interactive tool for "playing" with various classification methods and their parameters, rather

than just a single static classification. We made the same note earlier, when discussing classification according to the values of a single numeric attribute. In fact, this statement applies to any classification.



**Fig. 4.22.** Left and centre: classification of districts according to the dominant sector of employment using two different ways of specifying the dominance threshold: 50% of the total sum, and 20% relative difference from the next highest value, respectively. Right: the same data, represented by bar charts

As the next example of classification, let us consider the method of cross-classification according to the values of two numeric attributes. This method is rather popular in thematic cartography. The idea of the method is to divide the value ranges of each of two attributes into subintervals, analogously to classification on the basis of a single numeric attribute. If the value range of the first attribute has been divided into  $M$  subintervals and the value range of the second attribute into  $N$  subintervals, this potentially defines  $M \times N$  different classes. Each class is a combination of two intervals, one for the first attribute and one for the second attribute. As in other cases of classification, the classes are represented visually, for example by colouring. It is reasonable to define such a class-colouring scheme so that the analyst can easily identify the meaning of each colour.

Figure 4.23C (“C” indicates a colour figure; all colour figures are placed at the end of the book) demonstrates an example map showing a classification of the districts of Portugal according to the values of the attributes “% 25–64 years” (percentage of population aged from 25 to 64 years) and “%

15–24 years” (percentage of population aged from 15 to 24 years). The value range of each attribute (from 40.25 to 55.99 and from 8.82 to 21.32, respectively) has been divided into three subintervals using the optimal-classification algorithm mentioned earlier (see Fig. 4.18 and the accompanying text). The resulting class breaks are at 46.78 and 50.56 for the attribute “% 25–64 years” and at 13.76 and 16.99 for the attribute “% 15–24 years”. This division defines nine different classes. A colouring scheme based on combining colour hue and colour brightness has been created to represent these classes. Different hues are used for the attribute “% 15–24 years”: blue for the interval of low values, green for the interval of medium values, and red for the interval of high values. Different degrees of darkness correspond to the division according to the attribute “% 25–64 years”: the lightest shades represent low values, medium shades correspond to medium values, and the darkest shades correspond to high values. The division into classes and the assignment of colours of the classes is illustrated graphically in the top left corner of Fig. 4.23C.

Like other maps using classification, cross-classification maps are used as tools to support an overall view of the spatial distribution of attribute values. It can be noted that the cross-classification map in Fig. 4.23C is more difficult to interpret than the previous examples of classification maps. This complexity results from the composite meaning of each colour. However, the analyst does not need to think about the meanings of the colours right from the beginning. It is reasonable first to look at the map and try to perceive the general pattern of the distribution of the colours.

The abstract drawing to the right of the map in Fig. 4.23C represents our perception of the principal structural features of the image conveyed by the map. In the north, we see something like a group of concentric rings with their centre somewhere near Porto. On the left (central west), around Lisbon, there is a dark green figure, which is connected to the group of rings by a green band stretching along the coast. In the southern half of the territory, we see two belts stretching in the direction of the meridians. The dark blue belt on the left is more conspicuous, while the one beside it in lighter blue is spotted with intrusions of green and is therefore not so readily seen.

After grasping the main structural features of the distribution, we can start on translating the figures that we have detected into meaningful descriptions in terms of proportions of young and adult people. The concentric rings radiating from Porto are characterised by high proportions of young people, which decrease at a larger distance from Porto. However, in Porto itself, as well as one of its neighbours, the proportion of young people is smaller than around it. The proportion of adults, which is very high in the innermost circle, decreases rapidly with increasing distance from

Porto and then increases slightly at the eastern edge of the country's territory. The coastal belt between the north and the centre is mostly characterised by medium to high proportions of adults and medium proportions of young people, except for an intrusion of two adjoining districts with high proportions of young people in the north of the belt. In the area around Lisbon, the proportions of adults are high and the proportions of young people are medium. The belt beginning east of this area and stretching to the southern coast has high proportions of adults and low proportions of young people. In the eastern direction from this belt, the proportion of adults decreases. The proportion of young people varies in this area between low and medium.

As in the case of the employment structure, we could represent the same data by bar charts and thereby avoid information loss. However, the bar chart representation would not give us such a general view of the principal features of the distribution as is possible with classification.

As an example of a classification according to a temporal attribute, let us consider Fig. 4.24, which represents forest fires that occurred in an area during a 7-year period. By inspection of the distribution of the dates of the fires over the whole period, we have found that there were 14 time intervals of frequent occurrences of fires, with gaps between them when no fires occurred. Each year contains two such periods of fires. The first period begins approximately in February and lasts mostly until the end of April; in some years it lasts until the end of May or even the beginning of June. The second period covers July, August, and September, and in some years the beginning of October. There are also a few fires that occurred in winter, in December or January.

On the basis of this finding, we have divided the whole set of fires into three classes: spring fires (153 fire occurrences), summer fires (523 occurrences), and winter fires (7 occurrences). The result of the classification is presented in the maps in Fig. 4.24. The map at the top left shows all the fire occurrences, which are represented by small circles coloured according to the classes that the fires belong into. Despite the overlap of the symbols, we can see even on this map that the spatial distribution of spring fires differs from that of summer fires. In particular, spring fires rarely occur within the east and south-east of the territory, while summer fires are nearly evenly spread over the whole territory. To see the differences better, we can look at the distribution of each class separately from the others. The map at the top right shows us the winter fires (which are very few), the map at the bottom left shows the spring fires, and the map at the bottom right shows the summer fires. The difference between the lower two maps is rather salient.



**Fig. 4.24.** Forest fires that occurred over a 7-year period, classified according to the season of occurrence: winter (December to January), spring (end of February to May), and summer (July to September). The fires are represented on the maps by small circles coloured according to the class. Top left, all fires; top right, winter fires; bottom left, spring fires; bottom right, summer fires

It is quite possible that the distinction between the spatial distributions of spring and summer fires may be well known to any inhabitant of the region that the data refer to (or at least to any specialist in forest fires), but we have arrived at this finding without any prior knowledge, with the help

of classification. However, in order to find the principle to be used for the classification, we first analysed the properties of the temporal behaviour of the phenomenon. This example shows that classification is not necessarily as straightforward as breaking the value range of a numeric attribute into intervals or determining the maximum of the values of several attributes. Classification may be quite sophisticated and involve supplementary information from a previous analysis and/or from domain knowledge. Another example of this kind could be a classification of tree species, such as pine, spruce, oak, and birch, into coniferous and deciduous, or into hardwood and softwood, depending on whether the classification is made by a botanist or a forest manager.

We believe that we have provided enough examples of classification to demonstrate its important role in exploratory data analysis. Let us now look at the remaining groups of methods for display manipulation and then consider other categories of tools for EDA.

#### 4.4.4 Zooming and Focusing

A common problem of all data displays is the relatively small space available, whereas the volumes of data that need to be analysed are usually rather large or even huge. As a result, it is often impossible to represent all data at the same time with sufficient legibility, precision, and level of detail. To fit much data into a small space, one needs to apply a high degree of generalisation, use very small marks, and/or permit symbol overlap. All of this greatly reduces the amount of information that can be perceived from the display or even makes it unintelligible.

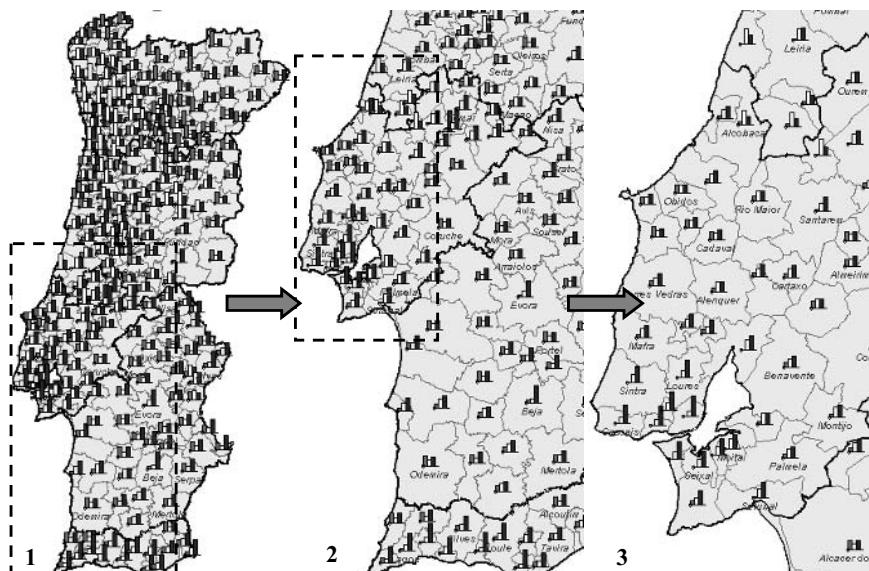
Displays on a computer screen are typically supplied with tools for zooming: the user chooses a fragment of a display, and it is enlarged to the maximum size permitted by the available space. The corresponding portion of data can now be represented in a better way: with higher precision, larger symbols, and less overlap. An evident shortcoming is that this improvement applies only to part of the data. The result of zooming can be seen metaphorically as the user watching a large picture through a small window, which does not permit a full view. Further zooming is similar to replacing the current window glass with a lens with higher magnifying power: as a result, the user can see a smaller fragment of the picture better.

Besides zooming, most visualisation tools support the operation of panning, which can be viewed as shifting the position of the window so that another fragment of the picture becomes visible.

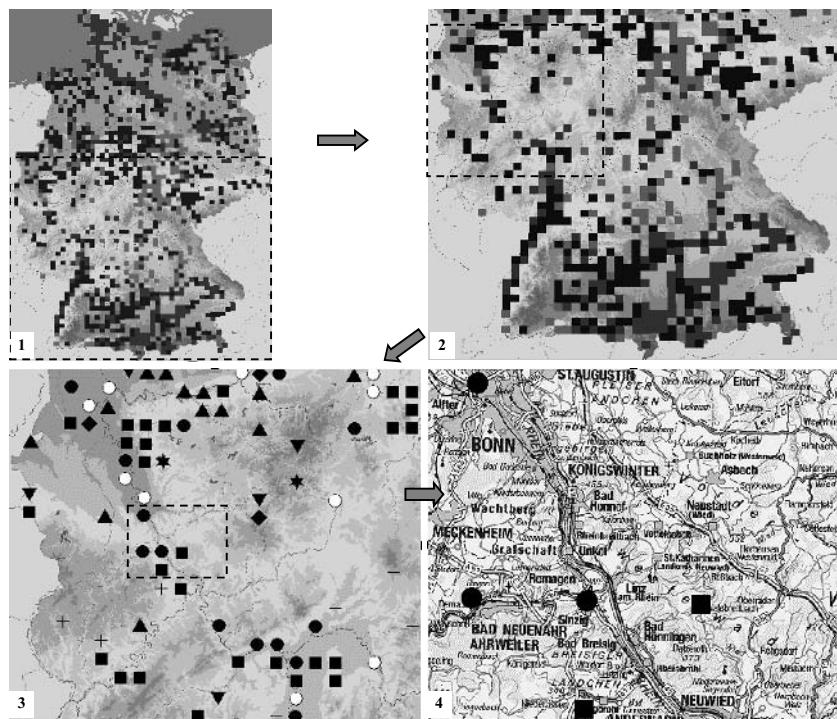
Zooming tools can be divided into two broad groups. One group consists of tools that, after selection of a display fragment, show only this

fragment in a larger size and discard the other parts of the display. The other group includes tools which show the selected fragment “in context”, i.e. with its surroundings. The supposed benefit of such an approach is that it supports the user’s orientation and navigation better. In order to make the space for increasing size of the selected fragment, the surroundings need to be substantially reduced in size. Hence, context-preserving techniques for zooming necessarily involve a distortion of the original display. One of the best known distortion-based zooming techniques is the “Fish-eye view” (Furnas 1986). A survey and taxonomy of distortion-based techniques can be found in Leung and Apperley (1994).

Zooming, in particular, map zooming, can be done in two general ways. One is simply to increase the sizes of the marks and symbols and/or the distances between them, without changing the amount and precision of the information represented in the fragment being zoomed in. The other approach is to change the information content by decreasing the degree of generalisation and including more detail; for example, by adding small objects that could not be made visible at the previous scale. Sometimes, this operation is accompanied by changing the symbolisation: a circle that has originally represented a city may be replaced by a group of area marks portraying major parts of the city or even blocks of houses.



**Fig. 4.25.** Simple map zooming: enlargement of the sizes of geographical objects and the distances between them is not accompanied by increasing the level of detail



**Fig. 4.26.** Zooming of a map with addition of more detail and replacement of the representation method. Source: [http://www.floraweb.de/datenservice/cg\\_floramap/cg\\_floramap.html](http://www.floraweb.de/datenservice/cg_floramap/cg_floramap.html)

Figures 4.25 and 4.26 illustrate the difference between the two methods of zooming. In Fig. 4.25, zooming in on a part of the territory of Portugal results almost entirely in enlarging the figures representing the districts of Portugal and increasing the distances between the bar chart symbols. Only labels with district names are added, when the contours of the districts become large enough to allow them to be placed.

In contrast, Fig. 4.26 demonstrates zooming accompanied by increasing the level of detail and changing the representation method. This illustration was produced using a Web applet that visualises the distribution of rare and endangered plant species in Germany. This applet, together with a Web interface for accessing a database containing data about the plants, is available at the URL [http://www.floraweb.de/datenservice/cg\\_floramap/cg\\_floramap.html](http://www.floraweb.de/datenservice/cg_floramap/cg_floramap.html).

The applet works as follows. The user sends a query to the database to retrieve data about the distribution of some plant species. In response, the database sends the requested data, and the applet shows these data on a

map of Germany in a generalised form (see map 1 at the top left of Fig. 4.26). Specifically, the locations where plants of the designated species have ever been observed are classified into three classes, according to the time when the plants were last seen in those locations: before 1950, between 1950 and 1980, and after 1980. The locations are marked by squares coloured according to the class.

In the series of zoom operations shown in Fig. 4.26, the first operation simply increases the size of the selected part of the territory and the squares corresponding to the locations at which the plants were found. However, the next operation results in replacing the classification by symbolic signs that convey much detailed information about the findings. Let us explain some of the symbols:

- indigenous occurrence, last observed before 1950;
- indigenous occurrence, last observed between 1950 and 1980;
- indigenous occurrence, last observed after 1980;
- ◆ doubtful indigenousness, last observed after 1980;
- ▼ fickle;
- ▲ naturalised, last observed after 1980;
- ★ domesticated, last observed after 1980;
- + extinct;
- correction: absent.

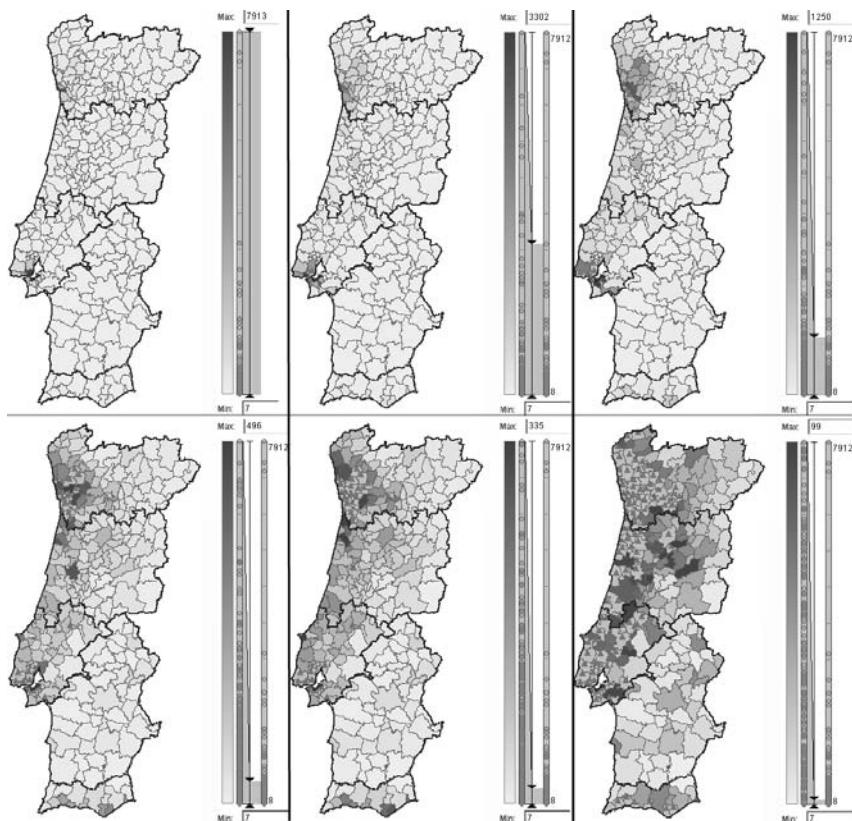
The next zooming operation results in replacing the highly generalised background image of the relief of Germany by a topographic map of the selected area; the representation of the data concerning the plant species remains the same as at the previous zoom level.

An operation similar to zooming is focusing, which involves selection of a data subset rather than a display fragment and results in this subset being portrayed with the maximum possible expressiveness. Owing to the conceptual similarity to display zooming, focusing may also be called “data zooming”. Let us give some examples.

Figure 4.27 presents a series of screenshots of one and the same map display portraying the population densities in the districts of Portugal by means of a gradual scale of decreasing brightness or, in other words, increasing darkness, i.e. darker shades correspond to higher values. In the map at the top left, the darkness scale is matched to the full range of attribute values, specifically, from 7 to 7913.

The population in Portugal is distributed very unevenly. There are a few small districts with very high population densities, among them the largest cities Lisbon and Porto, with population densities of 7913 and 7261 inhabitants per square kilometre, respectively, and a satellite of Lisbon called Amadora, with a density of 7455 inhabitants per square kilometre.

The distance from these values to the next highest value, 3302, is very large. As a result, the few very high values (outliers) and the bulk of the dataset fit into the opposite ends of the darkness scale. All but a few districts in the map are shown in very light shades and look almost identical, while a large part (about half) of the darkness scale remains unused.



**Fig. 4.27.** Focusing applied to a map of population densities in the districts of Portugal. The densities are represented using a gradual scale of increasing darkness, i.e. darker shades correspond to higher values. The map at the top left portrays the full range of attribute values, from 7 to 7913. A few very high values (outliers) are represented by dark shades. Owing to the great distance from these values to the rest of the data, the bulk of the dataset fits into the light end of the darkness scale. As a result, all districts look similar; no differences can be seen. Focusing allows the full extent of the darkness scale to be matched with shorter value intervals, so that the values in these intervals may be represented with the maximum possible expressiveness. The screenshots, from left to right, correspond to the following maximum values matched to the dark end of the darkness scale: 7913, 3302, and 1250 in the upper row, and 496, 335, and 99 in the lower row

The distribution of the population density values present in the dataset within the range from 7 to 7913 is shown in the dot plot display to the right of the map. The values are represented by circles positioned along the vertical dimension, with the lower end of the display corresponding to the minimum value, i.e. 7, and the upper end to the maximum, 7913. It may be seen that the density of the circles is very high at the lower end and decreases in the upward direction. At the top, there are three circles corresponding to the three highest values. These circles are separated from the rest by a wide gap, which “eats” more than half of the length of the darkness scale. The remaining part of the scale provides a rather small number of perceptually different shades. It therefore a natural idea to try to remove the outliers from consideration and “stretch” the reduced value range so as to match the full extent of the darkness scale. As a result, it will be possible to use more different shades for portraying the values in the selected subinterval. Hence, these values may be represented with greater expressiveness, so that differences between them can be better perceived.

The sequence of screenshots in Fig. 4.27 demonstrates the effect of a recurrent application of the operation of selecting a value subinterval to be matched to the full extent of the darkness scale. In the second screenshot (i.e. in top centre), the three highest values have been removed from the representation, and the dark end of the darkness scale corresponds to the a value of 3302. The districts corresponding to the removed values are not “coloured” (shaded) on the map any more. The software tool used for producing the screenshots replaces colouring by triangular symbols, which indicate that the corresponding values lie beyond the focusing range. The map looks less uniform than the original display: we can see areas around Porto and Lisbon shaded slightly darker than the rest of the territory. Nevertheless, the map is not sufficiently expressive.

The next screenshot (top right) corresponds to an even shorter subinterval, specifically, from 7 to 1250. The clusters around Lisbon and Porto have become more apparent. Between these clusters, one can detect a band of relatively high population density (identified owing to a slightly darker shading than in the remaining territory) stretching along the coast. In the next screenshots (in the lower row), one can observe a spatial trend of decreasing population density from the coast towards the inland areas, as well as a large area of low population density in the southern part of the country, except for the southern coast. In the last screenshot (bottom right), with the interval from 7 to 99 represented by shades, the differences between individual values within this interval are very salient.

In this example, we dealt with a single numeric attribute. Focusing was done by selecting subintervals of the whole value range of the attribute. How do we do focusing if we have two or more numeric attributes?

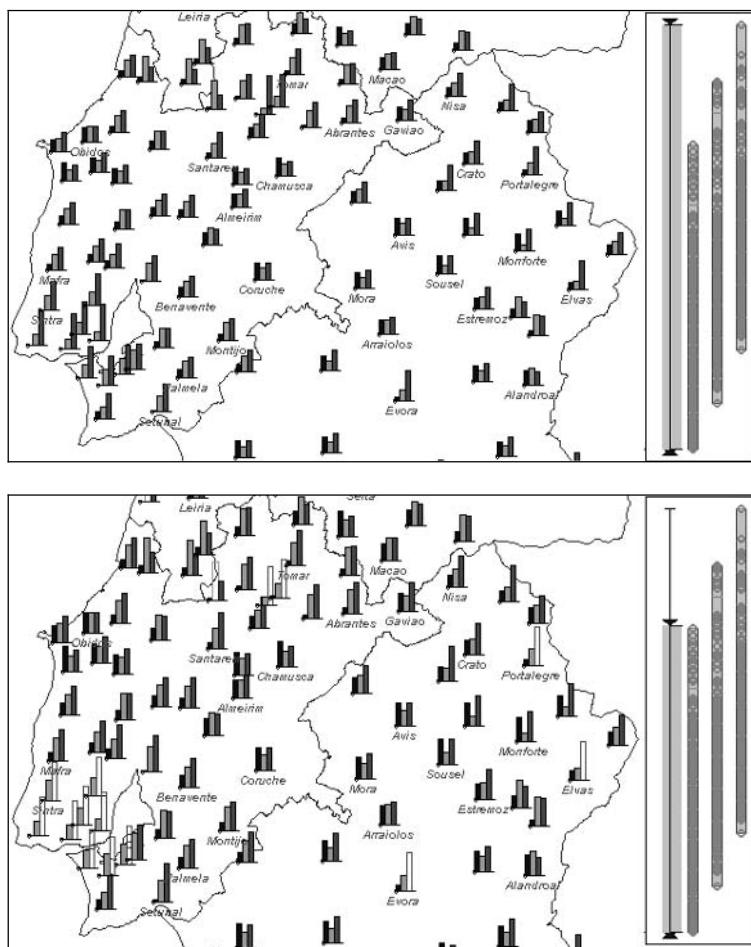
Figures 4.28 and 4.29 demonstrate two possible approaches. In Fig. 4.28, the percentages of people employed in agriculture, industry, and services in the districts of Portugal are represented by bar charts with bar heights proportional to the values of the respective attributes. For better visibility of the bar charts, we have zoomed in on the central part of the country and switched off the drawing of the district boundaries.

The three attributes, i.e. “% employed in agriculture 1991”, “% employed in industry 1991”, and “% employed in services 1991”, have different value ranges. The maximum value of the first attribute is 61.41, the maximum value of the second is 74.20, and the maximum value of the third is 85.57. The bar charts have been constructed so that a unit of bar length has the same meaning (i.e. the same corresponding value) in all three bars representing different attributes. This has been done by matching the maximum bar size chosen for the charts to the maximum of the values of all three attributes, i.e. the value 85.57. Any other value  $x$  of any attribute is represented by a bar height computed according to the formula  $x$  divided by 85.57, multiplied by the maximum bar height. Since 85.57 is the highest value in the dataset, the resulting bar height will definitely not be more than the maximum bar height.

The bar chart visualisation allows one to do focusing by choosing a smaller value to match the maximum bar height. Then, the height of the bar that represents the value  $x$  is determined by the formula  $x$  divided by  $N$ , multiplied by the maximum bar height, where  $N$  is the chosen value. Since  $N$  is smaller than the maximum value in the dataset, it may happen that the computed bar height  $h$  for the value  $x$  is greater than the chosen maximum bar height  $H$ . In this case, instead of a bar with a height  $h$ , a rectangular frame with a height  $H$  appears at the corresponding position in the chart. Such frames can be seen in the lower map fragment in Fig. 4.28.

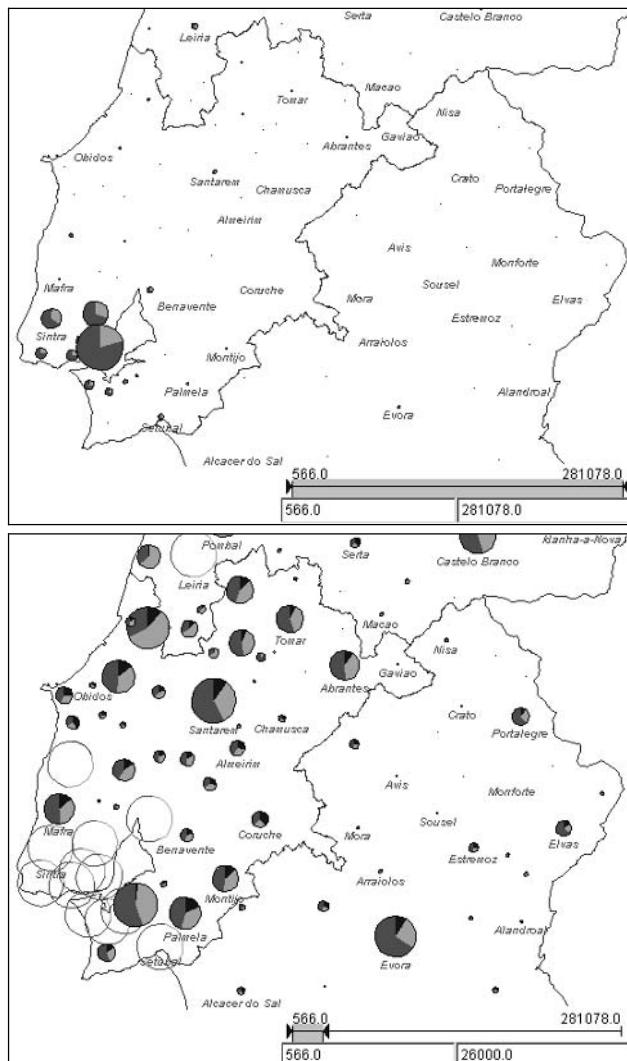
However, the appearance of the frames is not the main effect of the focusing operation. The main result (and the reason for applying focusing) is that the values in the interval from 0 to  $N$  can now be represented more expressively than before: the value corresponding to a unit of bar height has become smaller and, hence, finer differences may be perceived.

Analogously to the previous example, the dot plots to the right of the maps represent the distribution of the values of the three attributes within their value ranges. The vertical line shows the combined value range of all three attributes. The triangular marks indicate which portion of this range is currently matched to the maximum bar height. In the upper picture, this is the entire combined value range, i.e. from 0 to 85.57; in the lower picture, this is about three-quarters of the entire range; more precisely, the interval from 0 to 62.



**Fig. 4.28.** Focusing by reducing the attribute value to be matched to the chosen maximum bar height. Top: the maximum bar height corresponds to the maximum value 85.57 of the values of the three attributes “% employed in agriculture 1991”, “% employed in industry 1991”, and “% employed in services 1991”. Bottom: the maximum bar height corresponds to the value 62.0. Values greater than this are represented by frames with a height equal to the maximum bar height. Values below or equal to 62.0 are represented by bars of proportional height

In this example, focusing is done by specifying a constraint common to all attributes concerning the value interval to be represented by the chosen visual means (values that do not fit into this interval are shown in a special way). Another approach is to constrain the range of some integrated characteristic derived from the values of individual attributes, such as the sum of the values. This approach is demonstrated in Fig. 4.29.



**Fig. 4.29.** Focusing by constraining the maximum sum of attribute values. Pie charts represent the total number and the proportions of people working in three sectors of the economy. The size (area) of a pie is proportional to the sum of the number of people working in all three sectors. At the top, the maximum pie size corresponds to a value of 281 078. The pies in the districts with little working population cannot be seen, because of their very small sizes. At the bottom, the maximum sum of values to be represented by the pie size has been limited to 26 000. As a result, many of the formerly invisible pies have become sufficiently large to be seen. In the districts where the working population exceeds 26 000, hollow circles are drawn instead of pies

The map fragments in Fig. 4.29 are also related to the employment structure in the districts of Portugal. However, instead of the percentages of people employed in agriculture, industry, and services, the maps represent the absolute numbers of such people, which constitute the values of the attributes “total employed in agriculture 1991”, “total employed in industry 1991”, and “total employed in services 1991”. To represent these values, the pie chart method is used: for each district, there is a circle (“pie”) divided into three sectors (“pie slices”) proportional to the numbers of people working in agriculture, industry, and services. The size (area) of the circle is proportional to the total working population in the district, i.e. the sum of the values of all three attributes. Analogously to the bar chart representation method discussed above, a certain maximum pie size is chosen. This size is initially matched to the maximum of the sums of the values of the three attributes computed for all districts. For any district, the size of the corresponding pie is defined by the following formula: the sum of the three attribute values divided by the maximum sum multiplied by the maximum pie size.

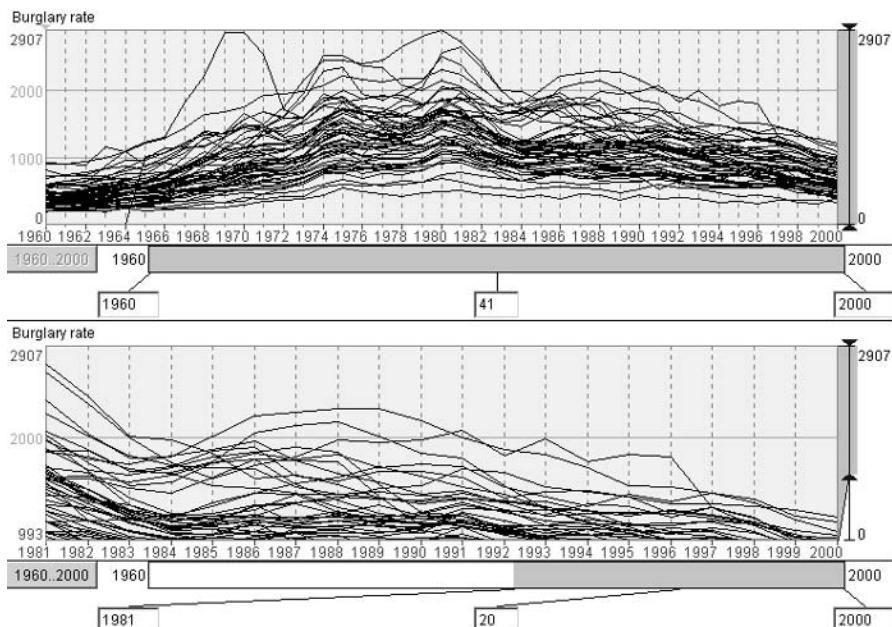
The resulting representation is shown in the map fragment at the top of Fig. 4.29. We have already mentioned that the population of Portugal is distributed very unevenly, i.e. the number of inhabitants in a district varies greatly. This also applies to the working population (i.e. the sum of the number of people employed in the three sectors of the economy), which ranges from 566 to 281 078. The maximum number, which is attained in Lisbon, is much higher than the next highest number 153 319. Only 5 of 275 districts (1.8%) have a working population over 85 000, and only 23 districts (8.4%) have more than 35 000 working people.

As a consequence of such an uneven distribution, only a few pies are actually seen in the upper map fragment in Fig. 4.29. This is because the working population in most districts is so small in comparison with the maximum value, 281 078, that the sizes of the corresponding pies are close to zero, and the pies cannot be seen.

The lower map fragment demonstrates the result of focusing on the interval of value sums from 566 to 26 000. This means that the maximum pie size corresponds to the value 26 000, instead of the 281 078 in the original map. Hence, the value of a unit of pie size has become much less than before, and it is now possible to represent smaller numbers of working people by pies with quite discernible sizes. In the districts where the working population exceeds the limit of 26 000, hollow circles are drawn instead of pies. The proportions of people employed in the various fields can still be determined from the division of the circles into coloured arcs (which is not clearly visible on the greyscale reproduction but can be seen better on a computer screen).

The line below each map represents the entire range of the sums of values, i.e. from 566 to 281 078. The triangular symbols indicate the current focus interval. At the top, this corresponds to the full range; at the bottom, it corresponds to the interval from 566 to 26 000.

On a map, the difference between the operations of zooming and focusing is evident. Thus, both map fragments in Fig. 4.29 have been produced by means of zooming, but the lower fragment is also the result of focusing applied to the upper fragment. In some visualisations, however, zooming and focusing have exactly the same meaning and effect. This applies to visualisations in which attributes or referrers are represented by spatial dimensions of the display. In such a visualisation, selection of a display fragment to be enlarged is equivalent to choosing a value subset to be shown with the maximum possible expressiveness.



**Fig. 4.30.** Focusing in a time graph display. The horizontal dimension represents the time interval from the year 1960 to the year 2000. The vertical dimension is used to represent the values of the attribute “Burglary rate” in the states of the USA. The upper display shows the whole time interval and the full range of attribute values, specifically, from 0 to 2709. The lower display results from focusing on the last 20 years and on the attribute values from 993 to 2907

Let us take a look at the time graph at the top of Fig. 4.30. The horizontal dimension of this display represents the time interval from the year

1960 to the year 2000. The vertical dimension is used to represent the values of the attribute “Burglary rate” in the states of the USA, which range from 0 to 2709. If, for example, an analyst wishes to consider in more detail the last 20 years of the entire time period and attribute values of around 1000 or higher, he/she may focus on the time interval from 1981 to 2000 and an interval of attribute values from about 1000 to the maximum value, i.e. 2907. The result of such focusing is presented in the lower part of Fig. 4.30. The same result could be achieved by means of zooming in on the part of the display corresponding to the selected intervals.

In general, zooming and focusing are used in order to see more detail or to better differentiate elements of a visualisation. Such needs are typically more pertinent to elementary tasks than to synoptic ones. However, the example concerning the population density in Portugal (see Fig. 4.27) shows that focusing can also be very helpful for grasping the distinctive features of a behaviour as a whole (in our example, the distribution of the population density over the entire territory). This usually happens when the data contain outliers, i.e. extremely high or extremely low values, which differ very much from the rest of the data. In such cases, removing the outliers from the consideration allows an analyst to get a clearer overall view of the bulk of the dataset.

#### 4.4.5 Substitution of the Encoding Function

Any data visualisation is based on encoding elements of data, i.e. values of referrers and attributes, by graphical features, i.e. values of display dimensions and graphical variables. This encoding may be specified by means of a mathematical formula, for example, the following formula for encoding values of numeric attributes by the heights of bars in a bar chart:

$$\text{height} = \mathbf{H} \cdot x : \mathbf{M} \quad (\text{Example 1})$$

where  $x$  is the value to be encoded,  $\mathbf{M}$  is the maximum of the values of the attributes involved in the visualisation, and  $\mathbf{H}$  is the chosen maximum bar height. Another possibility is to define the encoding of data by graphical features through a set of rules; for example, in the case of encoding attribute values by colours,

$$\begin{aligned} &\text{if } x = \text{“spruce”} \text{ then } \text{colour} = \text{“dark green”}; \\ &\text{if } x = \text{“pine”} \text{ then } \text{colour} = \text{“light green”}; \\ &\text{if } x = \text{“oak”} \text{ then } \text{colour} = \text{“brown”} \end{aligned} \quad (\text{Example 2})$$

or

---

(Example 3)

$$\begin{aligned}
 & \text{if } x \leq \mathbf{b}_1 \text{ then } \text{colour} = \text{"green"}; \\
 & \text{if } \mathbf{b}_1 < x \leq \mathbf{b}_2 \text{ then } \text{colour} = \text{"yellow"}; \\
 & \text{if } x > \mathbf{b}_2 \text{ then } \text{colour} = \text{"red"}
 \end{aligned}$$

Whatever method is used to specify the visual encoding of data values, we shall generally say that the encoding is defined by a certain *function* (in the mathematical sense), which converts data values into graphical features. We shall call this function the *visual encoding function*. In fact, *any display manipulation is manipulation of its visual encoding function*. Let us consider how such a function can be manipulated.

Any visual encoding function involves variables, independent and dependent, and constants. The independent variables in such a function are variables representing attributes or referrers of the dataset, such as the variable  $x$  in all three examples given above. The dependent are visual variables or dimensions, such as *height* in example 1 and *colour* in examples 2 and 3. The constants are specific values of either data components, such as **M** in Example 1, “spruce” etc. in Example 2, and  $\mathbf{b}_1$  and  $\mathbf{b}_2$  in Example 3, or visual variables, such as **H** in Example 1 and the particular colours in Examples 2 and 3. We shall call specific values of visual variables involved in a visual encoding function “visual constants”, and values of data components used in this function “data constants”.

Visual constants are usually chosen more or less arbitrarily. Thus, one can choose the maximum bar height to be 20, 25, or 30 millimetres in Example 1 or select another set of colours to represent forest types in Example 2. Of course, one should bear in mind some practical limitations and legibility requirements: the maximum bar height should be neither too big nor too small, and the colours used to represent different categories must be visually well distinguishable. Within these limitations, the choice of visual constants is mostly a matter of taste; it does not substantially affect the process and outcomes of data analysis.

The situation with data constants is different. On the one hand, they cannot be chosen as freely as visual constants, but are typically determined by the values present in the dataset. Thus, the value of the constant **M** in Example 1 is determined by the maximum of the attribute values present in the dataset. There is no sense in choosing a higher value for **M** than the maximum attribute value – this would decrease the expressiveness of the visualisation. It is, in principle, possible to choose a smaller value for **M** than the maximum attribute value; however, in this case it is necessary to specify how to handle attribute values greater than **M**, for example by defining an additional encoding function to be used for such values. In Example 2, the values “pine”, “spruce”, “oak”, etc. are elements of the value set of a particular data component. These constants could be replaced by

“coniferous” and “broadleaved” (taking into account the semantics of the data) but not by “cigarettes” and “beer”. In Example 3, the constants  $b_1$  and  $b_2$  must be chosen from the value range of the attribute represented through this visual encoding function.

On the other hand, changing data constants typically affects the visualisation much more seriously than does changing visual constants. In this section, we have demonstrated and are going to demonstrate further that display manipulation is of great use in exploratory data analysis, and almost all examples of display manipulation that we have discussed so far have been based on the alteration of data constants in visual encoding functions. Thus, the bar chart display shown in Fig. 4.28 uses the function presented in Example 1 for encoding values of the attributes by bar sizes. Focusing of the bar chart display is actually changing the constant  $M$  in this function. A similar function is used for encoding numeric values by degrees of darkness:  $darkness = D \cdot x : M$ , where  $D$  is the maximum darkness to be used in the display. Focusing in this case is, again, changing the value of  $M$ , and Fig. 4.27 demonstrates its great value in data exploration. Example 3 corresponds to classification of references, such as districts in Portugal, on the basis of some numeric attribute, such as the percentage of elderly people. The constants  $b_1$  and  $b_2$  are the class breaks. Figure 4.18 shows how changing the class breaks may affect the pattern perceived from the map. One can also modify this visual encoding function by adding new breaks or removing some of the existing breaks. The encoding of qualitative attribute values by colours, as in Example 2, may be modified by means of grouping the values (e.g. “spruce” and “pine” into “coniferous”) and assigning the same colour to all members of a group.

However, not only constants in the definition of a visual encoding function may be changed, but also the function itself may be replaced by another function with the same independent and dependent variables. Thus, the linear function in Example 1 may be replaced by a logarithmic function

$$height = H \cdot \log(x) : \log(M)$$

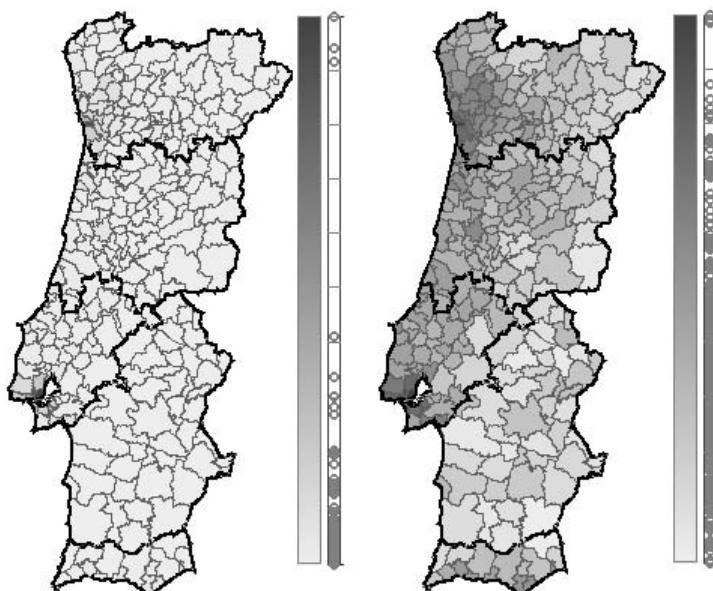
and the same can be done for the darkness:

$$darkness = D \cdot \log(x) : \log(M)$$

Non-linear encoding functions are often used when the statistical distribution of the attribute values is greatly skewed. In particular, logarithmic functions are helpful when a dataset contains a few very high values but the majority of values are quite small. When such data are graphically encoded through a linear function, only a rather small number of perceptually different values of a visual variable are available for representing the bulk

of the attribute values. As a result, the graphical representation of these attribute values is not sufficiently expressive, i.e. different values are visually indistinguishable. We had such a situation in the example concerning the population densities in the districts of Portugal (see Fig. 4.27). When a linear function was used to encode the attribute values by darkness, only focusing on a very small subrange of density values allowed us to differentiate between the shades of the districts in the greater part of the map.

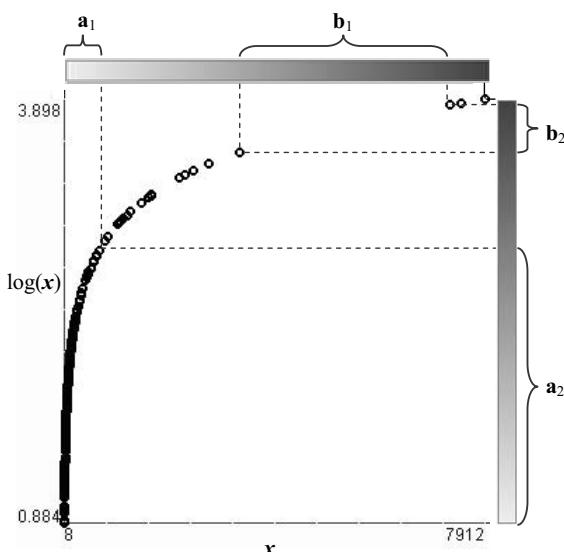
When a logarithmic encoding function is applied instead of a linear one, the map of population densities changes greatly. The difference is demonstrated in Fig. 4.31. The map on the left was built with a linear encoding function, and the map on the right with a logarithmic encoding function. As compared with the original map, the map on the right not only allows us to differentiate between population densities in districts in the east and in the south of the country but also demonstrates a salient pattern in the spatial distribution of the attribute values, specifically a decreasing trend in the direction from coastal to inland areas.



**Fig. 4.31.** A linear function for encoding population densities by degrees of darkness (left) has been substituted by a logarithmic function (right). The resulting map shows much more clearly the distinctive features of the spatial distribution of the population densities

Let us investigate what makes the map on the right more expressive than the one on the left. Beside each map, we have placed a dot plot aligned

with a darkness bar to demonstrate how attribute values are matched in this map to degrees of darkness. The lower ends of the dot plots correspond to the lowest population density value, and the upper ends to the highest value. In both cases, the lowest attribute value corresponds to the lightest shade, and the highest value to the darkest shade. However, the remaining values are distributed between the ends of the dot plots quite differently. On the right, the lower part of the dot plot looks stretched and the upper part compressed in comparison with the plot on the left. Owing to this, a longer interval of the darkness scale is available for representing the small values.



**Fig. 4.32.** The difference between linear and logarithmic encoding is demonstrated graphically here. The circles represent the districts of Portugal; their horizontal positions correspond to the population densities in the districts, and their vertical positions to the decimal logarithms of the population densities

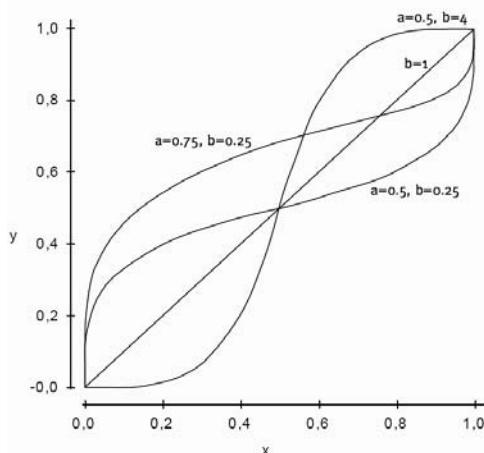
The effect of stretching the interval of low values and shrinking that of higher values is brought about by the properties of the logarithm function, as is visually explained in Fig. 4.32. Here, the horizontal positions of the circles correspond to the population densities in the districts of Portugal, and the vertical positions to the decimal logarithms of these values. It may be clearly seen that the small values of population density fit into a very short interval  $\mathbf{a}_1$  of the darkness scale, while their logarithms are distributed over a much extended interval  $\mathbf{a}_2$ . The gap between the highest three values and the rest is very wide on the linear (horizontal) axis and rather

short on the logarithmic (vertical) axis; these distances are denoted by  $\mathbf{b}_1$  and  $\mathbf{b}_2$ , respectively.

Not only linear or logarithmic encoding functions may be used in data visualisation; any other non-linear function may also be suitable for these purposes if it is monotonic, i.e. either increasing or decreasing but not increasing in one part of its domain and decreasing in another part. For example, Unwin and Hofmann (1998) suggest a parameterised visual encoding function with two parameters  $a$  and  $b$ . The function is defined as follows:

$$f(x) = \begin{cases} a \times \left(\frac{x}{a}\right)^b & \text{for } x \leq a \\ 1 - (1-a) \times \left(\frac{1-x}{1-a}\right)^b & \text{for } x \geq a \end{cases} \quad (4.1)$$

where  $0 < a < 1$  and  $b > 0$



**Fig. 4.33.** The graphical form of the parameterised visual encoding function suggested by Unwin and Hofmann (1998) for various choices of the parameter values. Reprinted from *Exploring Geovisualization*, ed. by Dykes, J., MacEachren, A.M., Kraak, M.-J., p.134, Copyright (2005), with permission from Elsevier

Figure 4.33 shows the graphical form of this function for different choices of the parameters  $a$  and  $b$ . It may be seen that the function becomes linear when  $b = 1$ . When  $b < 1$ , the function extends the margins of the attribute's value range and shrinks the middle part. When  $b > 1$ , the middle part is stretched and the margins are compressed. The value of the parameter  $a$  defines which of the margins will be stretched more. Increases-

ing  $a$  when  $b < 1$  enhances the extension at the beginning of the value range and decreases the extension at the end of it. The opposite takes place when  $b > 1$ .

As in the case of linear encoding functions, data displays based on non-linear functions can be manipulated further by changing the parameters used in the definitions of those functions. For example, an analyst can interactively change the values of the parameters  $a$  and  $b$  in Unwin and Hofmann's function until a satisfactory display expressiveness is achieved.

Nothing prevents any other monotonic non-linear function being used for the visual encoding of attribute values. However, the more complex the function is, the more difficult it becomes to interpret the resulting visualisation. In fact, whatever non-linear encoding function is used, the analyst must refrain from estimating values and differences between values from the appearance of the corresponding graphical features. He/she should treat the perceived differences in shades, sizes, positions, etc. just as indications of the existence of real differences. The explorer can find out which of two values is bigger but should be very cautious in judging how much bigger it is. Thus, the map on the right of Fig. 4.31 allows us to see that the population densities are higher at the coast than inland, but it is impossible, without using additional tools, to determine the real amount of difference.

In order to avoid misinterpreting data, it is recommended that one should not use non-linear encodings alone, but combine them with visualisations based on linear encoding functions. By comparing such complementary views, one can attain a truthful grasp of the data.

#### 4.4.6 Visual Comparison

The role of the display manipulation techniques discussed in this subsection is not in fact limited to facilitating comparison operations. We use the term “visual comparison” simply because we have not found a better way to express in a short but understandable phrase the main idea of these techniques, that is, to emphasise deviations of attribute values from a particular value, which can be specified interactively.

Let us explain this idea with an example. In Fig. 4.34C, we can see an unclassified choropleth map. Unlike the other unclassified choropleth maps that we considered before (see Fig. 4.16, left, and Fig. 4.27), this map uses shades of two colour hues to represent values of a single numeric attribute, specifically, the attribute “% female 1991” (the percentage of females in the population in 1991). Hence, two visual variables are involved in this visualisation, hue and brightness (i.e. darkness), while the previous choropleth maps employed only darkness.

When hue is used in an unclassified choropleth map, two different instances of this variable are used to indicate whether an attribute value is smaller or greater than some selected reference value. In Fig. 4.34C, this reference value is 50. Values greater than 50 are represented by a brown hue, and values below 50 by a blue hue. Brightness shows how much an attribute value differs from the reference value, i.e. the degree of darkness is proportional to the difference between the attribute value and the reference value (in our example, 50).

A colour scale constructed according to the principle described above is called a diverging, or double-ended, colour scale (Brewer 1994). The reference value is often called the “midpoint” of this colour scale. Usually, attribute values equal to the midpoint are represented by a specially chosen colour, for example white. The purpose of using a diverging colour scale is to show deviations of attribute values from a chosen midpoint.

The function for the visual encoding of numeric attribute values using a diverging colour scale is defined generally as follows:

$$\begin{aligned} \text{hue} &= \begin{cases} H_1 & \text{if } x < R \\ H_2 & \text{if } x > R \end{cases} \\ \text{darkness} &= D \times \frac{|x - R|}{\max(M - R, R - m)} \end{aligned} \quad (4.2)$$

Here,  $x$  is the independent variable that stands for the attribute value to be encoded; *hue* and *darkness* are the dependent variables, which determine the hue and darkness of the shade to be used to represent this attribute value,  $R$  is the reference value, or midpoint,  $H_1$  and  $H_2$  are two different hues chosen for values below and above the midpoint, respectively,  $D$  is the maximum darkness,  $m$  is the minimum attribute value to be encoded, and  $M$  is the maximum attribute value to be encoded. The expression  $|N|$  means the absolute value of the number  $N$ , which equals  $N$  when  $N \geq 0$  and  $-N$  (minus  $N$ ) when  $N < 0$ . The expression  $\max(N_1, N_2)$  means the maximum of the numbers  $N_1$  and  $N_2$ . In the example shown in Fig. 4.34C,  $R = 50$ ,  $H_1 = \text{blue}$ ,  $H_2 = \text{brown}$ ,  $m = 46.44$ , and  $M = 56.94$  (after removing the outlier 86.17 by means of focusing).

The same idea can also be applied to other visualisation techniques. Thus, Fig. 4.35C demonstrates the cartographic visualisation technique known as “graduated circles”. In traditional cartography, the sizes (more precisely, areas) of the circles are proportional to the numeric attribute values that they represent. In a computer implementation, the technique of graduated circles may be modified so as to show, instead of the attribute values as such, their deviations from some chosen reference value. Thus,

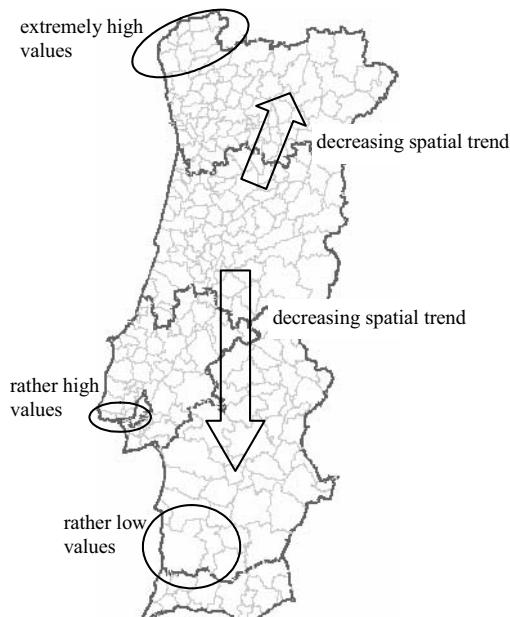
in Fig. 4.35C, the values of the attribute “% pop. no primary school education 1981” (the proportion of the population without a primary school education in 1981) in the districts of Portugal are compared with 50%. The sizes (areas) of the circles are proportional to the differences between the corresponding attribute values and 50. The circles representing values below 50 are coloured in cyan, and a red hue is used to represent values greater than 50. The encoding function in this case is almost the same as that specified in (4.2), except that the visual variable *size* is used instead of *darkness*, and the constant  $D$  denoting the maximum darkness is replaced by a certain maximum circle size  $S$ .

An apparent advantage of the maps in Figs 4.34C and 4.35C is the ease with which values below and above the reference value may be located. Thus, the groups of cyan circles in the north-west and central west of the map in Fig. 4.35C immediately attract the viewer’s attention. This excellent property can be enhanced by the possibility to easily change the reference value, which is not a problem in a computer display. Thus, the user can enter an exact value in a special field, choose a value by dragging a slider along an axis representing the value range of the attribute, or simply click on a district on the map to turn the corresponding value into the reference value. Any such operation results in the map being immediately updated taking the new reference value into account.

This responsiveness not only allows the user to perform quick and easy comparisons of values in all districts at once with particular values of interest (such as 50% for the proportion of women) or to compare a value in any district with the values around and over the whole country. It may also greatly help in a search for simple and understandable patterns in the spatial distribution. An analyst may gradually move the slider that controls the reference value and observe what happens on the map. The kinds of changes that may be expected are the emergence and evolution of various shapes resulting from the visual association of neighbouring objects or areas coloured in the same hue. The associative power of the visual variable “hue” is so great that human perception tends to unify objects with neighbours that have the same hue, regardless of any differences in their brightness. This makes the whole spatial distribution appear as a collection of integrated shapes, which are less numerous than the original set of districts or locations; hence, simplification without information loss is achieved. In the course of movement of the slider, the analyst notes the most prominent and clear-cut shapes that emerge on the map, and thereby advances his/her understanding of the data distribution.

Figure 4.36C presents several screenshots from such a process of investigating a spatial distribution by changing the midpoint of a diverging colour scale. As in Fig. 4.34C, the map represents the percentages of females

in the population in the districts of Portugal in 1991; the outlier 86.17 has been removed by means of focusing. In the leftmost map, the reference value is equal to the minimum attribute value 46.44; therefore, only shades of brown are present in the map, and the one district in which the minimum value is attained is white. The next screenshot (second from left) corresponds to the reference value 50, as in the map in Fig. 4.34C. A blue cluster of districts with values below 50% is easily detected in the south-west of the country. At a reference value of 51.06 (the third screenshot), a blue cluster in the north-east emerges, and the previously detected blue cluster in the south-west expands to the east and north. A small blue shape also appears in the west of the central part of the country. The fourth map corresponds to a reference value of 51.74. The former blue cluster in the north-east has extended to the west and now covers almost the whole northern part, except for the “horn” in the north-west, where the values of the attribute are, apparently, very high. The southern half of the territory is now mostly blue, except for a few brown “islands”. Brown shades still prevail between this half and the north, but they mostly disappear in the next screenshot corresponding to a reference value of 52.83. At this value, only the “horn” in the north-west, the district of Lisbon, and a few other scattered districts remain brown.



**Fig. 4.37.** A simplified representation of the spatial distribution of the percentages of females in the population over the districts of Portugal

From these observations, one can derive a pattern or mental image of the spatial distribution, which may look, for example, as is shown in Fig. 4.37. Hence, the “visual comparison” technique not only may help one to quickly find values below and above a certain threshold but also may be useful in studying the overall behaviour of an attribute over a reference set and approximating this behaviour by appropriate patterns.

A diverging colour scale can be used not only in an unclassified choropleth map but also in a choropleth map with classification. Compare, for example, the three maps in Fig. 4.38C. All maps represent the same classification of the districts of Portugal according to the relative change in the population from 1981 to 1991. We have defined the following classes: strong population decrease (the change is below  $-10\%$ ), moderate decrease (from  $-10$  to  $-2\%$ ), nearly stable population (the change is between  $-2$  and  $2\%$ ), moderate increase (from  $2$  to  $10\%$ ), and strong increase (over  $10\%$ ).

On the left of Fig. 4.38C, these classes are represented using a diverging colour scale, with shades of blue in one part, shades of red in the other part, and white colour in between. The colours are assigned to the classes so that the blue end of the colour scale corresponds to a population decrease (dark blue to a strong decrease and light blue to a moderate decrease), white corresponds to population stability, and red corresponds to a population increase (light red to a moderate increase and dark red to a strong increase). As in the case of the unclassified choropleth map, the use of the diverging colour scale exposes deviations from something, which in this case is a reference interval (from  $-2$  to  $2$ ) rather than a single reference value. The hue, blue or red, indicates the direction of deviation, i.e. lower or higher, and the darkness shows the amount of deviation. The selection of the reference class in such a choropleth map can be done interactively, like the selection of the reference value in an unclassified choropleth map.

By comparing the map based on the diverging colour scale with the two other maps in Fig. 4.38C, which represent the same classes by means of differently constructed single-hue colour scales, one can come to the conclusion that the former map supports much better the differentiation between areas with a population increase and those with a population decrease. It can be said that this map visually divides the territory into areas of increasing and decreasing population. In this respect, it differs from the maps in the centre and on the right, which do not impose any division of the territory but instead emphasise the coast-to-inland trend of decreasing population growth (centre) and increasing population loss (right). Hence, manipulation of the colour scale allows one to obtain complementary views of the spatial distribution and thereby arrive at a better understanding of the character of this distribution.

Let us also briefly discuss how the visual comparison technique can be applied to visualisations of several numeric attributes. Two different cases should be distinguished:

1. There is a common visual encoding function, applied equally to the values of all of the attributes.
2. Each attribute has its individual function for the visual encoding of its values.

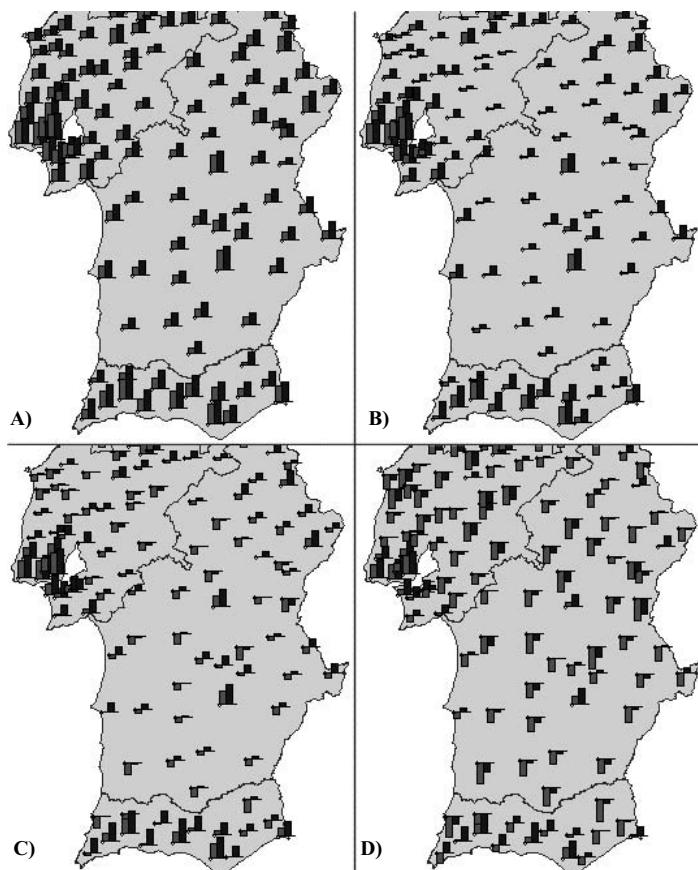
In the first case, the visual comparison technique is based on choosing a common reference value for all attributes. In the second case, an individual reference value may be specified for each of the attributes. Let us illustrate both cases by examples.

The map fragments in Fig. 4.39 demonstrate an application of the visual comparison technique to a bar chart representation of two numeric attributes, specifically, the percentages of the working population employed in services in the years 1981 and 1991 in the districts of Portugal. The values of the attributes range from 11.09 to 80.89 and from 20.12 to 85.57, respectively. For better visibility of the charts, we have switched off the drawing of the district boundaries and names. The map fragment denoted as A demonstrates the original appearance of the map, when no transformation has been applied yet. The heights of the bars are proportional to the values of the attributes (the visual encoding function presented in Example 1 in the Sect. 4.4.5 has been used to compute the bar heights).

The fragments B, C, and D result from applying the visual comparison operation with reference values of 30, 40, and 50, respectively. The operation changes the portrayal of attribute values so that downward-pointing bars represent values below the reference value and upward-pointing bars represent values higher than the reference value. The heights of the bars are proportional to the difference between the attribute value and the reference value. The visual encoding function thus takes the form

$$\begin{aligned} \text{orientation} &= \begin{cases} \text{up} & \text{if } x > R \\ \text{down} & \text{if } x < R \end{cases} \\ \text{height} &= H \times \frac{|x - R|}{M} \end{aligned} \quad (4.3)$$

where  $x$  is an attribute value,  $R$  is the reference value,  $H$  is the chosen maximum bar height, and  $M$  is the maximum of the values of all attributes. In fact, this is a more general form of the function for encoding numeric values by sizes of marks than that presented in Example 1 in Sect. 4.4.5. The latter is a special case of the function (4.3), which is valid when  $R = 0$  and all attribute values in the dataset are non-negative.



**Fig. 4.39.** The bar charts represent the percentages of working people employed in services in 1981 and 1991 in the districts of Portugal. The value ranges of the attributes are from 11.09 to 80.89 and from 20.12 to 85.57, respectively. A, original view; B, visual comparison with 30%; C, visual comparison with 40%; D, visual comparison with 50%

In the example just considered, the values of two numeric attributes with similar value ranges have been represented graphically using the same visual encoding function. Therefore, the reference value in the visual comparison operation has been common to both attributes. In the next example, we have four attributes with rather different value ranges:

1. “% pop. no primary school education 1991” (the percentage of people without primary school education in 1991), ranging from 7.33 to 37.86
2. “% pop. with primary school education” (the percentage of people with primary school education in 1991), ranging from 18.69 to 35.14

3. “% pop. with preparatory school education 1991” (the percentage of people with preparatory school education in 1991), ranging from 3.43 to 16.97
4. “% pop. with high school education” (the percentage of people with high school education in 1991), ranging from 1.29 to 13.83.

The smallest value of the second attribute is higher than the maximum values of the third and fourth attributes. If a common visual encoding function were applied to those data, all bar charts would look nearly the same. The second bar would always be very tall and the third and fourth bars very short. Therefore, it makes more sense to use an individual encoding function for each attribute.

The bar chart visualisation in Fig. 4.40 has been constructed differently from that in the previous example. First, in the “original” view (part A), an attribute value has been encoded by a bar with a height proportional not to that value itself but to the difference between the value and the minimum value of the attribute. In other words, the bar height portrays the distance from the current attribute value to the minimum value of that attribute. Hence, a bar of zero height represents the smallest value of the respective attribute available in the dataset, not the value 0. Second, while the maximum bar height is common to all four attributes, it corresponds to a distinct value for each of the attributes, specifically, the maximum value of this attribute. Hence, the “worth” of a unit of bar height differs from attribute to attribute.

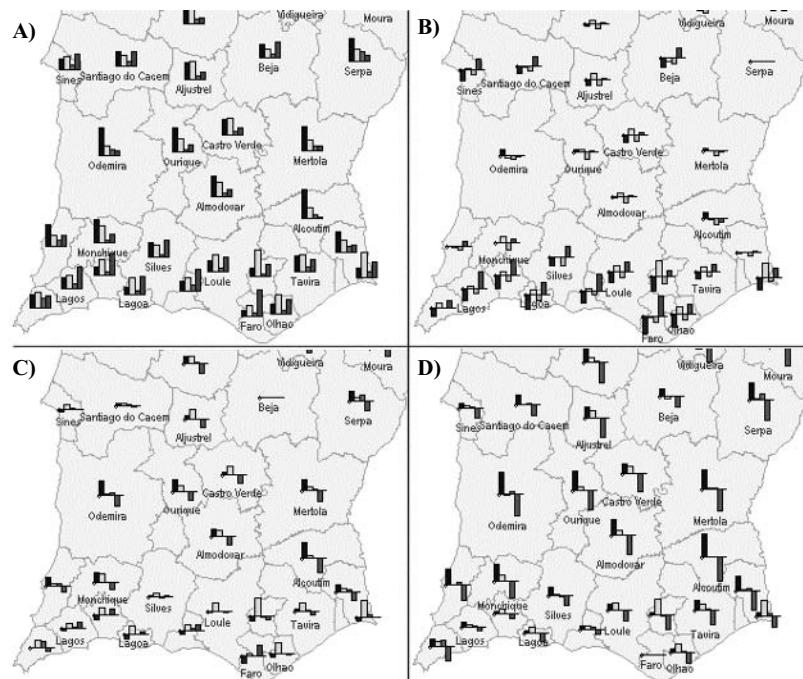
The visualisation can be manipulated by choosing an individual reference value for each attribute, which results in the bar charts being transformed in the same way as in the previous example. The visual encoding function differs slightly from the previous one:

$$\begin{aligned} \text{orientation} &= \begin{cases} \text{up} & \text{if } x > R \\ \text{down} & \text{if } x < R \end{cases} \\ \text{height} &= H \times \frac{|x - R|}{M - m} \end{aligned} \quad (4.4)$$

where  $m$  is the minimum value of the attribute, and  $m \leq R \leq M$ . The original view (fragment A in Fig. 4.40) corresponds to  $R = m$ .

The easiest way to specify the reference values for all attributes at once is to select a particular district, for example by clicking on it in the map. The attribute values characterising this district will be taken as the reference values, that is, all other districts will be compared with this district. Naturally, for the selected district, all bars will have zero height.

The fragments B, C, and D in Fig. 4.40 demonstrate how the original visualisation is transformed after the selection of three different districts: Serpa (in the north-east of the territory shown), Beja (west of Serpa), and Faro (in the southern coast). Within a fragment, upward-oriented bars represent higher values than in the selected district, and downward-oriented bars correspond to lower values than in the selected district. The heights of the bars show the “distance” to the selected district in the attribute space. Districts with short bars are close in their characteristics to the selected district.



**Fig. 4.40.** Manipulation of a bar chart map with individual encoding of each attribute. The bars represent values of the attributes “% pop. no primary school education 1991” (ranging from 7.33 to 37.86), “% pop. with primary school education” (from 18.69 to 35.14), “% pop. with preparatory school education 1991” (from 3.43 to 16.97), and “% pop. with high school education” (from 1.29 to 13.83). A, original view; B, comparison with Serpa, with attribute values 29.1, 24.73, 7.61, and 3.51; C, comparison with Beja, with values 20.08, 23.05, 4.94, and 7.59; D, comparison with Faro, with values 12.93, 24.20, 4.93, and 11.78

It can be noted that in both examples using multiple attributes, we used the visual comparison technique merely for performing elementary comparisons, i.e. for analysis on the level of individual elements of a reference

set. We have not demonstrated the application of this technique to an investigation of the overall spatial distribution of characteristics, as we did with visual comparison on a choropleth map (see Fig. 4.36C). In the case of the choropleth map, the visual comparison operation leads to display simplification owing to visual association of objects into larger shapes. This facilitates the perception of the distribution as a whole, which is required for synoptic tasks. In the case of a chart map, no visual association occurs. Each chart can easily be considered individually, but multiple charts can hardly be united into a coherent image. The visual comparison operation does not change the nature of the charts; in particular, it cannot suppress their resistance to unification. Therefore, both chart-based visualisations and the techniques for manipulating them are suitable primarily for elementary tasks rather than synoptic ones.

#### 4.4.7 Recap: Display Manipulation

Data visualisation is based on encoding values of data components by values of visual dimensions and variables. We call the mechanism of defining the correspondence between elements of data and the graphical features that represent them the “visual encoding function”. Such a function is often (but not necessarily) specified by means of a mathematical formula. It can also have the form of a set of rules or a decision table.

Irrespective of the form, a visual encoding function always involves some constants or parameters. Software implementation of a visualisation technique may be done in such a way that the user can interactively change the values of these constants or parameters or even substitute one visual encoding function for another. Such changes are called display manipulation. When the user has convenient and easy-to-use controls for modifying the visual encoding of data, and the display reacts promptly to the user’s action, display manipulation may become a powerful tool for exploratory data analysis.

In this section, we have considered several types of display manipulation techniques. By means of numerous examples, we have demonstrated what services these techniques can provide to a data explorer. Here is a brief summary of the techniques.

- *Ordering, reordering, and arrangement:* Changing the positions of individual marks or display fragments (groups of marks) within one or more display dimension. These techniques preserve all of the information originally available in a display. Ordering techniques may result in simplification of the display, and a clearer overall view of the distribution of characteristics over a dataset and of relations between different at-

tributes. Two-dimensional arrangement is helpful in the exploration of periodic data, such as time-related data with yearly, weekly, or daily cycles. It can also be used for detecting periodicity in data.

- *Smoothing and generalisation*: Elimination of irrelevant detail and random fluctuations in order to reveal the principal features of a behaviour, for example a spatial distribution or temporal variation. Simplification is achieved at the cost of a substantial reduction of the information.
- *Classification*: A kind of generalisation based on uniting references with close characteristics into groups and regarding members of a group as identical. Like other generalisation methods, classification simplifies the visualisation at the cost of information loss. Classification favours the perception of a display as a single image and supports pattern-building. As with any technique involving information loss, an explorer should “play” with the classes (i.e. redefine them in various ways) rather than base the analysis on a single variant of classification.
- *Zooming and focusing*: Reduction of the amount of data in a display so that a selected subset of the data can be represented with the maximum possible expressiveness. Display expressiveness includes such aspects as legibility (which depends on the sizes of the marks and the presence or absence of cluttering and overlap), the level of detail, and differentiability, i.e. whether different data values are converted into perceptually different graphical features. The term “zooming” is usually applied to the representation of data by spatial display dimensions. Zooming appears as the enlargement of a selected part of a display area in order to improve its legibility. This may be accompanied by an increase in the level of detail. The term “focusing” is mostly applied to the representation of data by retinal variables. The main idea is to use the whole set of perceptually distinct values of a visual variable for encoding a reduced subset of data values in order to improve the differentiation between those data values.
- *Substitution of a linear encoding function by a non-linear one* may be recommended when the statistical distribution of attribute values is greatly skewed. If, for example, a dataset contains a few very high values while the majority of values are quite small, a linear function may provide only a small number of perceptually different values of a visual variable for encoding the bulk of the attribute values. As a result, these attribute values are visually indistinguishable in the display. An analogous situation occurs when almost all attribute values are quite high, while there are a few very low values. In the former case, a logarithmic encoding function may improve display differentiability, and in the latter case, an exponential function may be used. Other monotonic non-

linear functions can be applied as well, depending on the peculiarities of the value distribution. However, an analyst should be cautious in interpreting the results of non-linear visual data encoding. It is better to use non-linear and linear displays in parallel and to verify any observation obtained from a non-linear display using a linear display or another tool that is free from data distortion.

- *Visual comparison:* Displaying and/or emphasising the deviations of numeric attribute values from a particular reference value. The reference value is specified and changed interactively, which results in the display being dynamically updated. We have considered several examples of visual comparison techniques, which represent the degree of deviation by either brightness (or darkness) or the size of a mark, and the direction of deviation (i.e. whether the value is smaller or greater than the reference value) by hue or orientation. The use of hue favours the visual association of neighbouring marks with a common direction of deviation from the reference value. This, in turn, supports the grasping of distinctive features of a behaviour, and pattern-building. With any variant of visual encoding, visual comparison techniques support comparisons on the elementary analysis level.

We have not considered changes of display properties that are made mostly for aesthetic reasons or for rhetorical purposes, such as producing a convincing argument or an emotional effect. We have reviewed and discussed only display manipulation techniques that are capable of supporting exploratory data analysis or improving the capabilities of data displays to support this kind of analysis. We cannot guarantee that our enumeration of the types of techniques is exhaustive. It is based mostly on our practical work experience, but we have not encountered, either in the literature or at software demonstrations, any other display manipulation tools that would not fit into any of these groups. Nevertheless, we leave this list open to the inclusion of new categories of display manipulation tools.

Let us now move to other types of exploratory tools. While display manipulation tools modify the way data are visually encoded but do not change the data themselves, the next group of tools that we are going to consider consists of tools that manipulate data, i.e. transform the data or derive new data from them.

## 4.5 Data Manipulation

As we have demonstrated, manipulation of graphical displays often allows one to see what was previously not evident, and to look at data from dif-

ferent perspectives to investigate various aspects. In other cases, however, display manipulation can make the view simpler and easier to comprehend. Data manipulation serves, in principle, the same purposes, and we can group data manipulation tools into two broad categories: *sophistication* and *simplification*. Sophistication means creating conditions for the thorough investigation of various aspects of the data by increasing the initial amount of data, i.e. enriching the original dataset with additional attributes or additional references (and corresponding characteristics). Simplification means decreasing the amount of data under consideration, i.e. the number of attributes or the number of references with their corresponding characteristics. Of course, this does not mean that part of data is simply thrown away. Simplification is achieved by means of generalisation and abstraction. Ideally, simplification should be done so that no valuable information is lost.

As we have mentioned, both sophistication and simplification may involve either attributes or references. Hence, one can distinguish four groups of data manipulation tools, which are summarised in Table 4.8.

**Table 4.8.** Four groups of data manipulation tools

	<b>Sophistication</b>	<b>Simplification</b>
Attributes	<i>Attribute transformation</i> : deriving additional attributes from existing ones, e.g. transforming absolute quantities into relative quantities	<i>Attribute integration</i> : combining several attributes into a single attribute, which substitutes for the original attributes in further analysis
References	<i>Interpolation</i> : inserting additional references between the original ones and deriving the characteristics of the new references from those of the neighbouring references	<i>Aggregation</i> : grouping references and considering the groups and their collective characteristics instead of the original references

Before discussing each of these groups of tools, we would like to point out that the most reasonable thing to do is to use data manipulation tools in combination with visualisation, rather than alone. First, it is always advisable to take a preliminary look at the data before starting to manipulate them: this may help one to choose appropriate manipulation techniques. Second, one needs visualisation in order to see the results obtained from the data manipulation. All our examples of the use of data manipulation tools cited throughout this section are illustrated by visual displays showing the results of the manipulation.

## 4.5.1 Attribute Transformation

### 4.5.1.1 “*Relativisation*”

As we have mentioned, attribute transformation is intended for looking at the same characteristics from various perspectives in order to investigate them more comprehensively. One of the most commonly used transformations is that from absolute to relative numbers. For example, the Portuguese census dataset initially contained only absolute numbers, such as the number of people in various age categories (0–14 years, 15–24 years, 25–64 years, and 65 years and more) and the numbers of people employed in agriculture, industry, and services. Nevertheless, many of our example visualisations given in the previous sections represent proportions rather than absolute numbers: the proportions of children, of elderly people, of people working in different sectors of the economy, etc. These proportions were obtained by transforming the original absolute attributes: the number of people in each age group was divided by the total population of the respective district, and the number of people working in each sector was divided by the total number of working population. The transformation allowed us to consider the age and employment groups as parts of certain wholes, i.e. either the entire population of the district or the working population. This aspect of the data could not be investigated only with the use of the original attributes.

Figures 4.41C and 4.42C demonstrate that attribute transformation indeed results in producing new attributes with properties and behaviours quite different from those of the original attributes. Figure 4.41C shows a visualisation of the attribute “number of people without primary school education” referring to the districts of Portugal. Figure 4.42C shows the result of transforming the values of this attribute from absolute numbers to relative numbers, specifically, to the proportions of people without primary school education in the populations of the districts.

In Fig. 4.41C, the values of the original (i.e. absolute) attribute are represented on a map of Portugal using the technique of graduated circles. In the initial display (left), the size of a circle is proportional to the attribute value in the corresponding district. Beside the map, the value range of the attribute is shown, from 525 to 56 442. We can also see from the dot plot to the right of the map that the maximum value lies far from the rest of the values. The image on the right in Fig. 4.41C results from applying two display manipulation operations: focusing, which has removed the outlier 56 442 from the representation, and visual comparison with the country mean, 5140, taken as the reference value (the techniques of focusing and visual comparison have been described in the previous section).

Figure 4.42C shows the proportions of uneducated people in the populations of the districts, i.e. the result of dividing the values of the attribute “number of people without primary school education” by the total numbers of inhabitants in the respective districts. The same visualisation technique as in Fig. 4.41C has been used. The new attribute has a value range from 16.69% to 57.44%, and the mean value for the whole country is 33.36%. The maximum value, 57.44, is not as distant from the rest of the values as the maximum of the original attribute is, and is attained in another district. On the right in Fig. 4.42C, a visual comparison operation has been applied, taking the country mean as the reference value.

It does not take much effort to notice that the original attribute (Fig. 4.41C) and the derived attribute (Fig. 4.42C) have quite different spatial behaviours. The original attribute has high values on the western coast, especially in the north-west (around Porto) and in the centre (around Lisbon), while the derived attribute exhibits the opposite pattern. This is especially well seen from the maps where visual comparison has been applied, i.e. the right parts of Figs 4.41C and 4.42C.

A question may be asked: which attribute is the “right” one, or which one should be analysed? The answer is, in general, both. If we want to understand better the situation concerning the education in Portugal, we need to consider both absolute and relative values. However, our study may have particular goals. In that case, it depends on the goals which of the attributes is more relevant. Thus, if we want to know where additional facilities for primary education are required, we should pay attention to the absolute values. If we want to evaluate the education level in each district, or to see how the situation changed from the census year 1981 to the census year 1991, or to relate education to employment structure, we should deal primarily with relative numbers.

Besides computing proportions of the parts of a whole, transformation from absolute to relative numbers may be done in a number of other ways. Thus, for spatially referenced data, it is common to compute densities, i.e. amounts per unit area, for example the population density. A density is computed by dividing a certain amount (e.g. the number of inhabitants) specified for a compartment (e.g. a district) of a territory by the area of this compartment. An implicit assumption is made that the amount is evenly distributed within the compartment, which may not always be the case. For example, all of the population of a district may be concentrated in a small part of it, owing to particular natural conditions (mountains, swamps, deserts, etc.) in the other parts. Hence, an analyst should be cautious in using the results of such transformations.

In demographic studies, certain absolute values are often transformed into “per capita” or “per household” values. For example, countries can be

characterised in terms of the gross domestic product per capita. On a more local level, an analyst may be interested in studying the behaviour of the number of cars per capita or the number of children per household.

One and the same absolute attribute can often be “relativised” in several different ways. For example, the numbers of unemployed women in the districts of Portugal can be divided by the total population of the district, by the female population, by the number of employed women, by the total number of unemployed people, or by the number of unemployed men. Each division gives us a different perspective in a study of female unemployment.

Actually, the term “relative” is very general, and a broad group of transformations can be viewed as converting absolute values into relative ones. Thus, the original values of an attribute may be transformed into their relative positions with respect to the minimum and maximum values of this attribute. This transformation may be applied when it is necessary to analyse jointly several otherwise incomparable attributes, such as fertility rate, infant mortality rate, and female life expectancy. Another possibility could be to convert the values of each attribute into relative deviations from the mean value of this attribute. Such deviations may be computed simply as ratios to the mean value, or in proportion to the standard deviation. The latter transformation is called the “standard normal transformation” (Burt and Barber 1996, p. 196) and is specified by the formula

$$z = \frac{x - \mu}{\sigma} \quad (4.5)$$

where  $\mu$  is the mean value,  $\sigma$  is the standard deviation,  $x$  is the original attribute value, and  $z$  is the transformed value, which is called the standard score, or  $z$ -score.

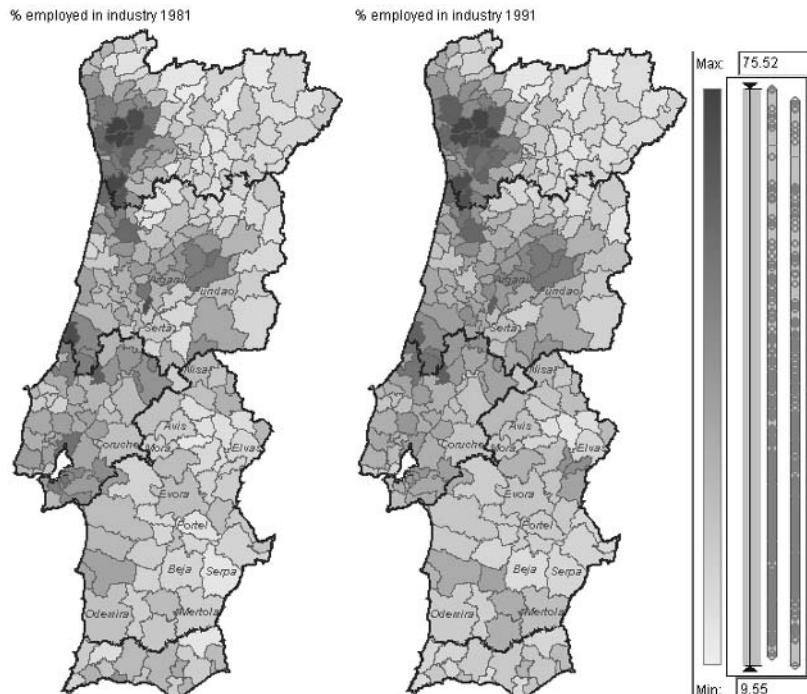
Transformations on the basis of the mean and standard deviation can only be recommended when the statistical distribution of attribute values is close to normal, and are certainly not recommended when there are outliers. For an attribute with outliers, it is more suitable to transform values in relation to other statistical measures, such as the median and quartiles (or other percentiles).

#### **4.5.1.2 Computing Changes**

When time-referenced numeric attributes are being explored, it is appropriate to consider not only the original attribute values, referring to different time moments, but also changes, i.e. differences from or ratios to values for preceding time moments or values pertaining to the beginning of the time series. In cartography, there is even a term “change map” to de-

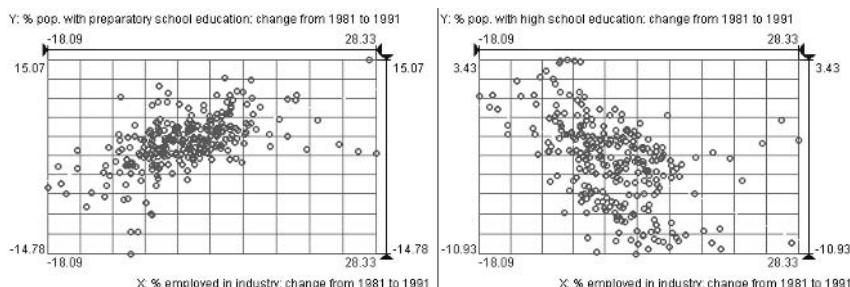
note a map portraying differences or ratios between attribute values for two time moments (Slocum 1999). Figure 4.43C contains an example of a change map, which represents the changes in the proportions of people employed in industry over the districts of Portugal from 1981 to 1991. A diverging colour scale is used to differentiate districts where the values increased (these are coloured in shades of brown) from those where the values decreased (coloured in shades of blue). The degree of darkness shows how much increase or decrease occurred in a district.

A change map is certainly much more convenient for detecting where an increase or decrease occurred and for estimating the amounts of change than is a representation of the original attribute values referring to the two years, for example on two choropleth maps, as is shown in Fig. 4.44. This does not mean, however, that comparison of maps representing states at different time moments is useless: it is necessary for seeing changes in the spatial distribution of attribute values. In particular, Fig. 4.44 shows us that the spatial distribution of the proportion of people employed in industry over Portugal did not change significantly from 1981 to 1991.



**Fig. 4.44.** The proportions of people employed in industry in 1981 and 1991 are represented here on two choropleth maps. From this representation, the changes that occurred between the two census years are quite hard to estimate

In order to demonstrate that change maps are not the only tools suitable for visualising and analysing changes, we have constructed two scatterplots representing the changes in the proportion of people employed in industry together with two other derived attributes: the changes in the proportion of people with preparatory school education and the changes in the proportion of people with high school education (Fig. 4.45). One can observe a slight negative correlation in the right scatterplot (the change in employment in industry against the change in the proportion of people with high school education). The pattern perceived from the left scatterplot (the change in employment in industry against the change in the proportion of people with preparatory school education) can be interpreted, though with less certainty, as a slight positive correlation.



**Fig. 4.45.** The changes in the proportions of people employed in industry are analysed together with the changes in the proportions of people with preparatory school education and with high school education

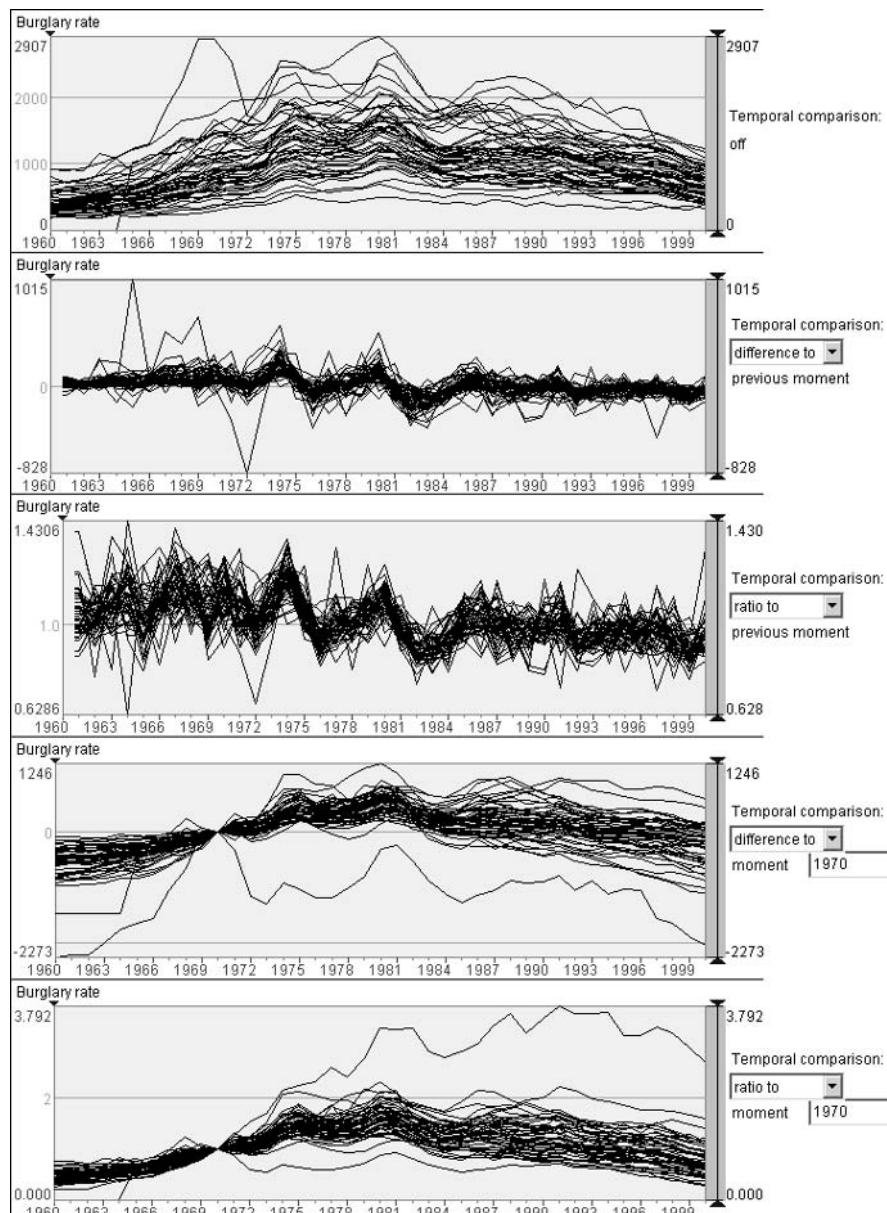
In the Portuguese census dataset, we have data for only two time moments, specifically, the census years 1981 and 1991. Hence, for any attribute, we can compute only the absolute or relative change (i.e. difference or ratio) from 1981 to 1991. When longer time series are explored, many more changes need to be computed and analysed. For example, the US crime dataset contains data for 41 time moments, specifically, the years from 1960 to 2000. An analyst needs a convenient tool to compute and visualise changes in each year in comparison with the previous year, as well as in comparison with the initial state (i.e. that in the year 1960) or with any selected time moment. A possible solution is the use of either map animation or “small multiples” for the visualisation of changes. When map animation is applied, the map at each display moment represents transformed attribute values, i.e. computed changes, rather than the original values. The same applies to a “small multiples” visualisation: each map in this display is a change map.

For example, Fig. 4.46C presents a “small multiples” display of the changes in the burglary rates in the states of the USA during the years from 1991 to 2000 (to save space, we have removed Alaska and Hawaii from the display and restricted the display to the last 10 years of the 41-year period that the data refer to). For each year, the changes with respect to the previous year have been computed, specifically, the differences between the values in the given year and the values in the same districts in the previous year. The changes are represented on the maps using a diverging colour scale. Brown shades encode positive differences, i.e. an increase in the burglary rate in comparison with the previous year, and blue shades correspond to a decrease in comparison with the previous year.

In the same way, differences or ratios with respect to any selected time moment can be visualised. In this case, each map in a “small multiples” display or each frame in a map animation represents changes with respect to the same year, for example, 1960, the beginning of the period under study. This is different from the visualisation in Fig. 4.46C, where each of the small maps represents changes with respect to its own “reference year”: the map for 1991 shows the changes as compared with 1990, the map for 1992 shows the changes as compared with 1991, and so on.

Changes over time can be explored using not only change maps but also other visualisation tools, for example time graphs. Thus, Fig. 4.47 shows five different appearances of a time graph display representing the dynamics of the burglary rates in the states of the USA over the period from 1960 to 2000. Initially (top of Fig. 4.47), the display represents the original values of the attribute “Burglary rate”. The horizontal axis of the display represents the time period that the data refer to. The values of the attribute are encoded by positions in the vertical dimension. Each sequence of positions corresponding to one state is linked into a line.

The other four images in Fig. 4.47 demonstrate the results of various transformations applied to the values of the attribute “Burglary rate”. The transformation method is indicated to the right of each image. First, we have computed and visualised the changes with respect to the preceding year. This means that the vertical position corresponding to a state  $S$  and a year  $Y$  represents the arithmetic difference between the value of the burglary rate in the state  $S$  in the year  $Y$  and the value of the burglary rate in this state in the year  $Y-1$ . Similarly, the next display (third from top) represents the ratios with respect to the previous year. It can be noted that the lines in both the second and the third display look shorter than in the other graphs. This is because there are no positions corresponding to the year 1960: the data for the year 1959 are not available in the dataset, and therefore the changes in 1960 with respect to the previous year cannot be computed.

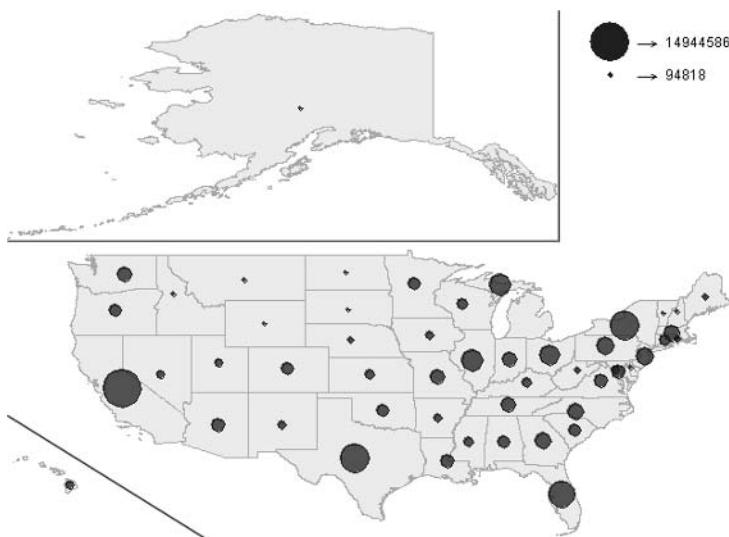


**Fig. 4.47.** A time graph can represent not only the original values of a time-related numeric attribute but also computed changes. Here, the time graphs represent, from top to bottom, the original burglary rates in the states of the USA, the differences from the preceding years, the ratios to the preceding years, the differences from the rates for the year 1970, and the ratios to the rates for the year 1970

The remaining two images at the bottom of Fig. 4.47 represent the changes with respect to the year 1970. The display at the very bottom represents the ratios, and the one above it shows the differences. It can be noted that all lines cross in one point corresponding to the year 1970, since the results of the computation for all states for this year coincide: the differences equal 0 and the ratios equal 1.

#### 4.5.1.3 Accumulation

Besides computing changes, it may be useful for certain type of time-related attributes also to sum (accumulate) values over time intervals. This transformation is applicable only to quantitative attributes, i.e. attributes whose values express counts or amounts. Moreover, a value referring to a time moment must represent some quantity that appeared only at that moment and did not exist before or after that moment (and, hence, is not included in the values referring to the previous and subsequent moments). Examples of this sort of attribute are counts of events such as the total number of burglary incidents that happened during a year in each state of the USA, the yearly gross domestic product or total imports and exports of various countries, the daily amount of rainfall, and so on. As a negative example, we could mention the population number: although this attribute is quantitative, it does not reflect an “added” quantity, in contrast, for example, to the number of newborn.



**Fig. 4.48.** The graduated circles represent the total numbers of burglary incidents that occurred in the states of the USA during the time period from 1960 to 2000

Since an attribute represents something added (or, sometimes, deducted, as in the number of deaths) at each time moment, it makes sense to count how much of this has been added (or deducted) in total over a sequence of time moments. Thus, the USA crime dataset contains a number of attributes representing the total numbers of crimes of different types in each year from 1960 to 2000 and each state of the USA. Hence, we can compute and analyse the total number of crime incidents that occurred in each of the states during the whole period from 1960 to 2000 or any of its subintervals. For example, the map in Fig. 4.48 represents the total number of burglary incidents that occurred in each state during the 41-year period. One can visualise not only the final count but also the dynamics of the accumulation of crimes, for example, on a time graph.

#### **4.5.1.4 Neighbourhood-Based Attribute Transformations**

In the examples of attribute transformations that we have considered thus far, two different types of transformation may be traced:

1. The value of a derived attribute is produced for each reference from the values of one or more source attributes corresponding to this reference. No attribute values associated with other references are involved.
2. The value of a derived attribute is produced for a reference using not only attribute values corresponding to this reference but also values associated with other reference(s).

Thus, in transforming absolute attribute values into relative values, the values of one attribute are divided by values of another attribute associated with the same reference. To compute the percentage of people without education in each district of Portugal, the number of such people in each district is divided by the total population of that district. No other districts have any influence on the computation; hence, this is a transformation of the first type. The second type of transformation takes place, for example, when changes of attribute values with respect to time are computed: each derived value is a difference or a ratio of values referring to two time moments; hence, one additional reference is involved in the computation. In value accumulation, the number of additional references varies depending on the length of the interval over which the values are summed.

For convenience in further discussion, let us introduce the term *target reference* to denote a reference for which the corresponding value of a derived attribute needs to be obtained. If this value is produced using attribute values corresponding to references other than the target reference, we shall call these other references *contributing references* and the corresponding attribute values *contributory values*. The values of the source

attributes associated with the target reference will be called *primary values*.

Using this terminology, we can characterise the first type of attribute transformation as involving only primary values of source attributes and the second type as based on primary and contributory values (in principle, there may be transformations using only contributory values). The second type may be subdivided further according to whether the same or different contributing references provide attribute values in the production of derived attribute values for different target references. For example, in computing the changes of the burglary rates in all years with respect to a fixed year (such as the year 1970 in Fig. 4.47), this fixed year is the contributing reference, used equally for all years. However, in computing the changes in each year with respect to the previous year, an individual contributing reference is used for each year: 1960 for 1961, 1961 for 1962, and so on. Such individual contributing references are chosen according to a certain rule, which specifies the relation that exists between the target reference and the contributing reference(s). Thus, in comparing with the previous year, the rule is “the contributing year equals the target year minus one”.

We are now reconsidering these examples of transformations of temporally referenced data (i.e. computation of changes, and accumulation) in order to present them as instantiations of a more general category of transformations and, on this basis, to draw a parallel with a certain class of transformations of spatially referenced data. However, before moving to spatial data, let us consider one more transformation of temporal data, which also involves contributory values.

We have already talked about smoothing as a technique for the simplification of data displays. In particular, we have discussed smoothing of a time graph representing temporally referenced data (see Figs 4.14 and 4.15). At that point, we considered time graph smoothing as just graphical simplification. However, this graphical simplification is based on a transformation of the underlying data. Let us now take a closer look at this transformation, which is also called “smoothing”, or “data smoothing” (in contrast to graphical smoothing). We shall introduce the general idea with an example of a specific smoothing technique, which is called the “simple moving average”. This technique is rather popular; in particular, it is intensively used in stock market analyses.

A simple moving average is formed by computing the average (mean) value of an attribute over a specified number of consecutive time moments, one of which is the target time moment.<sup>9</sup> Usually, the target time

---

<sup>9</sup> Different smoothing techniques may use other operations instead of the mean. However, any technique would involve a sequence of consecutive moments,

moment either is the last in the sequence of moments or is positioned in the middle of the sequence (the latter case is called the “centred moving average”). For example, the non-centred 5-year moving average of the burglary rate in a state of the USA for the target year 1970 is computed by adding the burglary rates in this state for the years 1966, 1967, 1968, 1969, and 1970 and dividing the sum by 5. The centred 5-year moving average for the target year 1970 is computed in the same way from the burglary rate values for the years 1968, 1969, 1970, 1971, and 1972. For the target year 1971, the non-centred moving average is derived from the values for the years 1967, 1968, 1969, 1970, and 1971, whereas the centred moving average is computed from the values for the years 1969, 1970, 1971, 1972, and 1973, and so on for other target years. It may be easily guessed how moving averages over 3-year or 7-year time periods are computed. Computing the centred moving average over a period consisting of an even number of years, such as 4-year or 6-year period, is slightly more complicated. Thus, the 4-year centred moving average for the year 1970 is computed as the average of two averages, one for the period from 1968 to 1971 and another for the period from 1969 to 1972. To state this more generally, the centred moving average of an attribute for a target moment  $t$  over a period of length  $2d$  is the average of the averages for the periods from  $t-d$  to  $t+d-1$  and from  $t-d+1$  to  $t+d$ .

We shall not discuss here the effect of the length chosen for the averaging period on the results of the data transformation. We have briefly touched upon these issues in comparing the smoothed time graph in Fig. 4.14, which was produced using a 5-year centred moving average, with the original time graph. We showed that the 5-year smoothing hid some distinctive features of the behaviour under analysis, while the smoothing on the basis of 3-year intervals applied in Fig. 4.15 preserved those features. A more detailed consideration of this topic can be found in statistical handbooks, for example Burt and Barber (1996). Interested readers can also refer to the Web, where there are several tutorials on stock market analysis explaining how to use moving averages computed over intervals of different lengths (look, for example, at [http://www.stockcharts.com/education/IndicatorAnalysis/indic\\_movingAvg.html](http://www.stockcharts.com/education/IndicatorAnalysis/indic_movingAvg.html)).

Besides simple moving averages, exponential moving averages, or exponentially weighted moving averages, are often used in analysing time-referenced data, in particular, stock market data. The basic idea is to

---

including the target moment. In the current context, the focus is on the selection of contributing references rather than on a specific operation to be applied to primary and contributory attribute values. Accordingly, most points of the following discussion can be related to arbitrary smoothing techniques.

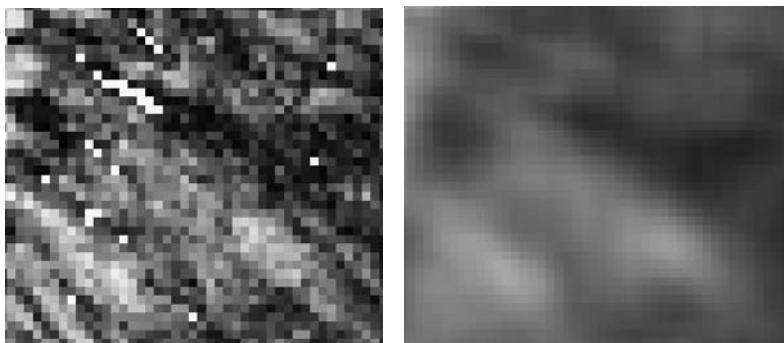
weight the contribution of each time moment to the derived value according to the distance from this time moment to the target time moment: the closer to the target time moment, the more influence the corresponding attribute value has on the resulting value.

The method of deriving moving averages can be formulated in more general terms as computing the average (or a weighted average) value of an attribute over a set of references consisting of the target reference and a specified number of contributing references, chosen in a specified way from the neighbourhood of the target reference. The advantage of this general formulation is that it may be applied not only to time-referenced attributes but also to numeric attributes defined on any reference set with distances (the presence of distances is essential for the notion of neighbourhood to be meaningful). In particular, we may apply it to spatially referenced attributes, which may also be smoothed using the moving-average technique. In this case, each derived value is produced from a value in a target location and a number of values taken from the neighbourhood of this target location. Of course, neighbourhood in space is defined differently from neighbourhood in time. Usually, one specifies a certain distance (radius), and all locations within this distance from the target location are taken as contributing locations. This method of choosing contributing locations applies not only to the spatial smoothing performed with the use of the moving-average technique but also to many other transformations of spatially referenced data. Such transformations are described in detail in the GIS literature. Here, we shall consider just two examples.

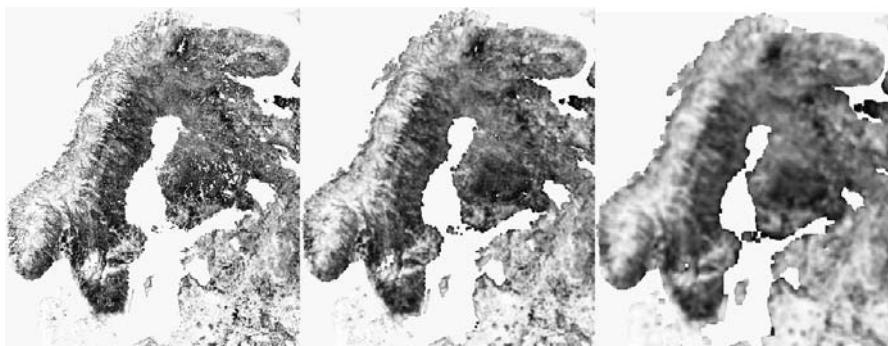
Figures 4.49 and 4.50 demonstrate the effect of applying a smoothing transformation to data concerning the proportion of land covered by coniferous forest. The data are specified using the raster model, i.e. as attribute values referring to cells of a regular rectangular grid.

The images on the left and on the right in Fig. 4.49 correspond to the same territory fragment. On the left, the original attribute values are shown, i.e. the proportions of coniferous forest in the grid cells. On the right, values derived by means of smoothing are portrayed. The value for each cell has been computed from the original value in this cell and the original values in the surrounding cells within a certain specified distance from the cell (the averaging radius). In both images, values are represented by shades of grey, with darker shades corresponding to higher proportions.

In Fig. 4.50, a larger territory fragment is shown. The map fragment on the left portrays the original data. In the centre and on the right, the results of smoothing these data with two different averaging radii are demonstrated. A smaller averaging radius has been chosen for the map in the centre than for the map on the right.



**Fig. 4.49.** The effect of smoothing is demonstrated here in an enlarged map fragment representing the proportion of coniferous forest in cells of a regular rectangular grid. Left, the original data; right, the smoothed data. Darker shades correspond to higher proportions



**Fig. 4.50.** Smoothing of spatially referenced data. Left, the original data; centre and right, the results of averaging over neighbouring locations within a smaller (centre) and a larger (right) distance from a target location

From comparing the representations of the original and transformed data, it may be noted that the effect of spatial smoothing is similar to that of temporal smoothing: small differences (often called “noise”) are suppressed so that an analyst may focus on larger structural features. In general, smoothing is defined as a technique that can be used to remove or reduce local noise within spatial or temporal data and therefore reveal the global pattern or trend. As in the examples of temporal data discussed earlier, the degree of smoothing depends on the number of contributing references involved in the transformation. In the spatial case, the number of contributing locations is determined by the smoothing radius chosen. The larger the radius, the less detail (small-scale features) remains in the result of the transformation. It is hard to say what level of detail is appropriate –

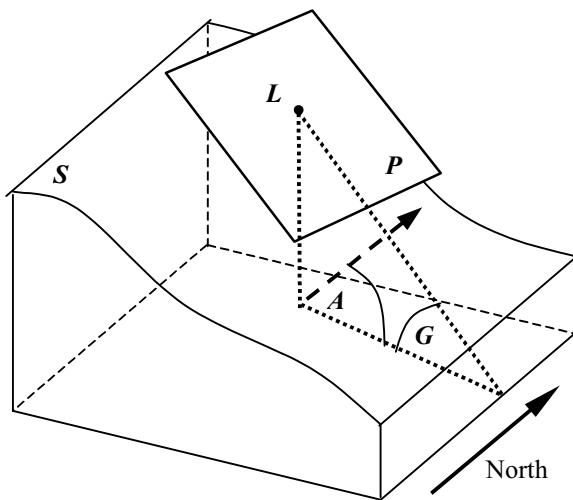
this depends on the data and analyst's goals – but the general recommendation is to consider the same data with different degrees of smoothing.

It is also worth adding that smoothing may be done using not only the method of the moving average but also other computational formulae or algorithms. Thus, in exploratory data analysis, it is typically recommended that one should smooth temporal or spatial data using medians rather than averages. The median of a set of numeric values is defined as the value that divides the set into two equal parts so that the values in one part are less than or equal to this value and the values in the other part are greater than or equal to this value. In particular, the median of a set of three values is the value that lies between the two others. The reason for using medians is that they are less subject to substantial variations due to a single outlier, as compared with means.

The simplest median-based smoothing method is known as “running medians”. The principle of this method is the same as for the moving-average technique. The statistical literature (see, for example, Burt and Barber (1996)) also describes a number of more sophisticated smoothing methods. Thus, it is possible to apply smoothing to the results of another smoothing operation. Such repeated smoothing may be applied until there is no change in the resulting data. In order to reduce the risk of oversmoothing, which may remove interesting patterns, it is suggested that one should use “compound smoothers”. The main idea can be explained as follows. After an elementary smoothing method has been applied to the original data, the differences between the original and the transformed values are computed. These differences are called the *rough*, or *residuals*, and the transformed values are called the *smooth*. Then, another elementary smoothing method is applied to the rough. The final result is the sum of the original smooth and the smoothed rough. Here, the term “elementary smoothing” is used as an antonym of “compound smoothing”, i.e. in the sense that smoothed values are derived directly from source values, without the separation of residuals. The method used for the elementary smoothing is not necessarily simple. Thus, repeated smoothing is also elementary smoothing in this context.

As with time series, the smoothing of spatially referenced data may be based on computing weighted averages of values in the target and contributing locations. The weights assigned to the contributing locations are usually inversely proportional to their distance from the target location. For various possible methods of assigning weights, as well as for further information related to smoothing, we refer readers to the literature on general and spatial statistics and on the mathematical foundations of GIS, for example Burt and Barber (1996), Cressie (1991), and Fotheringham and Rogerson (1994).

Neighbourhood-based operations in space are not limited to smoothing. There are, in particular, spatial analogues of the process of computing changes with time. In such a computation, the values of a spatially referenced numeric attribute are treated as altitude values on a continuous surface. The rate of change in such a surface could be estimated as the difference between the altitudes of two neighbouring locations. However, since space, unlike time, is not a linearly ordered set, it is possible to choose a neighbouring location in any direction from a target location. For each direction, the altitude difference will be different. Therefore, the rate of change in space is computed using the plane tangential to the surface at the target location. The orientation of this plane is referred to as the *slope*. It may be characterised by two components, vertical and horizontal, known as the *gradient* and the *aspect*. The gradient of the slope describes the rate of change as a function of the angle of the tangent plane with respect to the horizontal plane. The aspect is the direction of the slope, measured as an angle from some arbitrary bearing such as north or east. The notions of the gradient and aspect are illustrated graphically in Fig. 4.51.



**Fig. 4.51.** For computing changes in space, the values of a spatially referenced numeric attribute are treated as measures of the altitude of a continuous surface  $S$ . The change at a target location  $L$  is estimated using the plane  $P$  tangential to the surface  $S$  at the point  $L$  (i.e. it touches the surface at this point). The orientation of the plane is characterised by the gradient  $G$  and the aspect  $A$

The view of a spatially referenced numeric attribute as a surface may be exploited further in an attempt to reveal topological, or morphometric, fea-

tures on this surface, such as peaks, pits, ridges, channels, and passes. Interested readers are referred to the general discussion of the notions related to surface topology in Chrisman (1997) and to examples of the use of the extraction of topological features in the exploratory analysis of spatio-temporal data described in Sadahiro (2002) and Rana and Dykes (2003). Chrisman also proposes a taxonomy of neighbourhood-based transformations of spatial data (Chrisman 1997, pp. 218–231).

Since the topic of neighbourhood-based transformations of spatially and temporally referenced data is covered quite well in the literature on (spatial) statistics and GIS, we believe that our brief and superficial discussion should suffice in the context of this book. In addition to what has already been said, we would like to mention that neighbourhood-based transformations can be performed not only on numeric attributes but also on attributes with qualitative values. For example, the “majority filter” replaces the value of a qualitative attribute associated with a target reference by the most frequent value (mode) in the neighbourhood of this reference. It is also possible to compute various measures of the diversity of the values in the neighbourhood, the simplest measure being the number of different values.

At this point, we would like to close this subsection, in which we have considered several types of attribute transformation that are frequently used in exploratory data analysis. We do not claim that these are the only possible or the only useful transformations. Of course, there are countless possibilities for producing derived attributes, and there may be countless cases requiring transformations different from the ones we have described. Therefore, a data analyst needs tools that are sufficiently powerful and flexible for performing various transformations, depending on the meaning and properties of the data under analysis and the goals of the analysis.

## 4.5.2 Attribute Integration

It may seem very tempting to simplify data analysis work by integrating several attributes into one and considering this single attribute instead of the original multiple attributes. For example, for studying the variation of the age structure of the population over the territory of Portugal, it would be much more convenient to have a single attribute that somehow reflected the age structure than the actual four attributes of the proportions of people in different age groups. However, despite the attractiveness of this idea, it is very difficult to realise in practice. The problem is the impossibility of combining several attributes into one without significant information loss.

Nevertheless, there are situations where attribute integration is necessary or strongly desirable. We are aware of two major contexts requiring attribute integration:

1. *Evaluation, comparison, and ranking.* These tasks play an important role in management. Thus, a company or a government needs to evaluate its accomplishments and compare them with past states, with what has been planned, with standards and legal requirements, with the performance of others, etc. in order to determine whether there has been progress or decline (and how much progress or decline) and to estimate the position of the company, organisation, or locality relative to others. For this purpose, it is often necessary to combine a number of performance indicators into a single score that can be used for comparing and ranking. When multiple objects are being managed, such as different departments, retail stores, or administrative districts, they need to be evaluated and compared in order to identify good and poor performers.
2. *Influence analysis*, i.e. an investigation of whether and how a certain group of characteristics is related to some other characteristic or group of characteristics. For example, one may be interested in whether children's diet affects their physical and intellectual development. A diet may be characterised by a number of attributes reflecting the amounts of various products consumed. In order to make the problem manageable, the analyst may need to replace this set of attributes by a single attribute expressing the degree of healthiness of the diet. Another example could be an analysis of the growth of young trees in a forest, which is influenced by the number of older trees around, their heights and crown diameters, and the density of leaves. The role of all of these factors is that they determine the amount of sunlight that a young plant receives, and it is this amount that ultimately influences the growth of the plant. Therefore, researchers use the original set of characteristics to measure the illumination of young plants, and then study the dependency between the illumination and the growth of the plants.

The methods for integrating multiple attributes are often problem-specific and vary from case to case. Therefore, we do not intend to consider the topic of attribute integration in much detail. Instead, we are going to cite an example of the computation of problem-specific integrated index and then present a possible generic method for attribute integration as well as an interactive tool that realises this method. In fact, the main purpose of describing this tool is to introduce the idea of a dynamic attribute, the values of which change when the tool user modifies the parameters involved in the integration procedure.

#### 4.5.2.1 An Example of Integration

Once, in our practical work on data analysis, we explored a dataset characterising the administrative districts of a certain part of England. The dataset contained a number of census attributes, as well as four different indices of material deprivation. These indices, which are described at <http://www.swpho.org.uk/pat18discuss.htm>, are used in the UK to reflect the degree of poverty of the population in various areas. Each index is calculated in its own way by integration of specific attributes.

Our task was to discover links between the deprivation indices and the census attributes by applying various visual and interactive tools for exploratory data analysis.<sup>10</sup> Before doing the exploration, we had to learn how the deprivation indices were constructed and which of the census attributes were involved in this in order to avoid the discovery of self-evident links. Here, we shall give an example of attribute integration by describing briefly how one of the four deprivation indices, the Townsend Score, is computed. It is based on four census attributes:

- *Unemployment*: The percentage of economically active residents aged 16–59 or 64 who are unemployed.
- *No car*: The percentage of private households who do not possess a car.
- *Home ownership*: The percentage of private households not owner-occupied.
- *Overcrowding*: The percentage of private households with more than one person per room.

From these four attributes, the integrated score is computed in the following way. Two of the attributes, specifically unemployment and overcrowding, are first transformed using the logarithmic transformation  $y = \ln(x + 1)$  to produce more normal distributions. Then, the values of all attributes are converted to standard scores (*z*-scores).<sup>11</sup> Scores greater than zero indicate greater levels of material deprivation. The overall deprivation index is computed as the arithmetic sum of the four scores.

It is quite typical that attribute values are *standardised* prior to attribute integration. The purpose of this is to make the values of diverse attributes comparable. The methods used for the standardisation may vary. Thus,

---

<sup>10</sup> One of the results obtained was the detection of a link between material deprivation and the concentration of certain national minorities. It is interesting that national minorities of different origin are differently positioned with respect to poverty. More information can be found in Andrienko and Andrienko (2004).

<sup>11</sup> As a reminder, *z*-scores are computed according the formula (4.5) and show deviations from the mean.

besides the z-scores, the chi-square ( $\chi^2$ ) method is frequently applied. This method is based on raw values, i.e. the actual numbers rather than the proportions. It compares the observed value ( $O$ ) in an area with a certain expected value  $E$ . For example, in computing the deprivation scores for the districts of England,  $E$  may be the respective average rate for England as a whole. The computation is done according to the formula

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \quad (4.6)$$

where  $O_1$  is the observed value with the characteristic (e.g. unemployed),  $E_1$  is the expected value with that characteristic,  $O_2$  is the observed value without the characteristic (e.g. not unemployed), and  $E_2$  is the expected value without the characteristic.

Comparability of values of diverse attributes may also be achieved by means of other transformation methods, which may not necessarily be based on comparison with standard values. For example, the original attribute values may be converted into their relative positions between the minimum and maximum values of the attribute. This method is used when there is a specific need to compare actual values with either the smallest or the greatest possible value. Thus, in evaluating several options to choose the optimal one, high or low attribute values may be desirable or undesirable for a decision maker, and hence the distance to those values may be of interest.

In computing the Townsend index, all transformed attributes have equal influence on the resulting scores. There are also cases where the attributes to be combined have different importance. In order to take account of this different importance, the attributes are assigned different *weights*. To obtain the overall score, the values of the attributes are multiplied by the weights and summed. To compute such weighted sums, one may use an interactive tool such as that presented below. The tool allows the user to set and change the weights of the source attributes, and dynamically recomputes the values of the resulting integrated attribute after any change of the weights. This resulting attribute is an example of a *dynamic attribute*, i.e. an attribute whose values may change.

#### **4.5.2.2 Dynamic Integration of Attributes**

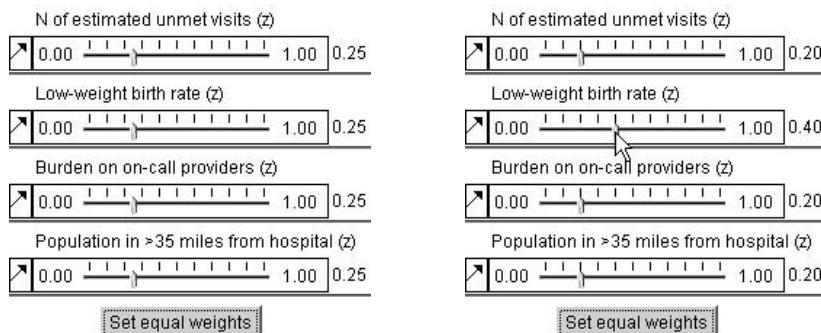
To demonstrate the operation of the tool, we shall try to evaluate the situation with regard to health care in various counties of the state of Idaho in the USA and determine which of them are most in need of support to improve the availability and accessibility of health care facilities for the

population. In our evaluation, we must combine multiple attributes characterising the situation, specifically the following:

- *N of estimated unmet visits*: The estimated number of unmet visits to the doctor (when people coming to see a doctor cannot be attended to because the doctor is overloaded).
- *Low-weight birth rate*: The percentage of infants born with insufficient body weight, averaged over a multiyear interval.
- *Burden on on-call providers*: The number of hours on call for each provider.
- *Population in >35 miles from hospital*: The number of individuals residing outside the influence zone (i.e. a radius of 35 miles, according to the national standard for rural areas) of the nearest hospital.

Before applying the tool, we converted the original values of all attributes into  $z$ -scores, which express the relative deviations from the respective means. Positive deviations signify that the original values are worse than the average value for the state of Idaho.

The attribute combination tool provides a direct manipulation interface for choosing the weights of the attributes used in the computation. The interface is shown in Fig. 4.52.



**Fig. 4.52.** A user interface for setting attribute weights for computing weighted linear combinations (weighted sums) of values of multiple attributes. Initially, all weights are equal (left). To change the weights, the user moves the sliders (right)

For each attribute, there is a ruler with a slider. The position of the slider corresponds to the current weight of the attribute, which must be a real number between 0 and 1. The sum of the weights of all attributes participating in the computation must equal 1. Immediately after activation of the tool, all attributes are assigned equal weights. In our example, we have selected four attributes; accordingly, each of them was assigned a weight

of 0.25. In order to change the default weights, the user moves the sliders along the rulers. When one of the sliders is moved, and hence the weight of the corresponding attribute changes, the tool automatically adjusts the weights of the remaining attributes so that the sum of the weights remains equal to 1. The changes are proportional to the values that the weights had before the slider was moved. Thus, on the right in Fig. 4.52, the slider corresponding to the attribute “Low-weight birth rate” has been moved to the right so that the weight of this attribute has become 0.4. The weights of the remaining three attributes have been automatically set to 0.2.

We have described the user interface of this tool in order to demonstrate the ease with which attribute weights can be changed. However, the important point is not this ease itself but its combination with the high reactivity of the tool: the weighted sums are dynamically recomputed as the user moves the slider. We have already mentioned that the tool produces a *dynamic attribute*. The values of this attribute, i.e. the weighted sums, change when the user modifies the computational parameters, i.e. the weights. A dynamic attribute may be visualised like the usual kinds of attributes; it is necessary only that the visualisation tools update the display when the values of the attribute change.

As a result, the user receives an excellent opportunity to investigate how altering the weights affects the computed scores. Such an investigation may be highly desirable when the scores are used as a basis for decision-making, for example to decide which of the counties should get financial support to improve their health care. A well-substantiated decision must be based on sufficiently robust evaluation results, i.e. results that are not affected by minor changes of the weights.

Let us illustrate the ideas of dynamic attributes and sensitivity analysis with an example evaluation of health care in the counties of Idaho. To reflect our initial understanding of the relative importance of the four attributes listed above, we assigned a weight of 0.3 to the first two attributes and a weight of 0.2 to the remaining two attributes. The table display shown in Fig. 4.53 represents the results of the computation (in the column headed “Evaluation score”) along with the source attributes. The rows of the table are arranged in order of decreasing evaluation score. To save space, only a fragment of the table representing the top 10 of the 44 counties of Idaho is shown. The rows with the three topmost scores are highlighted. They correspond to the counties of Washington, Payette, and Jerome. Hence, these counties have the most critical situation concerning health care, with respect to the weights assigned to the four evaluation attributes.

	N of estimated unmet visits (z)	Low-weight birth rate (z)	Burden on on-call providers (z)	Population in >35 miles from hospital (z)	Evaluation score	
<b>Washington</b>	<b>0.556</b>	<b>1.997</b>	<b>2.071</b>	<b>0.556</b>	<b>0.7350</b>	▲
<b>Payette</b>	<b>0.650</b>	<b>1.099</b>	<b>2.071</b>	<b>0.570</b>	<b>0.6841</b>	
<b>Jerome</b>	<b>0.678</b>	<b>0.873</b>	<b>2.071</b>	<b>-0.182</b>	<b>0.6430</b>	
Latah	1.172	1.617	-0.293	-0.182	0.5926	
Madison	0.912	-0.138	-0.341	3.246	0.5919	
Clearwater	0.540	-1.934	1.008	4.926	0.5913	
Gooding	-1.398	2.774	1.157	-0.182	0.5849	
Twin_Falls	1.198	1.010	-0.065	-0.182	0.5679	
Clark	0.632	-0.356	2.071	-0.411	0.5531	
Gem	-1.079	1.124	2.071	-0.182	0.5494	
...	...	...	...	...	...	

**Fig. 4.53.** This table display represents the results of evaluating the counties of Idaho on the basis of four attributes with weights of 0.3, 0.3, 0.2, and 0.2. The columns of the table show the values of the four source attributes (transformed into z-scores) and the resulting evaluation scores. The rows of the table are arranged in order of decreasing evaluation score. The rows with the three topmost scores are highlighted

	N of estimated unmet visits (z)	Low-weight birth rate (z)	Burden on on-call providers (z)	Population in >35 miles from hospital (z)	Evaluation score	
<b>Washington</b>	<b>0.556</b>	<b>1.997</b>	<b>2.071</b>	<b>0.556</b>	<b>0.7050</b>	▲
<b>Payette</b>	<b>0.650</b>	<b>1.099</b>	<b>2.071</b>	<b>0.570</b>	<b>0.6571</b>	
Clearwater	0.540	-1.934	1.008	4.926	0.6134	
<b>Jerome</b>	<b>0.678</b>	<b>0.873</b>	<b>2.071</b>	<b>-0.182</b>	<b>0.6105</b>	
Madison	0.912	-0.138	-0.341	3.246	0.5969	
Latah	1.172	1.617	-0.293	-0.182	0.5629	
Gooding	-1.398	2.774	1.157	-0.182	0.5556	
...	...	...	...	...	...	

**Fig. 4.54.** The result of changing the attribute weights from 0.3, 0.3, 0.2, and 0.2 to 0.28, 0.28, 0.19, and 0.25. The table rows are ordered as before, according to the computed evaluation scores. The same rows as in Fig. 4.53 are highlighted. It can be clearly seen that, after the weights have been changed, the county of Jerome has moved from the third to the fourth position

After some additional deliberation, we decided that the attribute “Population in >35 miles from hospital” should have a higher influence on the evaluation results, i.e. its weight should be increased. So, we increased its weight from 0.2 to 0.25. The weights of the other attributes were automatically adjusted: the first two attributes received weights of 0.28 each, and the weight of the third attribute was reduced to 0.19. The evaluation scores were immediately recomputed, and the table display reflected the changes,

as is shown in Fig. 4.54. Not only have the numbers in the column “Evaluation score” changed, but also the order of the rows, since the table display tool sorts the rows according to the evaluation scores, as before. It is easy to see, in particular, that the county of Jerome, which originally occupied the third position in the table, has moved to the fourth position. The third topmost score now belongs to the county of Clearwater, where the number of people living too far from any hospital is much higher than the average over the state (the  $z$ -score is 4.926).

To investigate the sensitivity of this evaluation result to minor changes of the weights, we increased the weight of the attribute “Population in >35 miles from hospital” by 0.01. To keep the sum of the weights equal to 1, the weight of the attribute “N of estimated unmet visits” was automatically decreased by the same amount. This rather slight change produced quite a noticeable effect, as may be seen from Fig. 4.55. The county of Jerome has moved from the fourth to the fifth position, and the county of Madison has ascended to the fourth position. This shows that our evaluation is quite sensitive to changes in the weights. This is an undesirable feature: if, supposedly, our limited resources were to allow us to provide support to no more than four counties, it would be unclear which of the counties, Jerome to Madison, should be preferred.

	N of estimated unmet visits ( $z$ )	Low-weight birth rate ( $z$ )	Burden on on-call providers ( $z$ )	Population in >35 miles from hospital ( $z$ )	Evaluation score
<b>Washington</b>	<b>0.556</b>	<b>1.997</b>	<b>2.071</b>	<b>0.556</b>	<b>0.6924</b> ▲
<b>Payette</b>	<b>0.650</b>	<b>1.099</b>	<b>2.071</b>	<b>0.570</b>	<b>0.6457</b>
Clearwater	0.540	-1.934	1.008	4.926	0.6227
Madison	0.912	-0.138	-0.341	3.246	0.5991
<b>Jerome</b>	<b>0.678</b>	<b>0.873</b>	<b>2.071</b>	<b>-0.182</b>	<b>0.5969</b>
Latah	1.172	1.817	-0.293	-0.182	0.5603
Gooding	-1.398	2.774	1.157	-0.182	0.5433
...	...	...	...	...	...

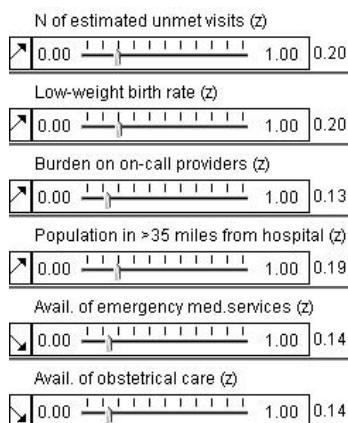
**Fig. 4.55.** To probe the sensitivity of the evaluation, the weights were changed from 0.28, 0.28, 0.19, and 0.25 to 0.27, 0.28, 0.19, and 0.26. As a result, Jerome has moved from the fourth to the fifth position

A possible way of solving this problem is to involve additional attributes in the evaluation. Let us use the following two attributes (also transformed into  $z$ -scores):

- *Avail. of emergency med. services:* The availability of emergency medical services, calculated by dividing the total number of ambulances and quick-response units by the population in each county.

- *Avail. of obstetrical care:* The availability of obstetric care, expressed by the number of providers offering obstetric delivery services.

These attributes (referred to as the availability attributes from now on) differ from the previous four attributes in that higher values of them indicate better health care conditions than do lower values. Since our goal is to see where the situation is the worst, the values of the availability attributes must be reversed. Our tool for computing weighted sums will do this if we indicate that these attributes have the opposite orientation. For this purpose, the user interface of the tool contains arrow-shaped “switchers” positioned to the left of the slider bars used for setting the attribute weights (see Fig. 4.56). Clicking on a switcher reverses the orientation of the corresponding attribute. The arrow also changes its orientation from north-east to south-east. Hence, the appearance of the controls shows whether the original or reversed values of each attribute are used in the computation. Thus, it is clear from the screenshot in Fig. 4.56 that the availability attributes have an orientation opposite to that of the other four attributes. This means that negative values increase the integrated score and positive values decrease it.



**Fig. 4.56.** Two additional attributes have now been included in the evaluation: the availability of emergency medical services and the availability of obstetric care. Their opposite orientation in comparison with the other four attributes is indicated by the downward-oriented arrows to the left of the slider bars used for setting the weights

The table in Fig. 4.57 shows the result of evaluating the counties of Idaho on the basis of the six attributes, i.e. the initial set of four attributes plus the availability attributes. The scores contained in the rightmost column of the table correspond to the attribute weights shown in Fig. 4.56, i.e.

0.2, 0.2, 0.13, 0.19, 0.14, and 0.14. As before, the rows of the table are arranged in order of decreasing evaluation score. The topmost four rows are highlighted. These rows correspond to the counties of Washington, Payette, Jerome, and Madison.

	N of estimated unmet visits (z)	Low-weight birth rate (z)	Burden on on-call providers (z)	Population in >35 miles from hospital (z)	Avail. of emergency med.services (z)	Avail. of obstetrical care (z)	Evaluation score
<b>Washington</b>	<b>0.556</b>	<b>1.997</b>	<b>2.071</b>	<b>0.556</b>	<b>0.131</b>	<b>-0.892</b>	<b>0.7294</b> ▲
<b>Payette</b>	<b>0.650</b>	<b>1.099</b>	<b>2.071</b>	<b>0.570</b>	<b>-0.532</b>	<b>-0.892</b>	<b>0.7176</b>
<b>Jerome</b>	<b>0.678</b>	<b>0.873</b>	<b>2.071</b>	<b>-0.182</b>	<b>-0.311</b>	<b>-0.892</b>	<b>0.6742</b>
<b>Madison</b>	<b>0.912</b>	<b>-0.138</b>	<b>-0.341</b>	<b>3.246</b>	<b>-0.311</b>	<b>-0.558</b>	<b>0.6728</b>
Latah	1.172	1.617	-0.293	-0.182	-0.975	0.027	0.6428
Clark	0.632	-0.356	2.071	-0.411	-0.754	-0.892	0.6260
Twin_Falls	1.198	1.010	0.065	-0.182	-0.754	-0.140	0.6226
Jefferson	1.355	-0.542	0.493	-0.411	-0.975	-0.558	0.5919
Clearwater	0.540	-1.934	1.008	4.926	2.783	-0.725	0.5875
Power	0.358	1.997	-0.766	-0.182	0.131	-0.140	0.5762
...	...	...	...	...	...	...	...

**Fig. 4.57.** The counties have been evaluated here on the basis of six attributes using weights of 0.2, 0.2, 0.13, 0.19, 0.14, and 0.14. The rows with the four topmost evaluation scores are highlighted

Again, we probed the sensitivity of this selection by slightly varying the weights of the attributes. This time, the four highlighted counties remained stably at the top of the list; only their relative positions changed from time to time. This looks good, but we realised soon that manual variation of the weights requires much time, and it is hard to ensure that all admissible weight combinations have been tested. Fortunately, the tool can help us in performing the sensitivity analysis: we can specify the limits for the weight of each attribute, and the tool will compute the integrated scores, taking various weights from the specified intervals. The results of this automated sensitivity analysis are summarised by providing the minimum, maximum, and mean position of each county and the variance of the position.

To test the robustness of our choice of four counties, we specified the parameters for the automated sensitivity analysis as is shown in Fig. 4.58. Besides the minimum and the maximum admissible weight for each attribute, it is possible to choose how many different values from this interval will be tested. In our case, we specified that the tool must test 20 values for each attribute weight.

The results of the automated sensitivity analysis are presented in the table in Fig. 4.59. This table shows the minimum position that each county received during the tests, the maximum position, the mean position, and the variance of the positions received in all test runs. The rows of the table are arranged in order of increasing mean position. It may be seen that the

four counties chosen as the top candidates for receiving funding appear at the top of the table, i.e. their mean positions are the highest. Even the relative order of these counties remains the same as in Fig. 4.57. From the ranges of the positions and the variances, it may be seen that Washington and Payette are indubitably in need of support: Washington always ranked first or (rarely) second, and Payette occupies positions from first to third, being second on average. The rank variances of these two counties are the smallest of all.

N of estimated unmet visits (z)	0.20		
<input type="text" value="20"/> iterations between weights	<input type="text" value="0.15"/>	and	<input type="text" value="0.30"/>
Low-weight birth rate (z)	0.20		
<input type="text" value="20"/> iterations between weights	<input type="text" value="0.15"/>	and	<input type="text" value="0.30"/>
Burden on on-call providers (z)	0.13		
<input type="text" value="20"/> iterations between weights	<input type="text" value="0.12"/>	and	<input type="text" value="0.25"/>
Population in >35 miles from hospital (z)	0.19		
<input type="text" value="20"/> iterations between weights	<input type="text" value="0.14"/>	and	<input type="text" value="0.29"/>
Avail. of emergency med. services (z)	0.14		
<input type="text" value="20"/> iterations between weights	<input type="text" value="0.12"/>	and	<input type="text" value="0.28"/>
Avail. of obstetrical care (z)	0.14		
<input type="text" value="20"/> iterations between weights	<input type="text" value="0.12"/>	and	<input type="text" value="0.28"/>
<input type="button" value="OK"/>		<input type="button" value="Cancel"/>	

**Fig. 4.58.** To run an the automated sensitivity analysis, the user specifies the minimum and maximum admissible weight for each attribute and the number of different values from this interval to be tested

	Minimum position	Maximum position	Mean position	Variance of positions
<b>Washington</b>	<b>1</b>	<b>2</b>	<b>1.1</b>	<b>0.301</b> ▲
<b>Payette</b>	<b>1</b>	<b>3</b>	<b>1.95</b>	<b>0.339</b>
<b>Jerome</b>	<b>3</b>	<b>5</b>	<b>3.2333333</b>	<b>0.514</b>
<b>Madison</b>	<b>1</b>	<b>6</b>	<b>3.9333334</b>	<b>0.764</b>
Latah	4	8	5.275	0.698
Clark	4	9	6.0	0.961
Twin_Falls	6	9	7.008333	0.667
Jefferson	8	12	8.716666	1.047
Clearwater	4	28	10.666667	4.801
Blaine	8	14	10.958333	1.292
...	...	...	...	...

**Fig. 4.59.** The results of an automated sensitivity test with the parameters specified in Fig. 4.58

The county of Jerome is, on the average, in third place. Its behaviour in the test was rather stable: it never ascended to a position higher than third but also never dropped below fifth. Its rank variance is also rather low, which indicates that the county was mostly among the top four throughout all test runs. Hence, it would be quite safe to conclude that this county must also receive financial support.

The county of Madison fluctuated between the first and the sixth position. It has the highest variance among the four counties. However, if the hypothetical funding has to be distributed among four counties, Madison is obviously a better candidate for receiving support than the counties positioned below it in the table. None of these counties ascended to a position higher than fourth, and the mean position of Madison differs quite substantially from the mean position of the next county in the list.

This example evaluation should not be considered as an introduction to the methods for supporting multicriteria decision-making but rather as an illustration of the ideas of dynamic attributes and sensitivity analysis. It happens quite often that methods applied for data transformation involve parameters that can be assigned more or less subjectively chosen values. It is highly desirable that the tools realising such methods allow users to understand how their choice affects the results obtained. One possible way is that the tool produces dynamic attributes with values that change as the user modifies the parameters of the method. In order to observe the changes, the user needs to visualise these dynamic attributes, which, in turn, requires the visualisation tool to be sensitive to the changes in the attribute values. The table display that we used in the course of our evaluation of the counties of Idaho is an example of such a visualisation tool. The display is updated in response to any change in the values of a dynamic attribute. Not only do the new values replace the old ones in the corresponding table column, but also the order of the table rows may change if the dynamic attribute is used for sorting the rows.

Hence, an interactive sensitivity analysis involves the following prerequisites:

- The user interface of the tool permits easy changing of the data transformation parameters.
- In response to changes in the parameters, the tool repeats the computations using the new parameter values and replaces the previously computed results with the new results.
- A tool must be used to visualise the results of the computation. The visualisation tool must be sensitive to changes in the results of the computation and must respond to the changes by updating the display.

Despite the importance of providing facilities for an interactive sensitivity analysis, this may still be insufficient. When the method used for data transformation involves multiple parameters, the procedure of interactive sensitivity testing may be too time-consuming and tiresome for the user. Therefore, it is good to supplement interactive facilities with tools for automated sensitivity testing.

### 4.5.3 Value Interpolation

The word “interpolate” is defined as “1) to introduce (something additional or extraneous) between other things or parts; 2) to insert, estimate, or find an intermediate term in (a mathematical sequence); …” (Random House 1996). Data *interpolation* may be described, in a broad sense and using the terminology of our data model, as inserting additional elements into the set of references of a dataset. For the new references, one needs to specify the corresponding values of the attributes of the dataset. These attribute values are initially unknown; they need to be estimated on the basis of known attribute values corresponding to the other references.

Practically, interpolation is done for attributes with continuous value sets and reference sets with distances between elements. The continuity of an attribute ensures the existence of intermediate values between any two different values. The existence of distances in the reference set is necessary for the notion of neighbourhood to be defined. An unknown attribute value corresponding to a reference is derived from known attribute values corresponding to its neighbouring references. This operation involves, explicitly or implicitly, certain assumptions concerning the nature of the phenomenon represented by the attribute. The minimum assumption is that the phenomenon is continuous and smooth, i.e. attribute values corresponding to close references are also close. Hence, the estimated attribute value for a new reference must be close to the actual attribute values corresponding to the neighbours of the new reference.

It is clear from this explanation that data interpolation has two aspects: first, how to define the appropriate neighbours for any new reference; and second, how to make appropriate estimates of the corresponding attribute values on the basis of the available attribute values. There are methods that work “locally”, i.e. they derive the estimated values for new references directly from the values associated with their neighbours, without involving any other references and their corresponding attribute values. Other methods use all of the original references available in the dataset and the corresponding attribute values to build an overall model (such as a mathematical equation or a system of equations), which may then be used to de-

termine the attribute value for any reference. Naturally, it is required that the values produced with the use of this model for the original references coincide with the corresponding original attribute values.

The first approach is, obviously, simpler and less computationally intensive. However, analysts are often not quite satisfied with the results obtained: these results are typically not sufficiently smooth, which is inconsistent with the assumption of the smoothness of the underlying phenomenon. The “global” methods usually work better but are more complex and resource-demanding.

We are not going to immerse ourselves deeply into the topic of data interpolation, which is quite well covered in the mathematical literature (more specifically, that on numerical analysis). Interpolation of spatial data is discussed in the literature on cartography and GIS; see, for example, Slocum (1999) and Chrisman (1997). In the context of our study, we shall briefly consider the two most common cases of interpolation: interpolation in a linearly ordered reference set, such as time, and interpolation in a two-dimensional space. We shall also enumerate some of the most popular interpolation methods. We believe this to be quite an appropriate level of detail for our general review of tools for exploratory data analysis.

As we have mentioned, data interpolation involves the problem of defining appropriate neighbours for a given reference. This problem is common to interpolation and to the neighbourhood-based attribute transformations considered earlier. Which of the pre-existing references can be taken as neighbours of a given new reference depends first of all on the properties of the reference set. If this is a linearly ordered set, such as time, for example, each new reference has two nearest neighbours, specifically, the pre-existing references preceding and following this reference in the order. In two-dimensional space, there is no ordering between elements. The neighbourhood relation may be defined by specifying a certain distance (radius). The references (i.e. spatial locations) lying within this distance from the target reference are considered to be the neighbours of this reference.

In the literature concerning the interpolation of spatial data, two different cases are considered individually: interpolation from scattered locations and interpolation in regular grids. In the latter case, the task of finding neighbours is easy: for each location, these are the nearest grid nodes. However, when the reference set consists of scattered locations, finding the neighbouring locations may require quite wasteful searching. There are several approaches to optimising this search process; they are described in the relevant literature.

Sometimes, when scattered locations are being dealt with, a limit is imposed on the number of neighbours to be taken into account. If the actual

number of neighbours exceeds this limit, the most distant neighbours are discarded. In some approaches, the number of neighbours involved in the interpolation is limited from the other side, i.e. by specifying the minimum number. In the case where there are not enough neighbours within the specified radius from the target reference, the radius is increased stepwise until the required number of neighbours is obtained.

In order to take into account the neighbours lying in different directions from a target location, a circle around this location may be divided into a specified number of sectors (typically four or eight; the respective search methods are called quadrant and octant strategies). Then, a specified number of neighbours are taken from each sector. One may also specify the minimum number of neighbours required and the maximum number of empty sectors permitted. When these criteria are not met, the initial radius of the circle needs to be increased.

Another approach to finding appropriate neighbours in a set of scattered locations is based on triangulation – connecting the original locations by lines so that the plane is divided into non-overlapping triangles. Then, any new location will fall into one specific triangle. One of the vertices of this triangle will be the closest neighbour of the new location, and the other appropriate neighbours can be found by inspecting the neighbouring triangles.

Once the neighbours have been found, the next step is to use the corresponding attribute values for the estimation of a value for the new reference. The simplest method is known as linear interpolation: the new value is found as a weighted average of the original values, i.e. according to the formula

$$\frac{\sum_{i=1}^N v_i \cdot w_i}{\sum_{i=1}^N w_i} \quad (4.7)$$

Here,  $N$  is the number of neighbours,  $v_i$  is the attribute value corresponding to the  $i$ th neighbour, and  $w_i$  is the weight, which is inversely proportional to the distance between the  $i$ th neighbour and the target location. Note that the expression (4.7) represents a linear function, which gives this interpolation method its name.

Linear interpolation can be used in a linearly ordered reference set (in this case, there are exactly two neighbours for each new reference), in two-dimensional space, and, in general, in any reference set with distances. In many situations this simple approach is quite appropriate, but there are

also many situations where the results of such interpolation are not sufficiently smooth, and then analysts opt for other interpolation methods.

For reference sets with linear ordering, two popular interpolation methods are polynomial and spline interpolation. Polynomial interpolation is a generalisation of linear interpolation by means of replacing the linear function by a polynomial of higher degree. Whereas linear interpolation uses, in the case of a linearly ordered reference set, only the two nearest neighbours of each new reference, a polynomial function of degree  $n$  requires  $n + 1$  values corresponding to consecutive references. Usually, a single polynomial is constructed using all original references and then used for deriving attribute values for intermediate references. Polynomial interpolation produces smoother results than does linear interpolation but is more computationally intensive. Besides, the results may be inexact, especially at the end point.

Spline interpolation can be characterised as piecewise polynomial interpolation: it uses low-degree (e.g. cubic) polynomials in each of the intervals between two successive references. The polynomial pieces are chosen so as to fit smoothly together. The resulting function is called a *spline*. This method produces quite smooth interpolation and is more precise than using higher-degree polynomials. Other interpolation methods are also applied to linearly ordered reference sets, for instance trigonometric interpolation, which uses trigonometric polynomials. A detailed review of all interpolation methods is, however, beyond the scope of our study.

It should not be thought that the interpolation methods mentioned above are applicable only to numeric attributes. A counter-example might be the interpolation of spatial positions of a moving object to construct a smooth trajectory line on a map, in a 3D view, or for the purposes of movement animation. In this example, there is a linearly ordered reference set, specifically, time. The positions of the object are initially known for sample time moments, and the task is to estimate its positions at intermediate moments.

Since the polynomial, spline, and similar interpolation methods produce smooth curves, these methods are also used in the interpolation of spatially referenced data to represent spatially continuous phenomena in the form of *isopleths* (equal-value lines, sometimes also called *isolines*). The idea of polynomial interpolation, both global and piecewise, is also extendable to spatially referenced data in another way. Similarly to the construction of polynomial curves or curve segments in the one-dimensional case, polynomial surfaces can be constructed from values specified at sample locations in a (two-dimensional) space. However, the most respected interpolation method for spatially referenced data is the method of *kriging*, which uses information about the spatial autocorrelation in the vicinity of each

location. The spatial autocorrelation is an assessment of the correlation between attribute values and the spatial locations that they refer to, or, simply speaking, how similar the values in neighbouring locations are. On this basis, kriging is believed to provide “optimal” interpolation in the sense of greater use of the information provided by the spatial arrangement. The kriging method is mathematically quite complex and will not be considered here in detail.

A concept related to interpolation is that of *extrapolation*, which means estimation of attribute values for references that are not between original references but are outside the given reference set. The results of extrapolation are often subject to substantial uncertainty.

At the end of our brief discussion concerning interpolation, let us demonstrate the effect of applying interpolation in the visualisation of spatially referenced data specified in a raster format. Let us recall that the raster model for spatially referenced data divides a territory into fine grid cells, called pixels, which are filled with attribute values. Data specified in a raster format are visualised by encoding the values of the raster cells by colours of the corresponding screen pixels. For example, Fig. 4.60C demonstrates a possible visualisation of the relief of Europe specified in a raster format, with the raster pixels containing altitudes.<sup>12</sup>

When a raster pixel is mapped onto more than one screen pixel, it is assumed that the value in the raster pixel corresponds to the centre of the area on the screen. The values corresponding to all other screen pixels are determined by interpolating values from the surrounding cells. Owing to the interpolation, the resulting image looks smooth. The effect of interpolation is demonstrated on the right in Fig. 4.60C. The image fragment at the top right has been produced without interpolation. Its appearance resembles a mosaic made of rectangular tiles. At the bottom right, the same data are shown with linear interpolation used. The resulting image contains no abrupt changes of colours, and no boundaries of raster cells are visible.

We would also like to mention another common case where interpolation is used, in the visualisation of time-referenced data by means of display animation. Animation may be thought of as the presentation of a se-

---

<sup>12</sup> In this example, a specific colour encoding is used: shades of blue represent values below zero, values from 0 to 200 are encoded by shades of green, and shades from light yellow to dark orange are used for values over 200, with darker shades corresponding to higher altitudes. This encoding is close to the traditional representation of relief in physical maps, which makes the image easily understandable. However, such a colourful representation is not highly recommended for arbitrary data, where a simple scale of increasing or decreasing brightness or a diverging colour scale may be more appropriate.

quence of images, or frames, one after another. This sequence consists of key frames and intermediate frames. Key frames are constructed from data originally available in the dataset; these data refer to a certain finite set of time moments. If only key frames are used, the resulting visualisation may be difficult to perceive owing to the lack of temporal continuity: abrupt local changes from one frame to another may obstruct one from seeing the general trend. Therefore, intermediate frames are constructed by interpolating the data to time moments in between the original moments. The more intermediate frames are used, the smoother the resulting animation appears. If the original set of time moments is not regularly spaced, it is reasonable to vary the number of intermediate frames depending on the length of the time interval between two consecutive key frames.

#### 4.5.4 Data Aggregation

Data aggregation tools reduce the amount of data under analysis by grouping individual references into subsets, which will be called “aggregates”, and computing some collective characteristics of the aggregates. Aggregates and their characteristics (jointly called “aggregated data”) are often explored instead of the original data, especially when one is dealing with very large datasets. Substitution of the original data by aggregated data facilitates the process of simplification and abstraction in the course of perceiving and characterising behaviours; hence, aggregation tools are appropriate for synoptic tasks. However, the simplification is achieved at the cost of information loss, specifically, discarding characteristics of individual references. Hence, aggregated data are not appropriate for elementary tasks.

The tools and techniques for data aggregation are rather numerous, which reflects their important role in the exploration of large amounts of data. In an attempt to provide some kind of systematic view of the variety of aggregation tools, we have extracted the following three aspects that characterise and differentiate existing approaches to data aggregation:

1. How individual references are grouped into aggregates (or, in other words, how the entire reference set is divided into subsets).
2. How the aggregates are characterised, i.e. what kind of characteristics are used and how they are derived.
3. How the aggregated data are visualised or, in a more general sense, presented to the analyst.

#### 4.5.4.1 Grouping Methods

Grouping of individual references into aggregates is done on the basis of relations between them (which includes the general relations of ordering and distances, as well as particular relations pertinent to specific reference sets) or on the basis of commonality or closeness of their characteristics, i.e. the attribute values corresponding to the references. In both cases, the grouping method depends on the number of referrers or attributes used as the basis for the grouping, and on the properties of the value sets of the referrers and attributes. Let us consider first how grouping on the basis of a single referrer or attribute is done.

For an attribute or referrer with a nominal value set, i.e. without ordering or distances, two possibilities exist. If the number of different values is not very large, references may be grouped together if they have the same value of the component that is used as the basis for aggregation. For example, athletes participating in a European championship may be aggregated according to the countries they are from, so that each aggregate consists of athletes from the same country. In this example, aggregation is done on the basis of a common value of the attribute “Country”. Data aggregation on the basis of a common value of a referrer makes sense only if the dataset also has other referrers. Thus, if we had data containing the results of the latest test on mathematics taken by a group of students and would like to aggregate the data according to common values of the referrer “Student”, we could not build any aggregates, since each value of the referrer occurs in the reference set only once. If, however, the dataset contained results of multiple tests, for example tests in different subjects or a sequence of tests in mathematics taken during a certain time period, we could build aggregates by uniting the results of different tests of the same student. In this case, the dataset has two referrers, “Student” and “Subject” (or “Time”), and hence each reference is a pair consisting of the name of a student and the name of a subject (or an indication of a time moment). The name of each student may occur in the reference set as many times as there are different subjects (or time moments),<sup>13</sup> and hence it is quite possible to group references containing the same student’s name.

Another case of aggregation on the basis of a component with a nominal value set arises when this component has very many different values. In this case, the explorer needs to divide the whole value set into equivalence classes. Then, references for which the values of this component belong to the same equivalence class are grouped together. For example, in a dataset

---

<sup>13</sup> We have written “may occur” rather than “occurs” because the data may be incomplete, i.e. the results of some tests may be missing.

containing the results of a world championship, the countries from which the athletes come may be very numerous. When the athletes are grouped according to common countries, the resulting number of aggregates is very large and difficult to analyse. Therefore, the analyst may decide to group the countries, for example into African, Asian, European, etc., and aggregate athletes from countries in the same group.

When aggregation is done on the basis of a component with an ordered (linearly or partly) value set without distances, the same approaches may be used as when dealing with an unordered value set. When the number of different values is moderate, it is possible to aggregate data items with coincident values of the component, and if the analyst finds the values too numerous, he/she may group them into equivalence classes. However, in defining the equivalence classes, it is desirable to preserve the order, i.e. the classes should be formed from consecutive values. For example, an explorer analysing data concerning a set of people serving in a military force may wish to aggregate these people according to their military rank. The ranks may have been previously grouped into equivalence classes such as "Private" (including all grades of privates), "Sergeant" (all ranks of sergeants), "Warrant officer", etc. The reason for wishing to use grouping may be not only the large number of different military ranks occurring in the dataset but also a wish to build aggregates of more or less equal size. It is not very likely that the set of servicemen under analysis includes as many generals as there are privates. Uniting generals with colonels and majors could make the sizes of the aggregates more balanced. However, it would hardly be reasonable to unite generals with corporals and colonels with sergeants.

In aggregation on the basis of a component that has a linearly ordered value set with distances, such as a numeric or temporal attribute or referrer, the value range of the component is divided into intervals, which play the same role as equivalence classes in the case of a nominal attribute or referrer. The intervals are often defined so as to have equal lengths, but other principles of division may be used as well. For example, one may aggregate the districts of Portugal on the basis of the attribute "Proportion of people aged from 0 to 14 years" by dividing the value range of this attribute into ten intervals of equal length (of course, the number of intervals may be changed arbitrarily). However, the analyst may also decide to define the intervals in such a way that the resulting groups of districts have approximately equal sizes. Hourly data concerning air pollution may be aggregated into 24-hour periods, but it may be more reasonable to separate night and day hours and to consider unequal intervals, for example, from 11 p.m. to 5 a.m. and from 6 a.m. to 10 p.m.

Aggregation on the basis of a spatial component is done by defining spatial compartments, which play the role of equivalence classes: each compartment includes several original values of the spatial component, and these values are treated in the same way. Such compartments may be defined either by joining individual values (typically, these values should be adjacent) or by space tessellation, which is most often done in a regular manner, for example a two-dimensional space may be divided into squares of the same size. Joining is often applied when the original values of the spatial component are already spatial compartments, i.e. have a certain spatial extent. For example, districts in a territory can be merged into larger districts. One can do this according to some predefined hierarchy of territorial division, such as administrative divisions, but in general this is not necessary. For example, the districts of Portugal may be aggregated on the basis of the provinces that they belong to, but the analyst may prefer to define geographical regions such as the western coast, the north-east, the central inland region, etc. Space tessellation may be recommended when the original values of the spatial component are points, i.e. have no spatial extent. Thus, in order to aggregate earthquake occurrences, one may divide the territory under analysis by introducing a regular grid. Then, earthquakes located within the same grid cell are united in an aggregate. The same technique is appropriate for aggregation of data specified in a raster format, which is often used to represent spatially continuous phenomena. In such data, attribute values refer to cells of a fine grid. For aggregation, one may introduce a coarser grid so that cells of the original grid that fit into the same cell of the new grid are united.

It may be noted that all cases considered thus far involve reduction of the original value set of the component used as the basis for the aggregation by means of treating some values of this component as equivalent. The same applies to aggregation on the basis of more than one component: the value set of each component is reduced, and the possible combinations of values from these reduced sets define the aggregates.

It is clear that equivalence classes may be defined in many ways for each of the components used for aggregation. In particular, one can even treat all values of a component as equivalent. Thus, in the example of aggregation of the data concerning the performance of students, references were grouped according to the value of the component “Student” with no regard to the value of the second referrer (i.e. either “Subject” or “Time”). This actually means that all of the values of the second referrer were united into a single equivalence class and therefore treated as the same.

It should be noted that when attributes rather than referrers are used as the basis for data aggregation, it may happen that some of the resulting aggregates are empty. The reason may be that the dataset contains no ac-

tual occurrences of values belonging to some equivalence class of one of the attributes, or that for some combination of equivalence classes of several attributes there are no corresponding combinations of original values.

We would also like to mention that the degree of aggregation, i.e. the sizes of the aggregates and their number, may vary greatly. An analyst chooses an appropriate degree of aggregation depending on the goals of the analysis; in particular, on how precise the pattern to approximate the behaviour of the phenomenon under analysis must be. The degree of aggregation depends also on the amount of data being analysed and some properties of the data, first of all the variability of the data: it makes sense to build aggregates such that the corresponding characteristics do not vary too much. The admissible degree of variation depends, in turn, on the required precision of behaviour characterisation.

It may be quite reasonable to explore data on different aggregation levels. The highest possible level is when the whole reference set is considered as a single aggregate. The analyst may start by considering aggregate characteristics of the entire dataset and then decrease the level of aggregation until it becomes possible to approximate the behaviour by a suitable pattern, in terms of simplicity and precision.

#### 4.5.4.2 Characterising Aggregates

One of the basic characteristics of an aggregate is the number of elements (i.e. individual references) included in it. Counts of elements are especially important when aggregates are defined on the basis of characteristics, i.e. attribute values. In this case, these counts show how many references with certain characteristics exist in the dataset. For example, after having aggregated the districts of Portugal on the basis of the attribute “Proportion of people aged from 0 to 14 years”, we would of course be interested in how many districts exist with values of that attribute that fit into each of the intervals into which we divided the value range of the attribute.

Besides counts, aggregates may receive other collective characteristics derived from the characteristics of the individual references included in them. The most frequently used operations for deriving characteristics of aggregates from the characteristics of their members are the following:

- *The sum of the attribute values* referring to individual members of an aggregate. For example, for the districts of Portugal aggregated on the basis of the proportion of children, we would like to know the total population of each group of districts. This can be ascertained by summing the values of the attribute “Population number” for all members of the group.

It should be noted that summing does not make sense for any numeric attribute and for any set of references. It is appropriate in some cases to add together values of an attribute that express absolute quantities, but not to add proportions, rates, or ranks. Even absolute quantities cannot always be summed. Thus, if we aggregate the time-series data in the dataset concerning crime in the USA by grouping the references on the basis of a common state irrespective of the time, it would be incorrect to compute the sum of values of the attribute “Population number” referring to different time moments. In this case, the same limitations apply as for the accumulation of values over time intervals discussed earlier.

- *The arithmetic mean, or average*, of the individual values of a numeric attribute. For example, when analysing a dataset concerning the performance of students, one might be interested in computing the average of the marks received by each student in different tests. It may be appropriate to consider means in combination with some measures of the variation of the values, such as the standard deviation or the variance.
- *A weighted average*, which differs from an ordinary average in the following way. In computing an ordinary average, the elements of an aggregate are treated equally: their characteristics are summed, and the sum is divided by the number of elements. In a weighted average, the contribution of each element to the combined characteristic of the aggregate depends on a certain measure, called the *weight* of the element. For example, in combining crime rates in all states of the USA into an aggregated figure for the whole country, it would hardly be appropriate to compute a simple average. It seems more reasonable to take into account the number of inhabitants in each state: the contribution of highly populated states to the overall characteristic should be greater than that of states with a smaller population. The number of population in each state is considered in this case as the weight of this state.

A weighted average is computed in the following way: the attribute value associated with each member of an aggregate is multiplied by the weight of this member (which must be non-negative). The sum of all such products is divided by the sum of the weights of all aggregate members (see (4.7)). In the example of the crime rates, the crime rate of each state would be multiplied by the population of that state. Then, the products computed for all states would be added together and the resulting sum divided by the total population of the country.

- *The value range*, i.e. the minimum and maximum values of an attribute that has an ordered value set, and possibly the difference between the minimum and maximum if the value set has distances. For example, in analysing the performance of students, one may wish to know the mini-

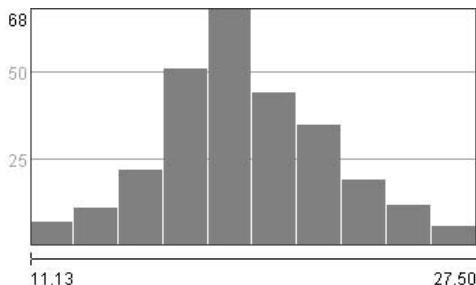
mum and maximum mark received by each student. However, we are not sure whether it would be appropriate in this case to compute the difference between the minimum and the maximum mark. This depends on whether the mark scale is supposed to be of interval type (i.e. with distances) or ordinal (i.e. with no distances).

- *The mode*, or the most frequent value. This operation is especially suitable for attributes with discrete value sets, in particular, values of nominal type. For example, in aggregating land use data specified in a raster format by means of a coarse-granularity grid, it may be useful to determine the type of land prevailing in each grid cell, for example urban, agricultural, or forest. For a more detailed analysis, the same aggregates may be characterised by the frequency of occurrence of each value or the proportion of references characterised by each value.
- *The median and other positional measures*. To recall, the median of a set of numeric values is the value that divides the set into two equal parts so that the values in one part are less than or equal to this value and the values in the other part are greater than or equal to this value. Analogously, each half of the set can in turn be divided into halves; the dividing values are called the first and the third *quartile*. By generalising this division procedure, *deciles* and *percentiles* are defined: deciles divide the value set into tenths, and percentiles divide it into hundredths. All of these values are called “positional measures”, because they locate the positions of particular values relative to a set of other values. Positional measures may be preferable to the arithmetic mean and the value range for characterising an aggregate when the set of original attribute values corresponding to this aggregate contains outliers. For example, we can aggregate the data about crime in the USA by uniting all references that have a common state but different years. Then, we can characterise the overall crime situation in each state during the 41-year period from 1960 to 2000 by computing the medians and quartiles of the values of attributes referring to the various years. In this way, we can disregard occasional very high or very low values that might have occurred in some years.

Unlike counts, all of the aggregate characteristics listed above are derived from the values of some attributes and have the same nature as the attribute values that they have been derived from. Hence, one can visualise such aggregate characteristics using the same visualisation methods as for the original attribute values. In contrast, counts, or aggregate sizes, require specific visualisation techniques. Therefore, we would like to consider the visualisation of counts separately from the visualisation of other aggregate characteristics.

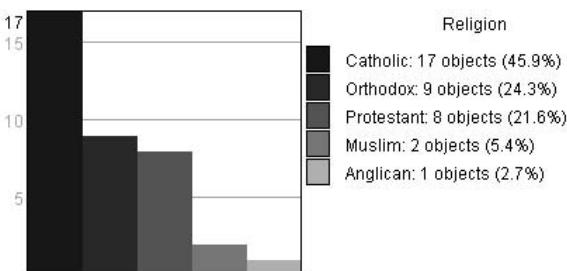
#### 4.5.4.3 Visualisation of Aggregate Sizes

Probably the most well-known and widely used visual display of aggregated data is the frequency histogram, which represents sizes of aggregates by heights of juxtaposed bars. For example, the histogram in Fig. 4.61 corresponds to aggregation of the districts of Portugal on the basis of the values of the attribute “Percentage of people aged from 0 to 14 years”. The value range of the attribute (from 11.13 to 27.50%) has been divided into ten equal-length intervals, and the aggregates have been formed by uniting districts with attribute values fitting in the same interval. The histogram in Fig. 4.61 contains one bar for each interval, with a height proportional to the size of the corresponding group of districts. In other words, the height of a bar shows how many values in the respective interval are present in the dataset. The bars are ordered according to the order of the intervals. In fact, the widths of the bars are also meaningful: they show the lengths of the intervals. In this case, all intervals are of the same length, but the data could also be aggregated on the basis of unequal intervals, so that the representation of the interval lengths might become important.



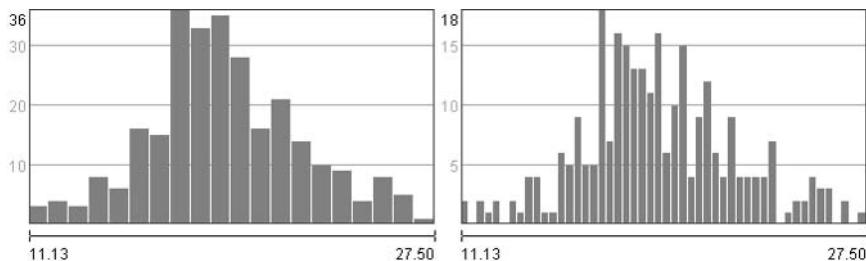
**Fig. 4.61.** This histogram shows the result of aggregation of the districts of Portugal according to the values of the attribute “Percentage of people aged from 0 to 14 years”, whose value range has been divided into ten equal-length intervals. The heights of the bars are proportional to the sizes of the aggregates

For comparison, let us look at the histogram in Fig. 4.62, which represents the results of aggregation on the basis of an attribute with a nominal-value scale. Specifically, a set of countries in Europe has been divided into subsets according to the values of the attribute “Dominant religion”. Five different values of this attribute are present in the dataset. The values are not ordered, and no distances between them are defined. Therefore, the order of the bars in the histogram may be arbitrary. In particular, the bars may be arranged in order of decreasing size, as has been done in Fig. 4.62. The widths of the bars do not convey any information.



**Fig. 4.62.** The countries of Europe have been aggregated here according to the values of the attribute “Dominant religion”. For each religion, the histogram shows the number of countries in which this religion is dominant

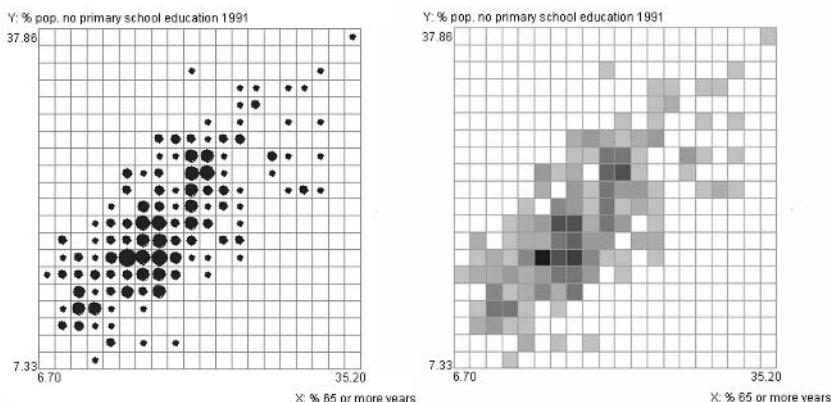
Let us now return to the aggregation of the Portuguese districts on the basis of the proportions of children (i.e. people aged from 0 to 14 years). In Fig. 4.61, the value range of the attribute has been divided into ten intervals. The choice of the number of intervals is quite arbitrary. One could also divide the value range into 20 intervals, as in Fig. 4.63, left, or into 50 intervals, as in Fig. 4.63, right. As we have mentioned, variation of the degree of data aggregation may be reasonable in data analysis. Therefore, it is desirable that data aggregation tools provide sufficient flexibility to an explorer for defining and redefining data aggregates.



**Fig. 4.63.** These histograms correspond to division of the value range of the attribute “Percentage of people aged from 0 to 14 years” into 20 (left) and 50 (right) equal-length intervals. The heights of the bars show the number of districts in Portugal with values of this attribute fitting in the respective intervals

As we have seen, a histogram representing the results of aggregation on the basis of values of a numeric attribute shows not only the aggregate sizes but also how the aggregation has been done, i.e. how the value range of the attribute has been divided into intervals. This idea can be extended quite easily to the case of aggregation on the basis of two numeric attributes. Thus, the horizontal display dimension may show the division of the value range of one of the attributes, and the vertical dimension may show

the same for the other attribute. Accordingly, the plane is divided into rectangular cells. In these cells, the sizes of the respective aggregates can be represented, for example, by the sizes of marks or by brightness (or darkness), as is demonstrated in Fig. 4.64. Such a display is known as a “binned scatterplot” or “two-dimensional histogram”. Carr et al. (1992) argue that aggregation in binned scatterplots or other binned displays (in particular, maps) is better to do using a grid of hexagons rather than squares or rectangles. These authors point out that using a rectangular grid results in the marks in a display being arranged in horizontal and vertical lines, which attract the viewer’s attention and thereby distract him/her from seeing the patterns generated by the data.

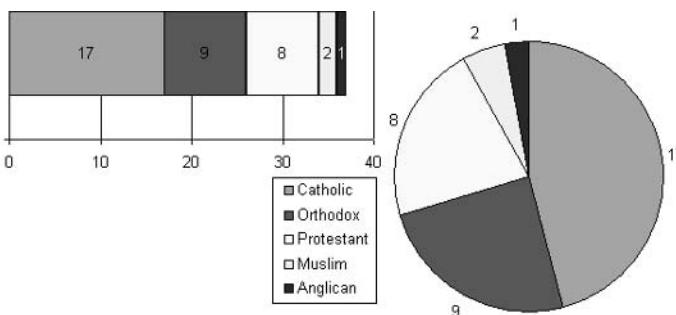


**Fig. 4.64.** Two possible visualisations of the results of aggregation of the districts of Portugal according to the values of the two numeric attributes “Percentage of people aged 65 or more years” and “Percentage of population without primary school education”. The value range of each attribute has been divided into 20 equal-length intervals. In both displays, the horizontal and vertical dimensions represent the division of the value ranges of the attributes. In the grid cells, the sizes of the aggregates are represented by circle sizes (left) and by darkness (right)

The idea of a two-dimensional histogram could be extended to the case of aggregation on the basis of three numeric attributes by involving the third spatial dimension of the display. However, there are obvious limitations on further extensions in this direction.

In addition to histograms, another popular way to visualise sizes of aggregates is segmentation of a figure in proportion to the sizes. For example, in Fig. 4.65 the sizes of the groups of European countries aggregated according to their dominant religion are represented in a segmented bar (left) and in a pie chart (right). As compared with the histogram in Fig. 4.62, a segmentation-based display is more convenient for estimation of

the relative size of each aggregate with respect to the whole reference set. In other words, segmentation is a rather straightforward representation of the division of the reference set into aggregates.

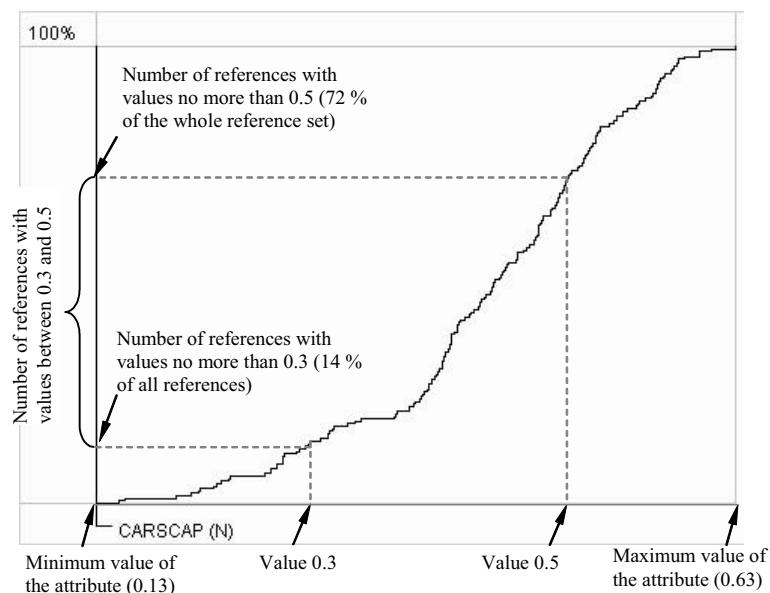


**Fig. 4.65.** The sizes of the groups of European countries aggregated according to their dominant religion are represented by segmented figures

The segmentation-based displays in Fig. 4.65 represent results of data aggregation on the basis of a qualitative attribute. It is also possible to do the same for numeric or other types of attributes. However, segmented representations have a disadvantage as compared with histograms: they do not show how the value range of an attribute is divided into intervals. Representation of the interval breaks may be not so important if the aggregation is done on the basis of equal-length intervals. In that case, if the order of the segments in a segmented figure is the same as the order of the intervals, it is relatively easy to identify the interval that each segment corresponds to (nevertheless, there may be problems if the aggregates corresponding to some of the intervals are empty).

There is a display technique that combines the advantages of a histogram and of a segmentation-based representation. It uses one of the spatial display dimensions to represent the value range of an attribute with an ordered value scale, and the division of this range into intervals. Another spatial dimension is used to show the corresponding division of the reference set into aggregates, i.e. the sizes of the aggregates and their proportions in relation to the whole reference set. The technique involves a graph known in statistics as the *cumulative frequency curve*, or *ogive*.

Figure 4.66 demonstrates the principle of the construction of a cumulative frequency curve. The horizontal dimension is used here to represent the value range of the attribute used for data aggregation. The vertical dimension is used to represent the number of references. The top edge of the display corresponds to the number of references in the entire reference set.



**Fig. 4.66.** The principle of the construction of a cumulative frequency curve

For each attribute value  $x$  between the minimum and maximum, one can count how many references in the reference set have corresponding values less than or equal to  $x$ . This count (let us denote it by  $N_x$ ) can be represented by a vertical position in the space of the graph, which corresponds to the horizontal position representing the value  $x$ . The cumulative curve is constructed by connecting all consecutive points with coordinates  $(x, N_x)$ , where  $x$  is assigned successive attribute values from the minimum to the maximum.

The ogive has certain useful properties that make it suitable for analysis of the distribution of the values of an attribute across a population<sup>14</sup> or any reference set considered as a population, i.e. considered irrespective of ordering, distances, or any other relations between the elements. Thus, steep segments of a cumulative curve correspond to groups of references with close values of the attribute. The height of such a segment shows the number of references with close values. Horizontal segments correspond to “gaps” in the sequence of values, i.e. where there are no references with values in an interval. Of course, the distribution of values can also be analysed using a histogram, which is easier to interpret than an ogive. However, a histogram display has a serious disadvantage: its shape depends

<sup>14</sup> We use the term “population” in the sense “statistical population”, i.e. a discrete reference set without ordering and distances between the elements.

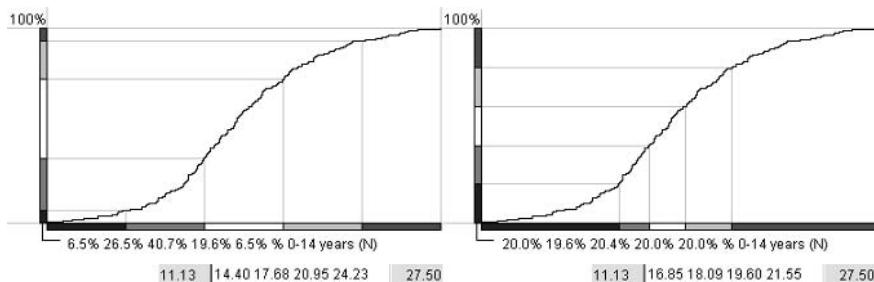
significantly on how the value range of the attribute is divided into intervals (compare, for example, the histograms in Figs 4.61 and 4.63, which are constructed on the basis of the same attribute). Unlike the case of a histogram, construction of an ogive does not involve dividing the value range of the attribute into intervals. Therefore, its shape may be regarded as a more objective summary of the value distribution.

We realise that this description of the properties of the ogive and its use in analysing statistical distributions is not quite in place here, in the middle of a discussion concerning methods of combined visualisation of the division of the value set of an attribute and the sizes of the resulting aggregates. However, since this is our first mention of a cumulative curve, we considered it reasonable to describe its construction and general properties before we started explaining any modifications of the standard technique. Let us now return to the main topic of our discussion.

As we have mentioned, the construction of an ogive is not based on a division of the value set of the attribute. However, the display can be modified so as to represent simultaneously a division of the attribute values into intervals and the corresponding division of the reference set into aggregates. This is done by segmentation of the horizontal and vertical axes. The horizontal axis is divided into segments in accordance with the division of the value range of the attribute into intervals: for each interval break, there is a corresponding position on the horizontal axis, which is used as a divider of the axis. For every such divider, a vertical line is drawn in the upward direction until it crosses the curve. From the crossing point, a horizontal line is drawn towards the vertical axis until it crosses that axis. Owing to the properties of the cumulative curve, this divides the vertical axis in proportion to the number of references with corresponding attribute values belonging to the intervals shown on the horizontal axis. Why this is so can be seen from the drawing in Fig. 4.66. On the horizontal axis, the positions of two attribute values are shown, 0.3 and 0.5. The vertical position corresponding to the value 0.3 shows the number of references with values less than or equal to 0.3; these references are about 14% of all references. Analogously, the vertical position corresponding to 0.5 shows the number of references with values less than or equal to 0.5 (about 72% of the reference set). The segment of the axis between these two vertical positions corresponds to the references with values of the attribute greater than 0.3 but less than or equal to 0.5, and the size of this segment represents 58% (i.e. 72% minus 14%) of the reference set.

Figure 4.67 shows two screenshots of a cumulative-curve display with a segmentation of the axis. The display represents the distribution of the values of the attribute “Percentage of people aged from 0 to 14 years” over the set of the districts of Portugal (these are the same data as were used for

constructing the histograms in Figs 4.61 and 4.63). The screenshot on the left represents the division of the value range of the attribute into five equal-length intervals. The corresponding segmentation of the vertical axis shows the sizes of the resulting aggregates in relation to the size of the entire set. Below the graph, the proportions of the aggregates in the entire set are shown as numbers. We can see, for example, that only 6.5% of all districts have values of the attribute belonging to the lowest one-fifth of the value range (i.e. from 11.13 to 14.40), and the same is true for the highest one-fifth, i.e. from 24.23 to 27.50. These subsets of districts are represented by quite small segments on the vertical axis, whereas the middle segment is the longest segment – it represents 40.7% of the whole set of districts, with the corresponding attribute values belonging to the middle interval from 17.68 to 20.95.



**Fig. 4.67.** Two screenshots of a cumulative-curve display represent different divisions of the value range of the attribute “% 0–14 years” on the horizontal axis, and the corresponding divisions of the set of the districts of Portugal on the vertical axis. On the right, the value range of the attribute has been divided so that the resulting subsets of districts have approximately equal sizes. Below each graph, the values of the interval breaks are shown

The screenshot on the right represents another division of the value range into five intervals: the interval breaks have been chosen here so that the reference set (i.e. the set of districts of Portugal) has been divided into subsets of approximately equal size<sup>15</sup>. Accordingly, the vertical axis is split into five segments of equal length. The horizontal axis is segmented according to the positions of the interval breaks. It can be seen that the first and last segments, which correspond to value intervals from 11.13 to 16.85 and from 21.55 to 27.50, respectively, are much longer than the rest; they are even longer than the total length of the remaining three intervals taken together.

<sup>15</sup> It is not always possible to divide a set in this way into subsets of exactly equal size, because attribute values for some references may coincide.

We would like to stress that the left and right parts of Fig. 4.67 differ only in the segmentation of the axes, while the curve remains the same. In principle, a cumulative-curve display can be implemented so that the segmentation of any of the axes may be changed interactively. This would allow a dual use of the tool:

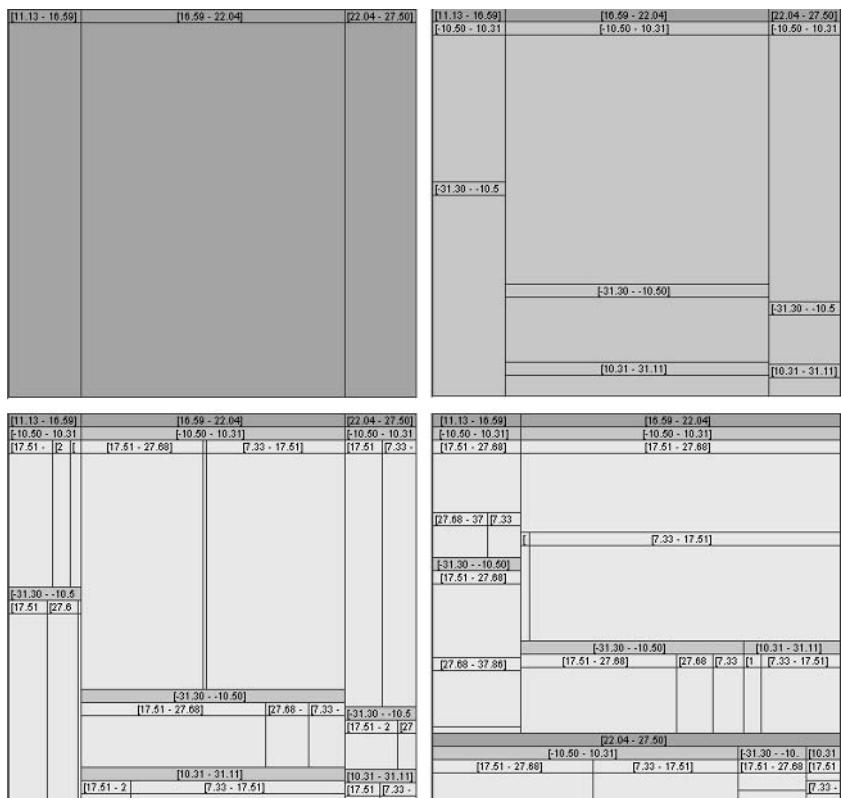
- The user could vary the division of the value range of the attribute and observe the variation of the sizes of the corresponding aggregates.
- The user could partition the reference set in any desired proportion (by splitting the vertical axis) and observe the corresponding division of the attribute value range.

Hence, the display may become not only a means for visualising previously defined aggregations but also a tool for defining various aggregations and, in this way, allow one to explore the distribution of attribute values over a reference set.

Up to this point, we have discussed the use of segmentation-based displays for representation of data aggregates formed on the basis of a single attribute. However, the idea of segmentation is, in principle, applicable to aggregation on the basis of any number of attributes: each segment of a figure representing the division of the reference set according to one attribute may, in turn, be segmented to show a division according to another attribute, and so on.

Among the visualisation techniques based on such recursive segmentation, the mosaic plot (Friendly 1994) and the treemap (Shneiderman 1992) are the best known. These two techniques look very similar. They both divide the two-dimensional display space into rectangles representing aggregates, so that the sizes of the rectangles are proportional to the sizes of the aggregates. In a mosaic plot, the horizontal and vertical display dimensions are handled independently, and segmentation according to each attribute is applied either to the horizontal or to the vertical dimension. The operations of horizontal and vertical division typically alternate: for the first attribute, the whole display area is split in the horizontal dimension (i.e. by vertical dividers); for the second attribute, each of the rectangles resulting from the first division is divided in the vertical dimension; for the third attribute, each of the rectangles resulting from the second division is segmented again in the horizontal dimension; and so on.

In a treemap display, several alternative methods can be used to divide the display area into segments with areas proportional to the sizes of the aggregates. One of the methods is the same as is used in mosaic plots; it is demonstrated in three of the four images shown in Fig. 4.68. The fourth image demonstrates a different segmentation technique. Let us consider this figure in more detail.



**Fig. 4.68** Treemap displays representing aggregation on the basis of one (top left), two (top right), and three (bottom left and right) attributes. The two displays at the bottom differ in the segmentation method applied. The illustration was produced using the demo version of the “Treemap” tool available at <http://www.cs.umd.edu/hcil/treemap>. These and further screenshots of the Treemap are used with permission of the Human-Computer Interaction Lab, University of Maryland, 2005

All four images were constructed by applying a treemap tool to the Portuguese census dataset. The image at the top left shows a division of the set of districts according to the values of the attribute “Percentage of people aged from 0 to 14 years in 1991”. The value range of the attribute has been divided into three equal-length intervals: from 11.13 to 16.59, from 16.59 to 22.04, and from 22.04 to 27.50. As a result, the set of districts is partitioned into three subsets, and the display space is accordingly split into three rectangles with areas proportional to the sizes of these subsets. For this purpose, two vertical dividers have been drawn. The rectangle on the left represents the subsets of districts with a low proportion of children (below 16.59%), the rectangle in the centre corresponds to districts with a

medium proportion of children (from 16.59 to 22.04%), and the rectangle on the right represents the subset of districts with a high proportion of children (22.04% and more).

In the next step, we have added another attribute to be used for aggregating the districts, specifically, the attribute “Population change from 1981 to 1991” (as a percentage of the population in 1981). The value range of this attribute (from -31.30 to 31.11) has also been divided into three equal-length intervals, by breaks at -10.50 and 10.31. As a result, each of the rectangles resulting from the previous segmentation (i.e. on the basis of the percentage of people aged from 0 to 14 years) is split by horizontal dividers into segments corresponding to the aggregates defined on the basis of the two attributes. This is shown at the top right of Fig. 4.68. It can be seen that the leftmost rectangle is divided into two segments rather than three. If we look at the labels, we see that there are segments corresponding to the intervals from -31.20 to -10.50 and from -10.50 to 10.31, but no segment for the interval from 10.31 to 31.11. This means that the corresponding aggregate is empty: there are no districts with a low proportion of children and high population growth.

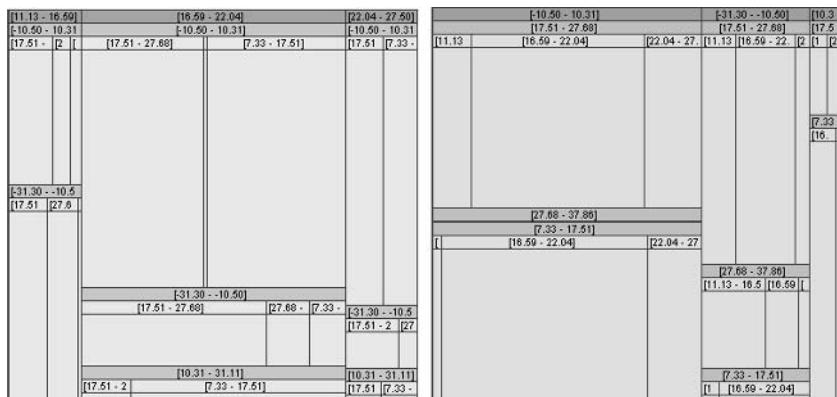
The image at the bottom left corresponds to one more attribute added to the definition of the aggregates, specifically, the attribute “Percentage of population without primary school education in 1991”. Again, the value range of the attribute (from 7.33 to 37.86) has been divided into three equal-length intervals, by breaks at 17.51 and 27.68. Accordingly, each of the rectangles resulting from the second division has been further segmented into smaller rectangles by drawing vertical dividers. Through an attentive examination of the resulting display, we can detect that there are no segments corresponding to certain combinations of attribute values, i.e. those combinations never occur in the dataset. Thus, for example, there are no districts with a medium or high proportion of children, high population growth, and a high proportion of uneducated people. Some of the segments are very narrow; this means that the sizes of the respective aggregates are quite small.

In the images described above, alternating horizontal and vertical segmentations of the display space have been applied to represent successive divisions of the set of districts. The same approach is used in mosaic plots. However, as we have already mentioned, the treemap tool offers several alternative space-partitioning methods. For comparison, the image at the bottom right of Fig. 4.68 shows the same division of the set of districts as in the bottom left but represented by means of another method of display space segmentation.

Since the space segmentation in a treemap display is done in a recursive manner, the appearance of the display depends greatly on the order in

which the attributes used for defining the aggregates are considered. This is demonstrated in Fig. 4.69. The image on the left is the same as in Fig. 4.68. It corresponds to the following order of the attributes:

1. Percentage of people aged from 0 to 14 years in 1991.
2. Population change from 1981 to 1991.
3. Percentage of population without primary school education in 1991.



**Fig. 4.69.** The segmentation of the space in a treemap display depends on the order in which the attributes are considered. These two images represent the same aggregates, defined on the basis of three numeric attributes. The difference between the images is due to the different order in which the attributes have been considered

The image on the right represents the same aggregates of districts, but the segmentation of the display area corresponds to a different order of the attributes:

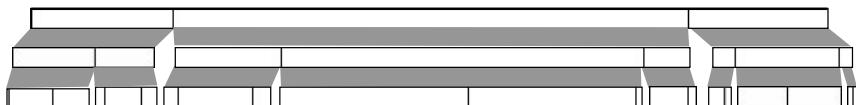
1. Population change from 1981 to 1991.
2. Percentage of population without primary school education in 1991.
3. Percentage of people aged from 0 to 14 years in 1991.

In both images, the same method of space segmentation has been applied, specifically, alternating horizontal and vertical division (this method is called “Slice and dice” in the user interface of the tool). The images differ only because of the different order in which the attributes were used for the segmentation. The treemap tool allows the user to change the order of the attributes interactively.

To finish the description of this treemap tool, we would like to mention that segments representing aggregates may be coloured to portray characteristics of those aggregates. The user may choose the attribute to be used for characterising the aggregates and the method for deriving the character-

istics of the aggregates from those of their members. The options available are the average, weighted average, minimum, and maximum. The same idea may be also applied to other types of displays of aggregated data, for example one- and two-dimensional histograms or mosaic plots.

Many people may find a treemap display rather difficult to understand. However, this is just one possible approach to portraying a hierarchy of divisions of a reference set. This hierarchy can be viewed as a tree, and a tree can be visualised in many different ways. Thus, we can represent the same division of the districts of Portugal as in Fig. 4.68 (bottom) by something like what is shown in Fig. 4.70. This drawing contains three horizontal layers, which correspond to the three levels of division of the reference set. At the top, the whole reference set is represented by a single bar, which is segmented in proportion to the division according to the attribute “Percentage of people aged from 0 to 14 years in 1991”. In the second layer, there are three bars, which have the same lengths as the segments of the top bar. Each of these three bars represents one of the subsets resulting from the first division, and hence all together they represent the entire set of districts. The segmentation of each of the bars corresponds to the division of the respective subset according to the values of the attribute “Population change from 1981 to 1991”. Analogously, the third layer demonstrates how each of the segments shown in the second layer is further divided according to the values of the attribute “Percentage of population without primary school education in 1991”. The grey parallelograms between the layers provide visual linking between corresponding fragments.



**Fig. 4.70.** An alternative representation of the division of the set of the districts of Portugal according to the values of three numeric attributes

It is possible to enhance this visualisation by representing additional information. In particular, it is appropriate to indicate what attribute has been used for the division at each level and into what intervals or subsets the value set of this attribute has been divided. The segments may be coloured according to some characteristic of the aggregates that they represent. Another possible enhancement is the inclusion of cumulative-curve displays to represent simultaneously the division of the value ranges of the attributes and the corresponding partitioning of the reference set.

An obvious problem is that of handling very small reference subsets: the corresponding graphical features need to be scaled in proportion to the

sizes of the subsets and hence may become too narrow and therefore barely legible, as in the graph at the bottom right in Fig. 4.70. In fact, this problem appears in other representations as well. Thus, some of the bars in a histogram may be so small that it is hard to say whether they exist or not. Some segments in a treemap display may also be hardly visible. A possible solution to this problem is interactive zooming or focusing.

#### **4.5.4.4 Sizes Are Not Only Counts**

Up to this point, we have discussed various methods for visualising the sizes of aggregates assuming that the size of an aggregate is the number of its members. However, a count of aggregate members is not the only measure of size. Thus, in analysing aggregates consisting of districts of Portugal, an explorer might be interested not so much in the number of districts in an aggregate as in the total population in these districts, the total area of these districts, or the total number of uneducated people. In aggregating data about earthquakes, it may be less relevant to know how many earthquakes are in each aggregate than how many people in total died or were injured, or how many buildings collapsed or were damaged. When data about traffic jams are aggregated by days of the week, an analyst might like to know not only the number of occurrences of traffic jams on each day but also their total duration and total length. Such “totals” may be viewed as a kind of measure of aggregate size. It may be noted that these measures are problem-specific, unlike the count of elements, which may characterize any aggregate irrespective of its nature.

In general, a measure of aggregate size may be based on any attribute if certain requirements are fulfilled:

- The attribute has a value scale of ratio type, i.e. with ordering, distances, and a true zero.
- All values of the attribute are non-negative.
- The value for a set of references is the sum of the values of the elements.
- The value for a union of two or more sets with no common elements is the sum of the values for these sets.

Any problem-specific measure of aggregate size may be visualised in the same way as counts of aggregate members. Thus, the heights of the bars in a histogram may be proportional not to the numbers of elements in the respective reference subsets but to some other measure of the sizes of these subsets. For example, the bars of the histograms in Figs 4.61 and 4.63 could represent the total number of children in the respective groups of districts. In an analogous way, the two-dimensional histograms in Fig.

4.64 could be modified. In segmentation-based displays, the segmentation could be done according to any meaningful measure of aggregate size. Thus, the segmented bar and the pie chart in Fig. 4.65 could be divided in proportion to the total populations in the groups of European countries formed according to their dominant religion. The treemap tool described earlier can divide the display space not only in proportion to the number of elements in each aggregate but also in proportion to the sum of the values of any user-selected attribute.

As we are cumulative-curve enthusiasts, we would like to demonstrate how problem-specific aggregate sizes can be represented in a cumulative-curve display and what opportunities this provides for data analysis.

Let us recall how a traditional cumulative frequency curve is built (see Fig. 4.66). The positions in one of the planar dimensions (e.g. horizontal) represent values of an attribute with an ordered value set, from the minimum to the maximum. The positions in the other dimension (in our case, vertical) represent possible sizes of reference subsets, which vary from 0 to 100%, where 100% corresponds to the entire reference set. The curve matches each value  $x$  of the attribute with the size of the subset of references that have attribute values less than or equal to  $x$ . Traditionally, the size means the number of elements in the subset. However, any other measure of the subset size may also be used. Moreover, we can construct cumulative curves for different size measures and overlay them in a common display area so that they can be easily compared. The horizontal (attribute) axis is common to all these curves. The same applies, in principle, to the vertical (set size) axis, which represents proportions of the total size of the entire reference set. However, we can use the axes to represent additional information by means of segmentation. The attribute axis is segmented according to the division of the attribute value range into intervals. This division does not depend on the measure of aggregate size used, and hence is common to all overlaid curves. The set size axis is segmented in proportion to the sizes of the reference subsets resulting from the division of the attribute value range. The sizes, and hence the proportions, certainly depend on the size measure used. Therefore, in order to show the divisions of the reference set corresponding to different size measures, we need to introduce additional axes. More specifically, there must be as many set size axes (in our layout, vertical) as there are curves overlaid in the display. These axes are drawn parallel to each other. Each of the axes is segmented individually, using the corresponding cumulative curve.

An example of such an enhanced cumulative-curve display is shown in Fig. 4.71. All four images are screenshots of the same display and differ only in the segmentation of the axes. The display represents an aggregation of the districts of Portugal according to the value of the attribute “% em-

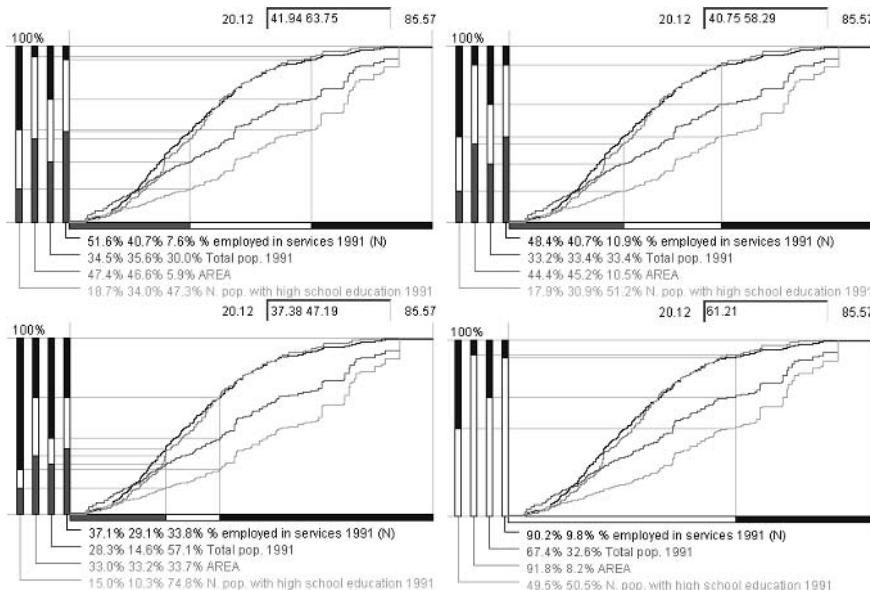
ployed in services 1991” (the percentage of the working population in each district employed in services). The values of this attribute range from 20.12 to 85.57%. The display contains an “ordinary” cumulative frequency curve for this attribute, i.e. one where the count of districts has been used as the measure of aggregate size. This frequency curve is drawn in black. In addition to the frequency curve, the display contains three more cumulative curves:

- a cumulative population curve, with the attribute “Total pop. 1991” (the total population of a district in 1991) used as the measure of aggregate size;
- a cumulative area curve, constructed on the basis of the attribute “Area” (i.e. the area of a district);
- a cumulative curve of the number of people with high school education, constructed by summing the values of the attribute “N pop. with high school education”.

Accordingly, there are four vertical axes on the left of the display.

It may be seen from Fig. 4.71 that the cumulative area curve almost coincides with the cumulative frequency curve, while the other two curves are quite different. It is therefore not surprising that we encounter differences in the segmentation of the vertical axes.

At the top left, the value range of the attribute is divided into three equal-length intervals. According to the cumulative frequency curve, the group of districts with low values of the attribute contains 51.6% of the total number of districts; medium values occur in 40.7% of all districts, and high values in only 7.6% of the districts. The division of the total area of the country between these district groups is quite close to these proportions: 47.4%, 46.5%, and 5.9%. At the same time, the total population is divided almost equally between the groups: 34.5%, 35.6%, and 30.0%. This allows us to conclude that the districts with high percentages of people employed in services are highly populated (30% of the total population lives in 7.6% of the districts) and, moreover, densely populated (the total area of these districts which have 30% of the population of the country, is only 5.9% of the area of the whole country). It is even more peculiar how the people with high school education are distributed between the district groups. 47.3% of these people live in the 7.6% of districts with high employment in services, while in more than half of the districts, where the proportion of employment in services is low, there are only 18.7% of all people in the country who have high school education. This indicates a correlation between the proportion of employment in services and the educational level of the population.



**Fig. 4.71.** An enhanced cumulative-curve display may be used to divide a reference set into subsets with desired sizes according to various measures of set size. Here, the districts of Portugal are aggregated according to the value of the attribute “% employed in services 1991”. At the top left, the value range of the attribute is divided into three equal-length intervals. At the top right, the value range is divided so that the corresponding groups of districts have approximately equal total populations. At the bottom left, three groups of districts have approximately equal total areas. At the bottom right, the value range is divided into two intervals so that each of the resulting two groups of districts contains approximately half of all people in the country who have high school education

An enhanced cumulative-curve display can be used not only for comparing previously defined aggregates in terms of different size measures, but also, as we have described earlier, an analyst may use any of the axes for defining aggregates. He/she may not only specify arbitrary intervals of attribute values by segmenting the horizontal axis but also divide the reference set into subsets of desired sizes by splitting the vertical axis. If there are multiple vertical axes, the analyst may use any of them. Thus, in our example, it is possible to divide the set of districts into subsets so as to obtain desired proportions between the numbers of districts in the subsets, or between the total populations, the total areas, or the total numbers of highly educated people.

A few such divisions according to different criteria are shown in Fig. 4.71. At the top right, the set of districts is divided into three subsets with approximately equal total population. This corresponds to interval breaks

at 40.75 and 58.29. At the bottom left, three groups of districts have approximately equal total areas. The interval breaks are at 37.38 and 47.19. At the bottom right, the set of districts is divided into two subsets such that each subset contains approximately half of all people in the country who have high school education. We can see that 50.5% of such people live in 9.8% of the districts of Portugal, in which the percentage of employment in services is more than 61.21. These districts occupy 8.2% of the total area of the country; their total population constitutes about one-third of the population of the whole country.

In this example, we have successfully used a cumulative-curve display for multiple purposes:

- comparison of relative sizes of aggregates defined by dividing attribute value ranges into intervals;
- analysis of various quantitative characteristics of these aggregates;
- detecting links between attributes;
- constructing aggregates with desired relative sizes and properties.

#### ***4.5.4.5 Visualisation and Use of Positional Measures***

We have demonstrated that certain attributes may be used for defining problem-specific measures of aggregate size. The resulting characteristics of the aggregates may be visualised and analysed using tools designed to represent aggregate sizes, such as histograms, segmented figures, and cumulative curves. However, not all attributes can be treated as measures of size. We have already discussed various methods of combining individual attribute values into characteristics of aggregates. Only characteristics derived by means of summing attribute values can be regarded as sizes and treated in the same way as counts of aggregate members.

We have mentioned that characteristics of aggregates, except for counts, have the same nature as the original attribute values that they were derived from, and, in principle, they can be visualised in the same way as these original values. In fact, by aggregating data, one obtains a new dataset, in which the reference set is the set of aggregates and the attributes are various integrated characteristics of these aggregates. Hence, one needs to choose appropriate display dimensions and visual variables to represent the new references (i.e. the aggregates) and the corresponding characteristics. In so doing, one should adhere to the basic principles of visualisation.

Of all the possible variants of combined characteristics of aggregates, we would like to consider especially how positional measures, such as the median and quartiles, can be visualised and used in data analysis. There are two reasons for this. First, John Tukey, the founder of exploratory data

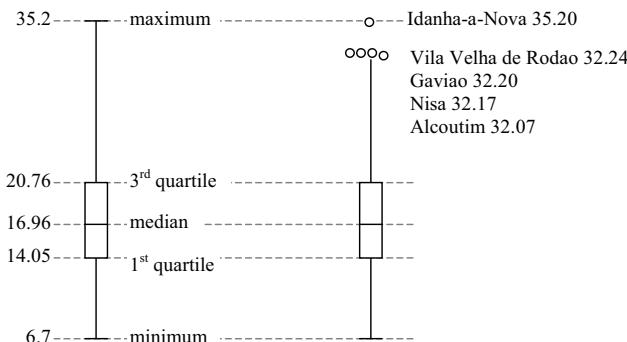
analysis as a research field and a philosophy, paid great attention to positional measures. He suggested a particular visualisation technique for these measures and described how to use it in data exploration (Tukey 1977). Second, positional measures have rather interesting properties: they simultaneously characterise a set as a whole, divide it into subsets in previously known proportions, and characterise these subsets. Let us explain this point with an example.

The median of the attribute “% 65 or more years” (the percentage of people aged 65 or more years) for the set of districts of Portugal is 16.96. This means that 50% of the districts have values of this attribute that are less than or equal to 16.96, and that the values in the remaining 50% of the districts are greater than or equal to 16.96. In other words, the median divides the set of districts in the ratio 50 to 50. Hence, we know the relative sizes of the aggregates and the value range of the attribute in each of them: from the minimum (6.70) to 16.96 and from 16.96 to the maximum (35.20). The difference between the lengths of these ranges (10.16 versus 18.24) shows that the variation of attribute values is much higher in the second half of the set of districts than in the first half. Since the median is closer to the minimum than to the maximum, it may be guessed that there is a tendency towards lower rather than higher proportions of elderly people. Just one number gives us quite a lot of information about the whole set and two particular subsets of it.

Adding two other numbers, the first quartile, 14.05, and the third quartile, 20.76, increases further our knowledge about this dataset. Thus, we know that the set of districts is divided into quarters, and we know the range of attribute values in each quarter. We also know that the values in 50% of the districts (the second and third quarters taken together) range over in a rather narrow interval, from 14.05 to 20.76, as compared with the whole range from 6.70 to 35.20, and there is a slight tendency in this subset towards smaller values (the median, 16.96, is closer to 14.05 than to 20.76). In 75% of the districts, the proportion of elderly people is not more than 20.76, i.e. even less than the number midway between the minimum and the maximum of the attribute (which is 20.95). The fourth quarter is characterised by high variation of the values, from 20.76 to 35.20.

Tukey invented a type of graph called the “box-and-whiskers plot”, which shows the positions of the median and quartiles of a set of numbers in relation to its extreme values and thereby provides a visual summary of the set conveying the information of the kind discussed above. The appearance and construction of a box-and-whiskers plot is demonstrated in Fig. 4.72 using the example of the proportions of elderly people in the districts of Portugal.

The minimum, maximum, median, and quartiles of a set of values are represented by positions in either the horizontal or the vertical display dimension. In Fig. 4.72, the vertical dimension is used for this purpose. The positions are linked visually by special graphical elements: lines between the minimum and the first quartile and between the maximum and the third quartile (called “whiskers”), and a rectangle (“box”) between the first and the third quartile, which is divided by a small line indicating the position of the median (see Fig. 4.72, left). As a result, a particular figure appears, and the shape of this figure reflects the distribution of values within the set. Tukey also suggests a modification of this visualisation: the extreme values and possibly some other values close to the extremes may be identified on the plot, for example as is shown on the right in Fig. 4.72. The corresponding “whiskers” are in this case cut off at the innermost values identified.



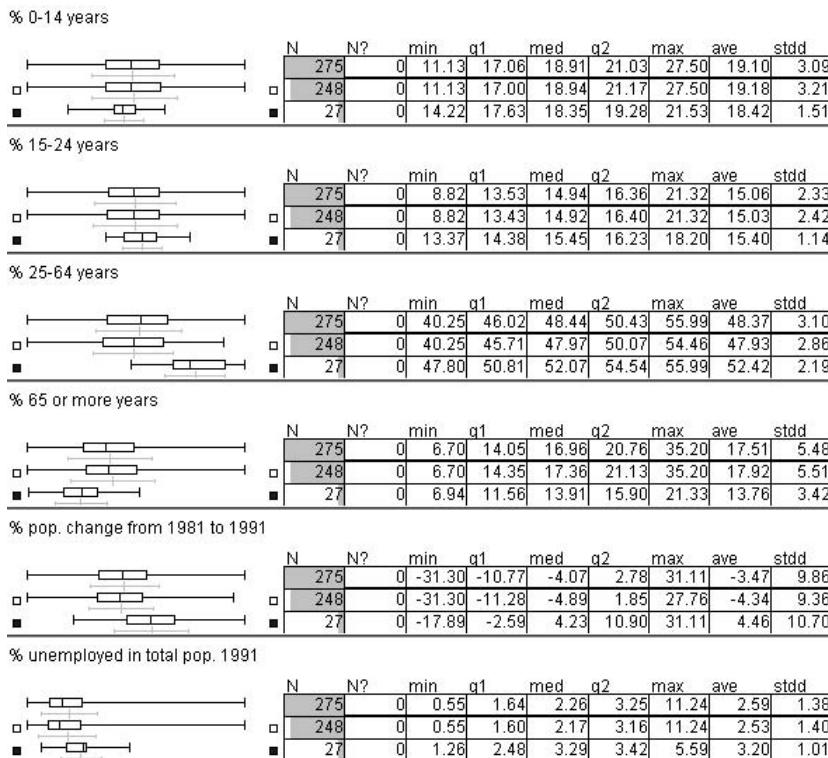
**Fig. 4.72.** Construction of a box-and-whiskers plot

Box-and-whiskers plots can be used not only for analysing a single set of numbers but also for comparing different sets. In particular, an analyst may compare characteristics of different subsets of references in terms of the same attribute, as well as compare the distributions of values of different attributes over the same (sub)set of references.

In Fig. 4.73, combined characteristics of three sets of districts of Portugal, in terms of six different attributes, are represented in a numeric and a graphical form. For the graphical representation of the medians and quartiles, box-and-whiskers plots are used. This time, these measures are represented by positions in the horizontal dimension.

The display is divided into six sections, in accordance with the number of attributes involved. In each section, characteristics of three sets of districts are shown: the set of all districts of Portugal, the subset of districts with a percentage of people employed in services below 61.21, and the

subset of districts with more than 61.21% of working people employed in services. Recall that each of these subsets of districts contains in total approximately half of the people with high school education in the whole country. The subsets were defined using a cumulative-curve display; see Fig. 4.71 (bottom right).

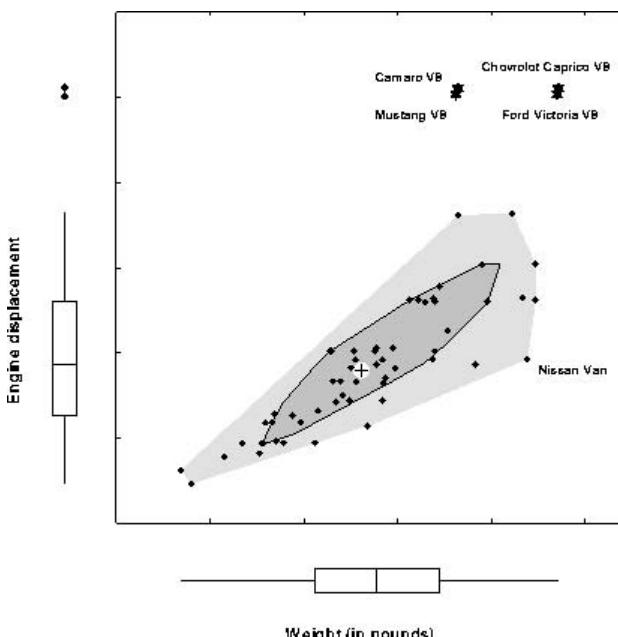


**Fig. 4.73.** Comparison of aggregate characteristics of the whole set of districts of Portugal, the group of districts with less than 61.21% of working people employed in services, and the group with more than 61.21% of such people

The display allows us to compare the age structures of the population, the population changes from 1981 to 1991, and the percentages of unemployed in the whole set of districts and in the two subsets. We can notice that the aggregate characteristics of the subset with less than 61.21% employment in services differ very little from those of the whole set. However, the subset of districts with high employment in services, which consists of only 27 districts, has quite different characteristics, especially with respect to the attributes "% 25–64 years" (the percentage of people aged from 25 to 64 years) and "% 65 or more years" (the percentage of people

aged 65 or more years). The percentages of people aged from 25 to 64 years are mostly higher in this subset than in the entire set and in the other subset. This is indicated quite clearly by the shift of the corresponding box-and-whiskers plot (the one at the bottom of the section representing the attribute "% 25–64 years") very much to the right in relation to the boxes for the other two sets. In contrast, the plot for the attribute "% 65 or more years" is shifted to the left. This means that the proportion of elderly people in this set of districts tends to be relatively low. Similar shifts can be observed for the attributes "% pop. change from 1981 to 1991" (the change in the population from 1981 to 1991 as a percentage of the population in the year 1981) and "% unemployed in total pop." (the percentage of unemployed people in the total population). In both cases, there are shifts to the right; hence, the values of both attributes tend to be relatively high. The precise values of the aggregate characteristics can be seen in the tables to the right of the plots.

In this example, we have compared characteristics of aggregates in terms of several attributes using multiple box-and-whiskers plots, each plot corresponding to a single reference (sub)set and a single attribute.



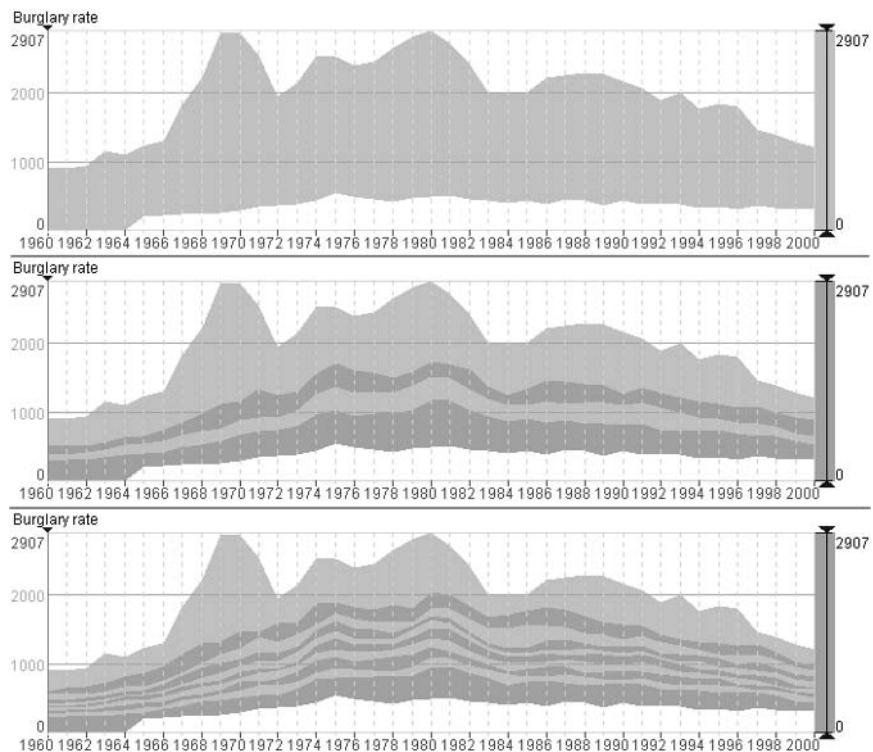
**Fig. 4.74.** A bagplot representing data about the weight and engine displacement of 60 car models (from Rousseeuw et al. (1999)). Reprinted with permission from *The American Statistician*. Copyright 1999 by the American Statistical Association. All rights reserved

A generalisation of the idea of a box-and-whiskers plot to two or more attributes is a collection of convex hulls in a two-dimensional or multidimensional space. The hulls are constructed in such a way that the outermost hull contains the whole set of references, and each successive hull contains fewer references by about 25%. An example of a bivariate box plot (i.e. a display constructed for two attributes) can be found in Wilkinson (1999). Since the construction of hulls is computationally intensive, a simplified version of the bivariate box plot, the *bagplot*, has been suggested in Rousseeuw et al. (1999). One of the example displays from that paper is reproduced here as Fig. 4.74. The main components of the display are a *bag* that contains 50% of the references (in this example, the references are different car models), a *fence* that separates the bulk of the set from outliers, and a *loop* indicating the references outside the bag but inside the fence. In Fig. 4.74, the bag is the polygon with a darker interior, and the loop is shown in a lighter shade. The four points in the upper right corner represent outliers.

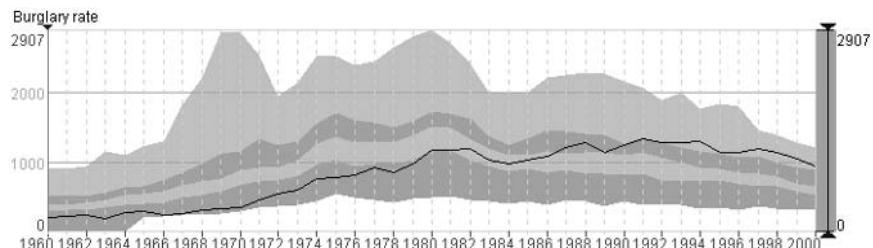
The box-and-whiskers plot (and its bivariate and multivariate extensions) is certainly not the only possible way to visualise medians and quartiles. Let us consider the display at the top in Fig. 4.75. This display was obtained from a time graph of the burglary rates in the states of the USA (see Figs 4.3, 4.30, and 4.47) by drawing an “envelope”, that is, a polygon enclosing all the lines, and then removing the lines. The envelope represents an aggregate characteristic of the set of states: for each year, it shows the range of burglary rates for the whole set.

In the middle of Fig. 4.75, the representation of the value range for each year is complemented by showing the median and quartiles. The positions of the corresponding positional measures in consecutive years are connected so that the original envelope is divided into four polygons. The polygons are shaded using alternating light and dark shades of grey, which makes them clearly visible and distinguishable. We would like to warn readers against considering these polygons as envelopes, i.e. containers of certain subsets of lines. They are just indicators of the positions of each year’s median and quartiles. As is shown in Fig. 4.76, an individual line may cross the boundaries of the polygons. Perhaps it would be less misleading to mark these positions somehow without connecting them, but such a display would be much more complex: instead of just four polygons (each polygon is perceived as a single figure), we would have  $3 \times 41 = 123$  separate marks indicating the median and two quartiles for each of the 41 years from 1960 to 2000. The four polygons are much easier to perceive.

As soon as we have understood how the polygons are constructed and what they mean, we can use them for data analysis without allowing their appearance to mislead us.



**Fig. 4.75.** This display represents an aggregation of time-series data. At the top, the ranges of attribute values in each year are shown. In the middle, the boundaries of the polygons indicate the positions of the median and quartiles in each year. The display at the bottom represents the deciles for each year



**Fig. 4.76.** Individual lines may cross the boundaries of the polygons indicating the positions of quantiles. Here, the line corresponds to the state of Mississippi

Thus, we have a summarised characterisation of the countrywide situation with regard to the burglary rates in each particular year and can compare situations in different years. We can also get an idea of the overall trend in the burglary rates over the country during the whole period from

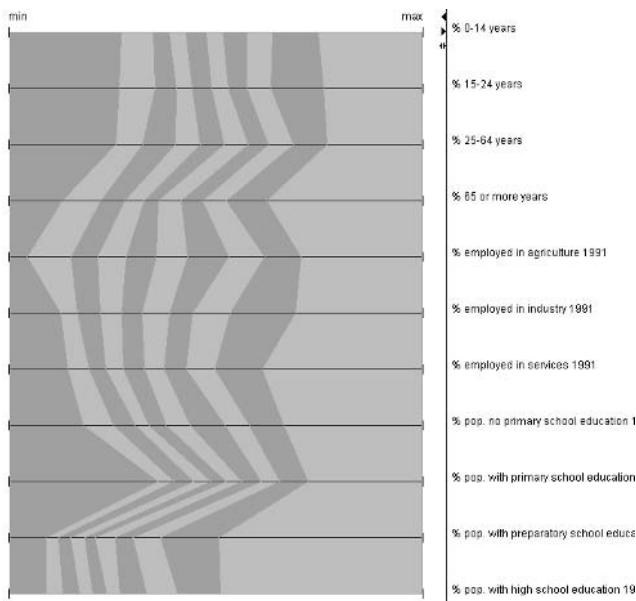
1960 to 2000 or any of its subintervals. An increase in the median and quartile values indicates an overall increasing trend in the burglary rates, and similarly for a decrease. For example, a clear decreasing trend can be observed in the interval from 1991 to 2000. Moreover, using the properties of positional measures, we can particularise this observation by giving some numerical estimates: in 1991, more than half of the states had burglary rates over 1000, whereas in 2000, the burglary rates in more than 75% of the states were below 1000. We can easily see the period of the highest burglary rates, from 1977 to 1982, when the rates in at least 75% of states were over 1000. The synchronous peaks in the values of the median and the quartiles in 1975 and 1980–1981 may also be worth attention, as well as the rather steep decrease from 1981 to 1984.

At the bottom of Fig. 4.75, the display represents the deciles of the burglary rates in each year. This allows us to refine our observations based on the representation of the quartiles. For example, we can see that 90% of all states had burglary rates below 1000 in the year 2000, and that more than 80% of the states had rates over 1000 in 1980 and 1981. The line connecting the ninth deciles shows us the maximum value in each year for 90% of the states. From the width of the upper polygon, we can see in which years the greatest outliers occur. In general, any positional measures may be represented in such an aggregate time graph and used for data analysis on the overall level. It is good when a tool allows an analyst to choose the measures that he/she wishes to use.

Analogously to how the time graph display has been modified to show aggregate characteristics, the parallel-coordinates display can also be revised. Figure 4.77 demonstrates a transformed parallel-coordinates display for 11 attributes characterising the districts of Portugal. Instead of lines for individual districts, the display shows the relative positions of the deciles for each attribute. Of course, any other positional measures can be shown instead of the deciles. The representation of such measures is analogous to that in an aggregate time graph. The positions of the corresponding quantiles on adjacent axes are connected by lines, which form polygon boundaries. As in a time graph, the polygons are not envelopes; they are drawn just to make the display simpler.

Nevertheless, it should be admitted that drawing such polygons in a parallel-coordinates display makes much less sense than in a time graph. In a case of time-series data, we are dealing with values of the same attribute that change over time. It is therefore quite reasonable to observe how the median or ninth decile of this attribute changes over time. Linking the positions of medians, etc. helps us to do this observation. In contrast, the parallel-coordinates display is meant for arbitrary attributes, and the same relative positions on different axes may correspond to totally different val-

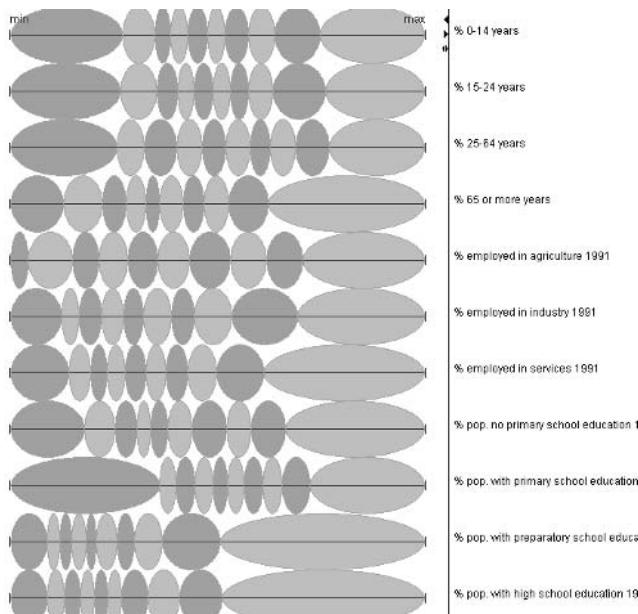
ues. Thus, one of the attributes may vary from 0 to 1 and another from 10 to 100 million. Therefore, connecting positions of the medians or other quantiles of these attributes is hardly useful.



**Fig. 4.77.** Representation of aggregate characteristics (deciles) in a parallel-coordinates display. On the axis for each attribute, the positions of the deciles of its value set are indicated. The positions of the corresponding deciles on adjacent axes are connected so that the lines form polygon boundaries

Therefore, we suggest another representation for positional aggregate measures in a parallel-coordinates display. The idea is demonstrated in Fig. 4.78. Each axis is transformed into a “necklace” – a sequence of ellipses drawn between the positions of successive quantiles. We believe that such shapes are perceived better than just ticks and allow easier identification of the quantiles. In fact, if we decided to limit the display to representing only medians and quartiles, we would draw Tukey’s box-and-whisker plots instead of ellipses. The advantage of the ellipses is that they can be used for any quantiles.

It should be borne in mind that a display such as that in Fig. 4.78 provides us with a visual summary of the distribution of the values of several attributes over a set but says nothing concerning the distribution of value combinations. Each attribute is represented independently of the others. We can compare the distributions of the values of different attributes but cannot detect any relations between the attributes.



**Fig. 4.78.** An alternative representation of positional measures in a parallel-coordinates display. Instead of connecting positions of corresponding quantiles on neighbouring axes, oval shapes are drawn between positions of neighbouring quantiles on the same axis

Besides analysing and comparing value distributions of different attributes over a single reference set, a “necklace display” may be used for comparing aggregate characteristics of several sets. For a convenient comparison, the representations of different sets may be overlaid in the same display, as is shown in Fig. 4.79C. Here, we have “necklaces” with “beads” of different colours. Grey “beads” correspond, as in Fig. 4.78, to the whole set of districts of Portugal; blue ellipses correspond to the subset of districts with a significant population decrease from 1981 to 1991 (i.e. with a population change between  $-31.3\%$  and  $-3\%$ ); and red ellipses correspond to the subset with a population increase of at least  $3\%$ . For each subset, the ellipses indicate the positions of the deciles, i.e. each ellipse includes 10% of the districts of the respective set. The interiors of the first and the last ellipse of each subset are not coloured; hence, the colouring indicates the positions of the central 80% of the values and helps us to disregard outliers. The vertical diameters of the ellipses are proportional to the sizes of the respective sets. Thus, we can see that the set of districts with a population decrease (blue) is about twice as large as the set with a population increase (red) and about half as large as the whole set of districts (grey).

By looking at the relative positions of the “beads” of different colours in the same “necklace”, we can compare aggregate characteristics of the groups of districts. Thus, we can note that the red ellipses on the axes for the attributes “% 0–14 years”, “% 15–64 years”, and “% 25–64 years” are shifted to the right in comparison with the blue ellipses. This means that the values of these attributes tend to be relatively high in the districts with a population increase. A remarkable difference can be observed on the axis for the attribute “% 65 or more years”: the districts with a population increase are characterised by much smaller values than are the districts with a decrease. More precisely, the values in 80% of the districts with a population increase are smaller than the values in 90% of the districts with a population decrease.

Concerning the employment structure, the most significant difference is in the percentage of people employed in agriculture: nine of the ten red ellipses, which represent 90% of the set of districts with a population increase, are compressed into the left one-third of the corresponding axis. This means that the percentage of people employed in agriculture is mostly low in this group of districts. The tenth ellipse stretches across the remaining two-thirds of the axis, thus indicating the presence of one or more outliers. In the group of districts with a population decrease, the values of the attribute “% employed in agriculture” are spread approximately evenly along the length of the axis. In comparison with the grey “beads”, the blue ones are slightly shifted to the right, indicating a tendency to higher values in the group of districts with a population decrease than in the entire set in general.

Concerning the educational level, we can detect that the districts with a population increase have significantly lower percentages of people without primary school education than do the districts with a population decrease. The percentages of people with primary school education are nearly the same. A tendency to higher values in the group with a population increase is observed for the percentages of people with preparatory and, especially, high school education. Hence, the educational level is, in general, higher in the districts with a population increase than in the districts with a population decrease.

We should stress that in this analysis, each attribute has been considered individually, independently of the others. This sort of visualisation does not say anything to us about the distribution of combinations of values of different attributes or about relations between the attributes. We believe that overlaid bagplots or bivariate box plots for two or more reference sets would be suitable for comparison of these sets in terms of combinations of values of two attributes; however, we have no tool at our disposal that would allow us to experiment with such a visualisation.

#### 4.5.4.6 Spatial Aggregation and Reaggregation

Up to now, we have considered methods for data aggregation and for visualisation of aggregate characteristics in which space is irrelevant. Although we used spatially referenced datasets in our examples, we did not take into account the spatial nature of the references, but dealt with them as with a (statistical) population. Now we would like to give some examples of the construction and visualisation of spatial aggregates.

In Fig. 4.80, we can see a map of the Marmara region of Turkey, in which circles indicate the locations of earthquakes that occurred during the period from 1 January 1976 to 30 December 1999. The total number of earthquakes represented on the map is 10 560; therefore, it is not surprising that the circles overlap greatly and clutter the map. Such a visualisation can hardly be used for any analysis.

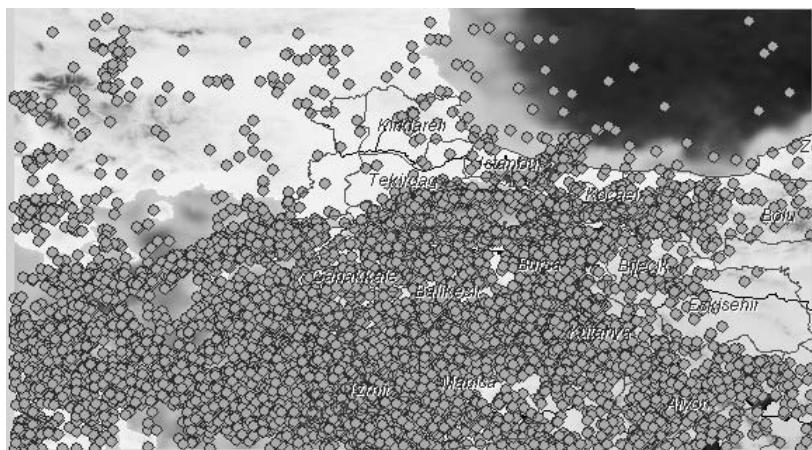
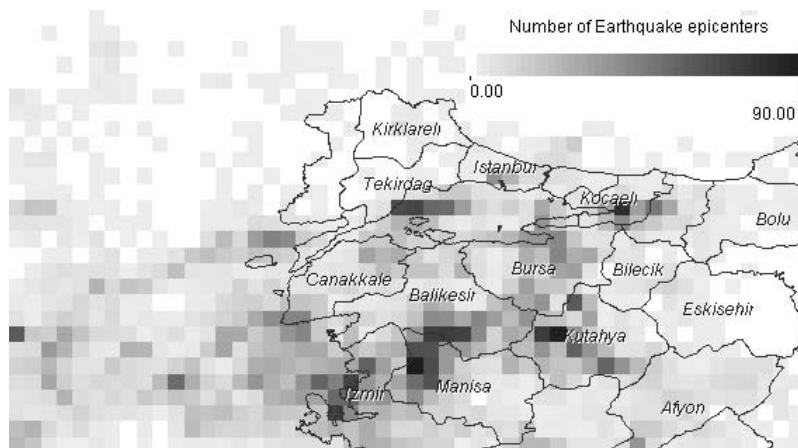


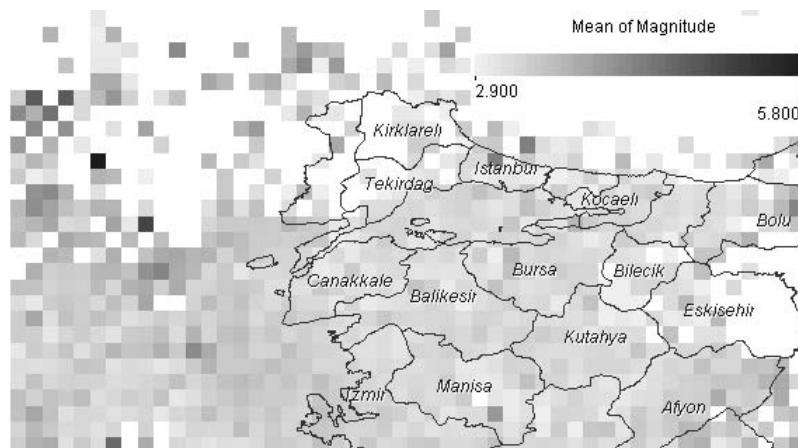
Fig. 4.80. Earthquake occurrences in the Marmara region of Turkey

Since the main problem with this display is too large an amount of individual data items represented on it, this is a typical case where one should try data aggregation. As we are primarily interested in the spatial distribution of the earthquakes (i.e. where earthquakes occur most frequently, where the strongest earthquakes take place, etc.), we need to aggregate the data according to spatial criteria. We can do this, for example, by introducing a regular rectangular grid to cover the territory under study. Then, for each grid cell, we can obtain and visualise various aggregate characteristics of the earthquakes whose locations fit inside that cell. Thus, the map in Fig. 4.81 visualises the number of earthquakes in each cell. The earthquake counts are represented by shading the cells so that the degree of darkness is proportional to the number of the earthquakes in a cell. The darkest spots

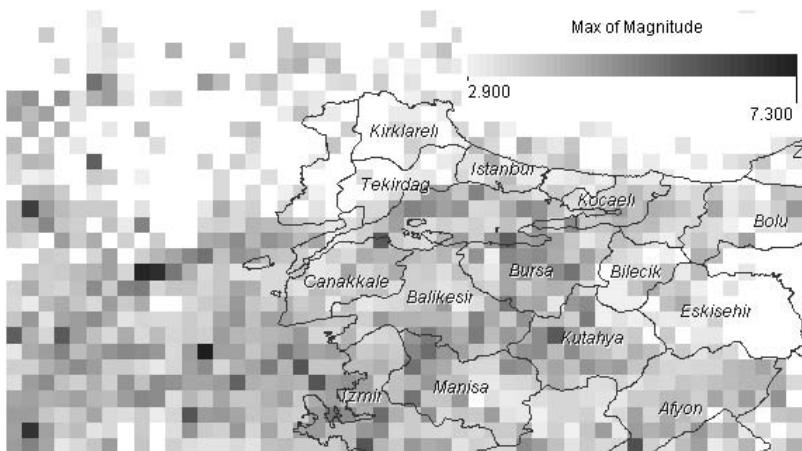
correspond to the places of most frequent earthquake occurrence. We can detect some features of the spatial distribution of the earthquakes: a diagonal belt of high earthquake frequency stretching between the provinces of Izmir and Kocaeli; a smaller strip with the same orientation to the north of it; a triangle-shaped area of high earthquake concentration in the province of Kutahya, which adjoins the diagonal belt; and so on. The diagonal formations stretch to the west beyond the territory of Turkey (i.e. the part of the map where the province boundaries are shown), into the Aegean Sea, where the bigger belt changes its orientation nearly orthogonally. Nothing like this could be observed in the original display representing individual earthquakes.



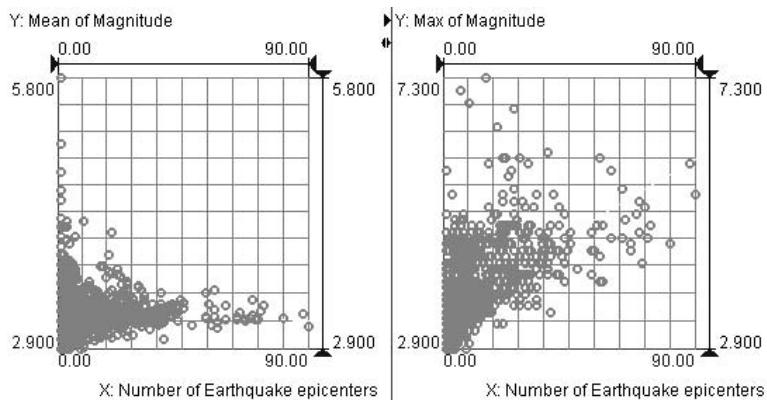
**Fig. 4.81.** Earthquakes aggregated: Visualisation of earthquake counts



**Fig. 4.82.** Earthquakes aggregated: Visualisation of the mean magnitudes



**Fig. 4.83.** Earthquakes aggregated: Visualisation of the maximum magnitudes



**Fig. 4.84.** These scatterplots allow us to investigate whether there are any links between the number of earthquakes and the mean and maximum earthquake magnitude

In Fig. 4.82, another aggregate characteristic of the grid cells is represented: the mean magnitudes of the earthquakes. The representation method is the same as was used for the earthquake counts. By comparing the maps in Figs 4.81 and 4.82, we can observe that the places with the most frequent earthquake occurrences have rather low mean earthquake magnitudes. Hence, most of the earthquakes that occur there are weak.

To estimate the possible danger from earthquakes in various places, it may be not so appropriate to analyse the mean magnitude as the maximum magnitude. The map in Fig. 4.83 represents the maximum earthquake magnitude for each grid cell. We can see a certain similarity between this map and the map in Fig. 4.81, in which the earthquake counts are repre-

sented. We can trace nearly the same diagonal shapes. The presence of a link between the number of earthquakes and the maximum magnitude can be also seen from the scatterplot on the right in Fig. 4.84. The scatterplot on the left shows us that the mean earthquake magnitude is not correlated with the number of earthquakes.

Returning to the maps in Figs 4.81 and 4.83, we can see that the shades within the territory of Turkey are lighter in the map representing the maximum magnitudes than in the map representing the counts, while in the sea the shades are darker. Unfortunately, we cannot be sure that the earthquakes in the sea are stronger but less frequent than on land. It may be that not all earthquakes that occurred in the sea were properly registered and recorded in the data.

Anyway, let us assume that we are concerned first of all about the possible dangers on land and need to identify places where strong earthquakes are likely to happen. To do this more conveniently, we can apply the visual comparison technique for display manipulation, with a diverging colour scale used to differentiate values below and above a selected reference value. In Fig. 4.85C, this reference value is 4.<sup>16</sup> Values below 4 are represented by shades of blue, and values over 4 by shades of brown. We can now easily see which areas can be judged as relatively safe and which are rather hazardous.

In this example, we have demonstrated how data referring to discrete spatial locations can be aggregated, and how aggregate characteristics can be represented using the usual cartographic visualisation methods. Now we shall consider another example, where the data characterise a continuous spatial phenomenon and are specified in the form of raster. An example visualisation of raster-based data has been considered earlier (Fig. 4.60C).

A problem with raster data is that it is difficult to visualise and analyse several attributes simultaneously. Thus, in our example dataset concerning European forests, we have data on the proportions of different types of forest: coniferous, broadleaved, mixed, and other wooded land. In addition, we have the relief data shown in Fig. 4.60C. All these data are specified in a raster format. We would like to analyse these data jointly in order to understand the variation of the land cover and forest structure over Europe, and its relation to the relief. However, when we apply colour encoding to

---

<sup>16</sup> We assume here that earthquakes with a magnitude over 4 may be strong enough to cause serious damage, since many buildings in Turkey are rather old and/or are not constructed adequately for the seismic conditions. In any case, this is only an example to show how such data can be explored. If the real danger threshold were different from 4, the reference value could easily be changed.

visualise each of these attributes, we can see only the visualisation that is on top of the others. Using colours to represent value combinations rather than individual values would require many different colours and make the image very difficult to interpret. At the same time, the areas on the screen corresponding to the raster pixels are usually too small to allow one to fit in any diagrams representing values of several attributes. Moreover, the number of pixels in a raster dataset is typically very large; therefore, an attempt to visualise the corresponding values on non-cartographical data displays, such as scatterplots and parallel-coordinates displays, would result in a severe overlap of graphical marks. Hence, this is again a case where one should try spatial data aggregation.

As we did before with the earthquakes, we shall use a coarse grid to aggregate the values of the attributes. For each cell of this grid, various aggregate characteristics can be computed from the original attribute values specified for the raster pixels fitting into this cell: the mean and standard deviation; the minimum, maximum, and their difference; the median and quartiles; the mode; and the sum of the values. As a result, we obtain a collection of attributes characterising the grid cells. These attributes can be visualised and analysed like usual attributes referring to compartments of a territory.

In Fig. 4.86C, we see the mean proportions of forests of different types, specifically coniferous, broadleaved, and mixed forests, and other wooded land, represented by pie charts placed inside the grid cells. The sizes of the pies are proportional to the sums of the values for the corresponding cell and hence indicate approximately the proportion of forest-covered land in each cell. We can see, for example, that there is much more forest in the north-east of Europe than in other places, and that this forest is predominantly coniferous. In the south of Europe, other wooded land prevails. We can also relate the amount and structure of the forest to the relief, which in this case is quite clearly visible beneath the pies. Thus, we can see that in central and southern Europe higher altitudes co-occur with more forest, while in the north an opposite relation can be observed.

A sophisticated tool for aggregation of raster data may allow the analyst to change the resolution of the aggregating grid interactively, which results in automatic reaggregation of the data and modification of the visualisation so that the new aggregated data are represented using the same visualisation technique as before. For example, the map shown in Fig. 4.87C is a result of modifying the map in Fig. 4.86C after the resolution of the aggregating grid has been increased. We can now verify and refine our observations made with the coarser grid. Actually, all our previous observations remain valid, and the increased resolution can only add some fine detail to our overall understanding of the variation of forest cover over Europe.

The aggregated data can also be represented in various non-cartographic displays. When the resolution of the grid changes, these displays can also be automatically updated to represent the new data.

#### **4.5.4.7 A Few Words About OLAP**

A brief consideration of OLAP (online analytical processing) tools seems appropriate here, since these tools involve data aggregation and are currently rather popular. OLAP tools typically function on top of relational databases and data warehouses. They are based on the idea of multiple hierarchical dimensions in data. *Dimensions*, in OLAP terminology, correspond to our notion of referrers, while what we call attributes are termed *measures*. Individual items within dimensions (i.e. values of referrers, in our terms) are called *members*. Dimension members may be organised in hierarchies. For example, in a temporal dimension, days may be grouped into months, months into years, and years into decades. In a spatial dimension, countries may consist of provinces, which, in turn, are divided into districts, and the districts into municipalities. A database of nature observations may include a dimension in which species of plants and animals are grouped into genera, genera into families, and so on. OLAP tools perform data aggregation according to these hierarchies: attribute values for members higher in the hierarchy are derived by aggregating the values associated with the subordinate members. The terms *roll up* and *drill down* denote increasing and decreasing the level of data aggregation.

OLAP tools are rather efficient in dealing with very large amounts of data. The results of data aggregation provided by OLAP tools may be visualised just like the usual kinds of data, i.e. attribute values associated with references. Therefore, OLAP is very suitable as an underlying technology for “multiscale data visualisations”. Such visualisations present the data at different levels of abstraction as the user zooms and pans; details and examples may be found in Stolte et al. (2002). A practical approach to the incorporation of the power of OLAP technology into a software system for exploratory data analysis is described in Hernandez et al. (2005). Hernandez et al. express the idea that the system can intelligently control the amount of information loss as the user navigates through different aggregation levels, for example by substituting one visualisation technique for another.

The aggregation operations typically offered by OLAP tools are the count, the mean, the minimum, the maximum, and the sum of values. Not all of these tools support positional measures. This has provoked a sceptical attitude to the entire OLAP technology on the part of some researchers in exploratory data analysis, especially statisticians (see, for example, Wil-

kinson 1999). The rational part of their criticism emphasises the importance of positional measures and necessity to involve these measures in data analysis. It is insufficient to base an analysis only on the most common statistics, such as means. The arithmetic mean can be viewed as a valid characteristic of an aggregate only when the distribution of the original attribute values is close to normal. When the distribution is skewed or there are outliers, the mean value may be completely misleading. The properties of the distribution (i.e. whether it is normal or skewed and whether any outliers exist) can be judged, from the relative positions of the quartiles with respect to the minimum and maximum values, for example, or from other positional measures.

For a data explorer, this entails two implications. First, the explorer should choose an OLAP tool that can provide various positional measures of data aggregates. Second, the explorer should use positional measures in the data analysis, in particular, for checking the validity of using a specific aggregation. If the distribution of the values inside the aggregates deviates significantly from normal, it is appropriate to try other aggregations.

Another criticism of OLAP tools is that they do not properly support work with qualitative attributes. Generally, there are many more tools and techniques intended for numeric data than tools and techniques capable of working with nominal and ordinal attributes. Hence, researchers in data-analysis-related areas, as well as tool designers, still have something to think about.

#### **4.5.4.8 Data Aggregation: a Few Concluding Remarks**

We have paid so much attention to data aggregation because of its high importance in exploratory data analysis. Aggregation is indispensable when the amount of data to be explored is large, which is almost always the case in real situations. It is therefore not surprising that research in data visualisation and EDA is currently strongly imbued with data aggregation. This has materialised in the invention of new aggregation-based tools and in the modification of traditional visualisation techniques to represent aggregated rather than atomic data. Contemporary tools for EDA are characterised by high user interactivity, which allows an analyst to choose and dynamically change the level of aggregation (i.e. how large the aggregates are), the method of aggregation (i.e. how individual items are grouped into aggregates), and the functions for deriving characteristics of aggregates from those of their members (i.e. sums, ranges, or various statistics).

When applying data aggregation, one should be very cautious about averaging, which can often produce meaningless or misleading figures. Thus, an average of an extremely low and an extremely high value is a quite or-

dinary value. For example, the mean surface temperature on the Moon may seem quite comfortable, but the actual temperature ranges from  $-230^{\circ}\text{C}$  to  $130^{\circ}\text{C}$ , and this should be taken into account in designing clothes for astronauts. The mean weight of a fruit in a basket filled with apricots and one watermelon is also not a very useful aggregate characteristic. In this case, even knowing the minimum and maximum weights may be insufficient. A general recommendation is to use any available means for investigating the distribution of attribute values over each aggregate. Suitable tools are histograms, cumulative curves, and positional measures, i.e. the median, quartiles, deciles, and so on. When any of these tools shows that the distribution is skewed or that there are outliers, averaging is inappropriate. If some aggregate characteristic is still needed for further analysis, one may try either to reaggregate the data so that the distribution of attribute values in each aggregate becomes close to normal (in this case, the use of the mean is valid) or to characterise the aggregates by positional measures, for example, medians or third quartiles, additionally to the minimum and maximum values.

Of course, it is not always possible to investigate the properties of the distribution of attribute values in each aggregate. For example, the aggregates shown in Figs 4.86C and 4.87C are too numerous to be analysed in detail. In such cases, it is recommended that one should verify observations and conclusions made on the basis of aggregated data by trying different aggregation levels with the same data. It is also meaningful to check whether the observations remain valid after means are replaced by medians.

In general, data reaggregation is always appropriate. A single, rigid data aggregation is inadequate for comprehensive data analysis. Varying the level of aggregation and changing the method of aggregation (e.g. choosing other data components as the basis for the aggregation, or altering the division of attribute values into subsets or intervals) are necessary for gaining a proper understanding and avoiding hasty conclusions and wrong decisions. Hence, it is important to have highly interactive tools for data aggregation, that provide sufficient flexibility in defining aggregates and a prompt response in computing and visualising their combined characteristics.

Besides trying different aggregation levels (i.e. aggregate sizes) and different ways of grouping references, it may be appropriate to apply a kind of adaptive aggregation, which varies the aggregation level depending on the internal variability inside each aggregate. Such an adaptive aggregation algorithm could easily be realised for temporally referenced attributes as well as for grid-based spatial data.

### 4.5.5 Recap: Data Manipulation

We have considered data manipulation techniques mainly as subsidiaries of other tools for exploratory data analysis, in particular, visualisation. We mean that data manipulation is not used independently and does not, by itself, provide answers to any exploratory questions. Two major purposes of data manipulation have been mentioned:

- To simplify data, to make it easier to perceive from a visual display, and to analyse it. The following operations are possible:
  - reduce noise and discontinuities (smoothing and interpolation);
  - reduce the amount of data under analysis (aggregation and attribute integration);
  - extract characteristic features, e.g. surface topology.
- To enrich data and consider its various aspects in the following ways:
  - involve additional references and estimate the corresponding attribute values (interpolation and extrapolation);
  - consider quantities in relation to other quantities: parts in relation to the total, values per capita, densities per area unit, etc.;
  - compute changes, e.g. in time, or deviations from standard values such as overall means;
  - standardise values to achieve comparability of several attributes.

In all our examples of data transformation, we have used visual displays to view the results and extract information from them.

The results produced by many transformation techniques depend on certain settings, or parameters, which may be specified more or less arbitrarily. Some examples of such arbitrary choices are the division of the reference set in data aggregation, the circle radius for neighbourhood-based transformations, the number of neighbours used in data interpolation, and the weights used in computing weighted averages and weighted linear combinations. In such cases, we always recommend that one performs the transformation several times using different settings and compares the results to evaluate their robustness. It may happen that such variation of settings helps one to uncover characteristic features of a phenomenon under analysis and, in this way, increase the knowledge about the phenomenon. However, it is generally a bad sign when observations and conclusions obtained from the results of different transformations are too diverse to allow them to be merged into a coherent overall picture (or mental model).

The possibility for the analyst to alter the settings of a method interactively and to observe the effect immediately is extremely supportive to

analysing the sensitivity of a data transformation to variation of the settings. We have demonstrated this possibility in an example where the attribute weights for computing weighted linear combinations of values of multiple attributes were varied. Unfortunately, such a highly dynamic realisation of a data transformation technique is not always achievable. Some data transformations are quite computationally intensive and require substantial time before the results of altering the settings may be observed.

We have devoted much space to data aggregation since it is being used more and more intensively in exploratory data analysis, which is currently being applied to large and very large volumes of data. When using data aggregation, it is important not to be fooled by averaging and not to lose important information. It is highly recommended that one uses positional measures rather than means for characterising aggregates.

Let us now move to the next group of tools on our list, that is, tools for querying and filtering.

## 4.6 Querying

Querying may be defined as a process in which software provides answers to users' questions about data under analysis. Consequently, in discussing query tools, we need to consider the following aspects:

- what types of questions may be supported;
- how questions may be asked;
- how the answers are presented to users.

As we have discussed before, any question (task) may be viewed as consisting of two major parts: a target and constraints. This also applies to questions intended for query tools. However, it should not be thought that query tools can, either actually or potentially, provide answers to all possible questions related to a given dataset (otherwise, no exploratory data analysis would be needed).

In our reasoning concerning possible questions, or tasks, we distinguished elementary questions from higher-level questions, which we called synoptic tasks. Elementary questions relate to individual elements of data, i.e. references and characteristics (attribute values). Synoptic questions relate to reference sets taken in their entirety and to behaviours of characteristics on such reference sets. Query tools are mostly intended to deal with elementary questions, which can be answered by means of searching through the source data. Some query tools can compute certain statistics, such as counts, sums, and averages. In principle, these statistics can be

viewed as giving a very rough approximation of a behaviour, i.e. as a kind of pattern. However, dealing with more precise patterns implies abstraction and generalisation, but query tools do not possess these abilities. Therefore, we have to conclude that query tools, in general, are not suited to answering synoptic questions.

Let us now consider the ways in which users can put questions to a software program.

### 4.6.1 Asking Questions

A classical category of software tools specifically designed to search through data and give answers to various (elementary!) questions is that of database management systems (DBMS). A DBMS is queried by means of a special formal (and therefore machine-readable) language. Currently, there is a standard database query language, called SQL. Since this language is not powerful enough with regard to spatial and temporal data, various extensions to SQL have been developed, and efforts to extend the SQL standard are under way.

Although many people know SQL (or other query languages) and can formulate their questions directly in a machine-readable form, there are many more people who cannot do this. Even for people who know SQL it may be far from fun to use it every time for obtaining any piece of information. Fortunately, there is a range of facilities intended to reduce users' effort and release them from the necessity to learn the constructs of formal languages. Somewhere behind the screen, SQL or something similar may still be used, but users do not need even to guess about this: they communicate with the program through a convenient interface, and it is the business of the program to interpret users' input and translate it into a form understandable by the search engine.

One possible approach is to use a “visual query language”. The basic idea is that the operators and constructs of a formal language are replaced by graphical elements, such as mnemonic icons and receptacles for those icons, and arrows, which can link things or indicate directions. A user expresses his/her information need by building a diagram from these elements. The diagram is then translated into an internal formal representation. Hence, rather than learn the lexis and grammar of SQL or something similar, the user needs to learn the lexis and grammar of the visual language, that is, the meaning of the graphical elements available and the rules for the construction of diagrams from these elements. To reduce the amount of learning needed, developers of visual languages strive to increase the “intuitiveness” of the graphical elements and the construction

rules. Ideally, the user should immediately recognise the meaning of any element and guess how to combine the elements to construct a visual query. To learn about the achievements in this direction, we refer readers to a survey of visual query languages by Catarci et al. (1997).

In general, a visual query language that has the same power and flexibility as the underlying formal language will necessarily have the same complexity. Hence, the advantage of a visual language is just that it is more appealing to a user and, supposedly, easier to learn. To achieve more substantial simplification, one needs to reduce the power and flexibility. One approach is to represent a few typical query types by predefined forms, which may look like texts with blank spaces. The user needs only to fill in these spaces with some particulars, such as attribute names, in order to turn a generic template into a specific information request.

Another approach is taken in a well-known querying technique called Dynamic Query (suggested and elaborated by Ben Shneiderman and his research team; see Ahlberg et al. (1992)). This tool is very easy to use, but it imposes quite serious limitations upon the questions that may be asked. Specifically, using Dynamic Query, one can only ask questions of the kind “What references correspond to the specified attribute values?”<sup>17</sup> Furthermore, not all kinds of attribute may appear in Dynamic Query. The tool allows only attributes with linearly ordered value sets. Some variants of the tool can deal with attributes whose values are strings, but in such a case an alphabetical order is imposed on the set of values.

The user interface of a Dynamic Query-like tool is organised around a set of slider lines or slider bars – interactive devices representing the value ranges of the attributes participating in the building of the query. A possible appearance of a slider line/bar is demonstrated in Fig. 4.88. Of course, the appearance may differ from implementation to implementation.



**Fig. 4.88.** A possible appearance of an element of Dynamic Query used for specifying constraints on the values of a single attribute

The left end of a line or bar corresponds to the minimum value (i.e. the first in the order) of the respective attribute present in the dataset, and the right end to the maximum value (i.e. the last in the order). Each line (or bar) is furnished with a pair of sliders, or delimiters – small devices that can be moved along the line. By setting delimiters in appropriate positions,

<sup>17</sup> Recall that we called such questions “inverse lookup tasks” and represented them by the formula  $?x: f(x) \in C'$  ((3.9) in Chap. 3).

a user may specify the interval of attribute values that the tool must look for.<sup>18</sup> Thus, in Fig. 4.88, the whole slider line corresponds to the number interval from 11.13 to 27.5 – these are the minimum and maximum values of the attribute “% 0–14 years” present in the Portuguese census dataset. With the use of the delimiters (black triangles), the subinterval from 14.90 to 25.04 is specified as a query constraint. This corresponds to the question “Which districts of Portugal have a percentage of children (i.e. people aged from 0 to 14 years) between 14.90 and 25.04?”

The specific features of the user interface of Dynamic Query account for the limitations that it imposes. First, while a line or a bar can serve as a quite intuitive representation of a linearly ordered set of attribute values, it is hardly suitable for representing a set with different properties. Second, since only a pair of delimiters is used, it is possible to specify only one subinterval of attribute values. Thus, it is impossible to express the question: “Which districts of Portugal have a percentage of children below 14.90 or above 25.04?” Furthermore, Dynamic Query does not provide any explicit means to specify how constraints for different attributes are supposed to be combined. For example, the user may have chosen the value interval of the attribute “% 65 or more years” to be from 6.69 (the minimum value available) to 15, in addition to the constraint on the value of the attribute “% 0–14 years”. This may have at least three possible meanings:

- Which districts of Portugal have a percentage of children between 14.90 and 25.04 **and** a percentage of elderly people below 15?
- Which districts of Portugal have a percentage of children between 14.90 and 25.04 **or** a percentage of elderly people below 15 **or both** (non-exclusive “or”)?
- Which districts of Portugal have **either** a percentage of children between 14.90 and 25.04 **or** a percentage of elderly people below 15 **but not both** (exclusive “or”)?

An ordinary query language allows the user to combine query constraints in various ways, whereas Dynamic Query, for the sake of simplicity, assumes that the constraints are always linked by the conjunction “and”.

Hence, the ease of use of Dynamic Query is achieved because of a number of restrictions:

- only one question type, specifically, “What references correspond to the specified attribute values?”;
- only attributes with linearly ordered value sets;

---

<sup>18</sup> Both the appearance and the use of a slider bar/line are similar to those of the display-focusing tool discussed earlier.

- only one value interval per attribute;
- only the conjunction “and” to combine several constraints.

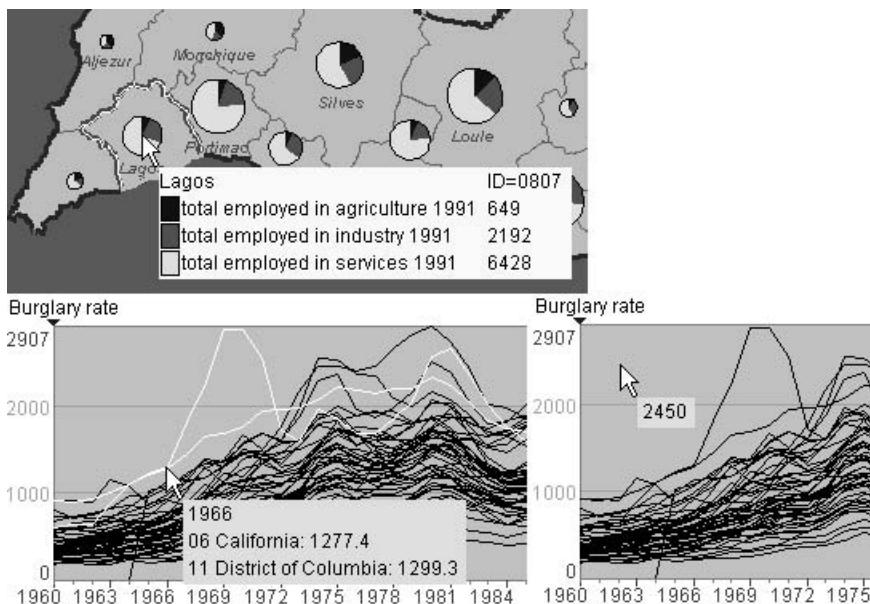
This analysis of the restrictions of Dynamic Query does not mean, however, that the tool is bad and should not be used. On the contrary, the tool is very good and is very well suited to exploratory data analysis, but users need to understand its limitations in order to use it properly. We shall consider the virtues of Dynamic Query, which strongly outweigh the limitations, a little later, but now let us continue with the possible ways of asking questions.

While the use of Dynamic Query is quite simple and does not require much learning, there are even simpler query interfaces, which, naturally, further reduce the flexibility in asking questions. Perhaps the simplest idea is to provide some information when the user just points on a display with the mouse. The mouse gesture is interpreted as the question “What’s this?” The exact meaning of the question depends on the type of display, the information represented in it, and the location of the pointer within the display. An answer to the question may be shown, for example, in a pop-up window or in a specifically dedicated part of the screen.

Figure 4.89 demonstrates how such “querying by pointing” may be done. At the top, there is a fragment of a map of Portugal with a division into administrative districts. The pie charts on the map represent the values of the attributes “Total employed in agriculture 1991”, “Total employed in industry 1991”, and “Total employed in services 1991” for each of the districts. When the user positions the mouse cursor on a district in the map, a pop-up window appears, in which the name of the district, its identifier, and the corresponding values of the three attributes are displayed.

At the bottom left, there is a fragment of a time graph showing the dynamics of the burglary rates in the states of the USA; each polygonal line corresponds to one state. The mouse cursor is located at a point where two lines come close together. The pop-up window shows information related to the cursor position: the corresponding time moment, specifically the year 1966; the names and identifiers of the states represented by the lines; and the burglary rate values for these states in the year 1966. The screenshot at the bottom right of Fig. 4.89 shows the situation when there are no lines near the mouse position on the time graph. In this case, the attribute (i.e. burglary rate) value corresponding to this position is displayed.

The flexibility of a “querying by pointing” tool may be increased by allowing the user to choose what position-related information will be displayed. Thus, when the user points on a map, he/she might get the values of not only the attributes that are currently represented on the map but also other attributes characterising the locations shown on the map.



**Fig. 4.89.** Simple “What’s this?” questions may be asked just by pointing on a display with the mouse

Obtaining information through pointing on a map display is an example of querying spatially referenced data. Of course, this is not the only possible way of asking questions about data that has a spatial component (referred to from now as “spatial data”, for the sake of brevity). Let us briefly explore the possibilities that exist. For this purpose, let us consider the possible types of questions concerning spatial data.

#### 4.6.1.1 Spatial Queries

In Chap. 3, we proposed a general typology of questions (tasks), in which elementary questions are grouped into the categories of lookup, comparison, and relation-seeking. Lookup implies either finding characteristics corresponding to specified references (direct lookup) or finding references corresponding to specified characteristics (inverse lookup). Comparison tasks ask about relations between characteristics (direct comparison) or between references (inverse comparison). Relation-seeking means looking for occurrences of specified relations between characteristics.

Spatial data are data in which at least one of the components, either a referrer or an attribute, has a spatial nature. The presence of a spatial referrer means that the values of some attributes are associated with spatial locations, spatial segments, or objects situated in space, for example rivers or

buildings. The presence of spatial attributes means that certain spatial characteristics, such as location, shape, or extent, pertain to the references. With this in mind, let us try to specialise the general categories of lookup, comparison, and relation-seeking to spatial data.

Direct lookup queries may ask either about various types of characteristics corresponding to spatial references or about spatial characteristics corresponding to arbitrary references. Inverse lookup queries may ask about spatial references corresponding to arbitrary types of characteristics or about arbitrary types of references corresponding to specific spatial characteristics. This gives four different subtypes of spatial lookup tasks. Examples of questions of these subtypes are given in Table 4.9.

**Table 4.9.** Examples of spatial queries of the lookup type

	<b>Direct lookup</b>	<b>Inverse lookup</b>
Spatial referrer	What is the percentage of children in Porto?	Where does the percentage of children exceed 25?
Spatial attribute	Where was the stork Prinzessin on 10 September? What is the distribution area of this plant species? What is the shape of a maple leaf?	Which storks were in the area around Lake Victoria on 10 September? Which rare plant species occur in this area? Of what tree do the leaves have the given shape?

It may be seen that some of the lookup questions specify constraints on values of non-spatial components (specifically, these are the cases of spatial referrer and inverse lookup, and spatial attribute and direct lookup). Hence, the formulation of such questions does not differ from that of questions about non-spatial data. In the other two groups of questions, the constraints may specify particular locations or spatial fragments (e.g. areas) or geometrical properties such as shapes. Probably the easiest and most natural way to specify a location is to point on a display representing the relevant space. For example, if we need information about a certain district of Portugal, we could specify this district by pointing to it on a map of Portugal. Analogously, if we need information concerning some area, or, more generally, fragment of a two-dimensional space, we can outline this fragment on an appropriate display using the mouse or another pointing device. Although this approach cannot easily be extended to three-dimensional space, some solutions can still be found, for example by the use of two-dimensional projections. There are also indirect ways to refer to a specific place (i.e. location or spatial fragment) in a query: in particular,

by specifying the name (if it exists) of the place or its coordinates. Geometrical constraints, such as a particular shape to look for, may also be specified in a direct, visual way, i.e. by drawing, or in an indirect way, i.e. by naming, description in a certain (formal) language, or selection from a restricted list of alternatives. The indirect ways of specifying spatial constraints do not differ in principle from the ways of specifying non-spatial constraints.

Sometimes the user needs to specify spatial constraints in a relative rather than an absolute way, i.e. by indicating a certain relation with respect to a specific place. Look, for example, at the question “Which storks were in the area around Lake Victoria on 10 September?” Here, the spatial constraint, “the area around Lake Victoria”, consists of a reference to a particular place, Lake Victoria, and a specification of a spatial relation, “around”. Here are a few more example questions where spatial constraints are specified in this way:

- What are the percentages of children in the districts bordering on Porto?
- Which plant species occur north of the Arctic Circle?
- Find restaurants within 1 km distance from the city hall.

In each of these questions, the user refers to locations or parts of the space. As we have discussed earlier, a tool may allow the user to do this in a direct way, by pointing or drawing on a representation of the space, or indirectly, by naming or specifying coordinates. Additionally, the user needs to specify a spatial relation, which may be topological (“border”, “inside”, “overlap”, “disjoint”, “between”, etc.), directional (such as “north”, “left”, or “front”), or metric, i.e. related to distances in the space. Spatial query languages (including spatial extensions of standard query languages) include special constructs for expressing such relations. For example, the SQL extension suggested by the OGC (Open Geospatial Consortium)<sup>19</sup> includes the predicates *Disjoint*, *Overlap*, *Touch*, *Cross*, and so on. A predicate denotes an operation with a possible result 1 (true) or 0 (false). This result shows whether the specified relation exists or not.

Using the OGC extension of SQL, the question “Find the names of all countries which are neighbours of the USA in the *Country* table” would be formulated as follows:

---

<sup>19</sup> The home page of the Open Geospatial Consortium has the URL <http://www.opengeospatial.org/>. A description of the suggested SQL extension is available (at the time of writing this book) at the URL <http://www.opengeospatial.org/docs/99-049.pdf> or can be found using links from the page <http://www.opengeospatial.org/specs/>.

---

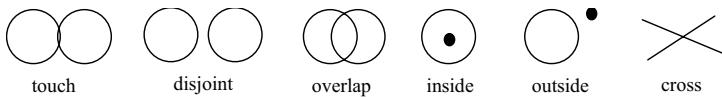
```

SELECT      C1.Name AS "Neighbours of USA"
FROM        Country C1, Country C2
WHERE       Touch(C1.Shape, C2.Shape) = 1 AND
           C2.Name = 'USA'

```

This example is borrowed from the book Shekhar and Chawla (2003), which gives a good introduction into spatial query languages in a dedicated chapter. In the SQL statement above, the expression “*Touch(C1.Shape, C2.Shape) = 1*” means that the relation *Touch* must exist between the shapes (contours) of the countries C1 and C2.

In visual languages for spatial queries, spatial relations may be specified by drawings, gestures (e.g. using the mouse or another pointing device), or appropriate positioning of some predefined icons. For example, the relation “touch” could be expressed by placing two icons so that they touch each other, the relation “overlap” by two overlapping icons, and the relation “disjoint” by two icons with space between them. Another approach is to introduce iconic representations of the spatial relations, such as the examples in Fig. 4.90 representing some topological relations. Visual specification of directional relations could be done, for example, using a widget in the form of an azimuth disc. It seems that distance relations can be specified more naturally through numbers than through pictures, although a slider bar/line could also be quite appropriate.



**Fig. 4.90.** Some possible iconic representations of topological spatial relations

Comparison and relation-seeking tasks deal with relations between characteristics or between references. As we have discussed before, the possible relations depend, in general, on the nature and properties of the set that the characteristics or references belong to. Spatial relations (topological, directional, and metric) are specific for spatial data.

In questions of comparison, which aim at determining what relation exists between some characteristics or references, one needs to specify these characteristics or references in the query constraints. Since spatial relations can exist only when these characteristics or references have a spatial nature, asking questions about spatial relations requires referring to particular places or spatial entities. This may be done as in lookup questions, i.e. either directly, by pointing or drawing on some representation of the space, or indirectly, by naming, description, or coordinate specification. Here are some examples of comparison tasks that target spatial relations:

- What were the relative positions of the storks Prinzessin and Moritz on 10 September?
- Do the migration routes of these storks cross?
- On what side of the river is the castle X situated?
- How far is the hotel Z from the city centre?
- Is the hotel Z within a pedestrian area?

The construction of relation-seeking queries is opposite to that of comparison queries: the user specifies a certain relation, and the goal is to find occurrences of this relation between characteristics or between references. As before, we are interested in spatial relations, which are specific to spatial data. Table 4.10 contains examples of spatial queries of the relation-seeking type. Some possible methods for the specification of spatial relations have been discussed above.

**Table 4.10.** Examples of spatial queries of the relation-seeking type

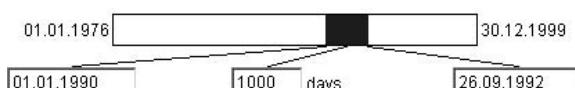
Relation type	Examples
Topological	Find all places where railways cross rivers. Which restaurants are located inside parks or gardens?
Directional	Find chemical plants situated north of populated areas. Which rivers are left tributaries of other rivers?
Metric	Find all towns situated more than 100 km from the nearest railway. Find stops of the same bus with less than 500 m distance between them.

We have recalled the classification of questions into lookup, comparison, and relation-seeking in order to see what questions about spatial data may be potentially asked. We have seen, however, that many such questions do not differ in principle from questions about non-spatial data. For example, to formulate the question “What is the distribution area of this plant species?”, one does not require any specifically “spatial” expressive means. In this sense, a question about the distribution area does not differ from a question about the typical habitat or the temperature interval for this plant species. The specifics of spatial data only come into play in questions where some places (i.e. locations or spatial fragments) and/or spatial relations are specified in the constraints. Hence, the methods for the specification of places and spatial relations are specific to spatial queries.

#### 4.6.1.2 Temporal Queries

Another area of special interest to us is queries about temporal data, i.e. data that have temporal referrers or attributes. Analogously to spatial queries, “temporal specificity” appears only in queries with constraints that specify times (i.e. time moments or intervals) or some temporal relations. Accordingly, we shall consider here the possible methods for the specification of times and temporal relations.

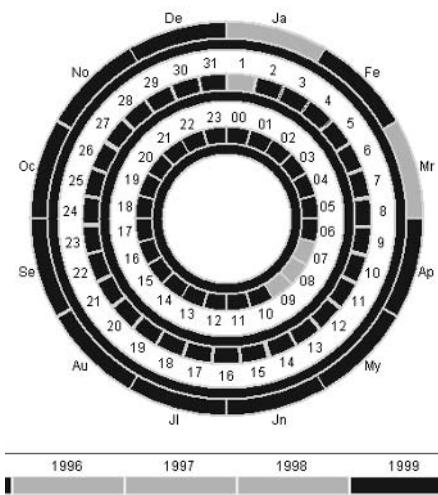
As we have mentioned earlier, time can be treated in two different ways: as a linearly ordered set, i.e. a sequence of time moments, or as organised into cycles. When time is treated as linear, the specification of time moments and intervals can be done in the same way as the specification of individual values and value intervals of any attribute or referrer with a linearly ordered value set, for example a numeric attribute. In particular, interactive devices such as those in Dynamic Query (see Fig. 4.88) can be used for this purpose. For example, the whole length of a slider line may represent the time interval from the year 1960 to the year 2000 in the dataset on crime in the USA. Using the delimiters, one can specify any subinterval of this interval: from 1960 to 1969, from 1991 to 1992, from 2000 to 2000, etc. In this example, the resolution of the time scale is one year, since the data in the dataset refer to years. Figure 4.91 demonstrates another interactive device that may be used for the specification of time moments and intervals. The basic idea is the same as in Dynamic Query, except for the additional possibility to specify explicitly the required length of the time interval. Thus, in Fig. 4.91, the user has selected a 1000-day interval starting from 1 January 1990. The black rectangle represents the selected interval, and the white bar corresponds to the period from 1 January 1976 to 30 December 1999 (this is the period that our Turkish earthquake data refer to). In this interface, the user may specify an interval not only by entering the start date and the end date or the length of the interval in the corresponding text fields but also by manipulating the black rectangle by use of the mouse. When the mouse cursor is positioned on the left side of the rectangle, dragging the mouse shifts the beginning of the query interval. Dragging with the mouse cursor on the right side changes the end of the interval, and dragging in the middle of the black rectangle shifts the whole query interval without changing its length.



**Fig. 4.91.** A time interval of 1000-day length starting from 1 January 1990 is specified here using a slider device for time

In Fig. 4.91, the resolution of the time scale is one day. It is convenient when a query tool allows the user to choose the most appropriate resolution. Thus, in analysing the earthquake data, one might prefer to specify intervals with an accuracy of one month (e.g. from January 1990 to September 1992) or even one year (e.g. from 1990 to 1992). Or, oppositely, it may be necessary to increase the precision and consider hours, minutes, or even seconds. It should be noted that purely visual query devices do not support high precision in the specification of times. Thus, if the query tool shown in Fig. 4.91 did not contain the text fields but only the slider bar, it would be difficult to specify an interval of exactly 1000 or 100 days length starting exactly on 1 January of some year. The reason is that the slider bar which has a length of just a few centimetres, represents the whole period from 1 January 1976 to 30 December 1999, consisting of 8765 different dates. Hence, a millimetre may correspond to several months. If the resolution was increased from days to hours, the same slider length would correspond to  $24 \times 8765 = 210\,360$  different values, and hence exact positioning by mouse dragging would become impossible.

Another limitation of a slider bar/line as a tool for the specification of times is that it does not allow one to express such queries as, for example, "find the earthquakes that happened in the hours from 6 a.m. to 8 a.m." (on weekends, in January, etc.). The formulation of such queries requires a user interface that incorporates a cyclic treatment of time, such as the "Time Wheel" query device described by Edsall and Peuquet (1997). This tool contains three concentric circles divided into segments. The innermost circle is divided into 24 segments representing hours of a day, from 00 to 23. The intermediate circle represents days of a month, from 1 to 31. The outer circle, which is divided into 12 segments, corresponds to months of a year, from January to December. Accordingly, the Time Wheel allows the user to select arbitrary combinations of months within a year, days within months, and times of day. The selection is done by clicking on the corresponding segments of the circles. Thus, the



**Fig. 4.92.** "Time Wheel" query tool in the software system TEMPEST (Source: <http://www.geovista.psu.edu/products/demos/edsall/Tclets072799/cyclicaltime.htm>)

selection shown in Fig. 4.92 (the selected segments are marked by lighter shading) would allow one to investigate what happened on the first days of January and March during the hours from 7 a.m. to 10 a.m. in the years 1996, 1997, and 1998. The years are specified using the linear device below the circles.

We would like to note that the organisational principle of the “Time Wheel” is very felicitous. First, the tool can easily be extended to other temporal cycles. For this purpose, one needs only to add corresponding circles, for example a circle for days of a week, a circle for minutes of an hour, etc. Second, the accuracy of time specification can easily be regulated. Thus, if a precision of one month is sufficient, the user can work only with the outermost circle. When higher precision is required, the user may move to the inner circles.

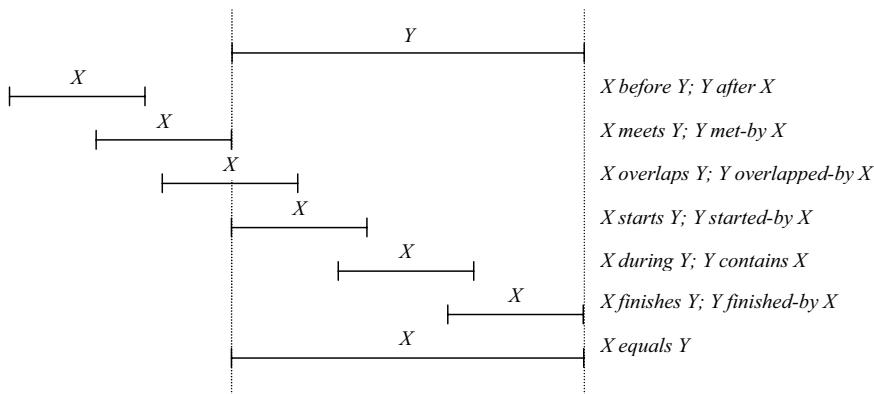
At the same time, the specific implementation described in Edsall and Peuquet (1997) imposes serious limitations on the choice of time intervals. For example, it is impossible to choose the interval from 15 January 1996 to 25 March 1996 or the set of intervals from 15 January to 25 March in the years 1996, 1997, and 1998. The reason is the discreteness of the time scale: the circle representing a year is divided into 12 segments, and it is possible to choose only an entire segment, not a part of it. Therefore, one can choose the whole of January but not the period starting from 15 January. This problem might be tackled by allowing continuous selections on the elements of the “Time Wheel”.

It should not be concluded that a query tool based on a cyclical treatment of time is always superior to a tool incorporating a linear time model. There are “purely linear” queries, which simply cannot be expressed using the “Time Wheel” or similar devices. For example, the “Time Wheel” does not allow one to specify the interval from January 1996 to March 1998. Therefore, an appropriate tool for temporal queries must combine linear and cyclic models of time.

Temporal relations may be divided into topological relations (before, after, overlap, during, simultaneously, between, etc.) and metric relations, i.e. those involving a distance in time, or duration. Unlike the case of space, there are no directional relations (except for “before” and “after”, which can also be regarded as topological): time is a linearly ordered set in which only two directions exist, from the past to the future, and back. Even when a cyclic time model is used, this actually means that time is treated as a linear sequence of cycles, and each individual cycle is a linearly ordered sequence of time moments.

Metric temporal relations are specified through numbers, and hence any query tools devised for specifying numbers may be suited to distances or durations in time, for example elements of Dynamic Query. Topological

relations may be specified using predicates or visually. Thus, Fig. 4.93 provides a visual illustration of the possible relations between time intervals considered by Allen in his theory of temporal reasoning (Allen 1983). Such drawings could be used for the specification of temporal relations in temporal queries.



**Fig. 4.93.** Allen's interval relations, after Allen (1983)

#### 4.6.1.3 Asking Questions: Summary

Querying implies imposing some constraints on values of data components (attributes or referrers) and/or relations between values or value subsets. In fact, various relations are often involved in constraining values: equal or not equal, greater than or less than, in or not in (referring to elements in a set), etc. The difference is that values are constrained by specifying a relation to some constant(s), while relations are constrained using two or more variables. A variable in a query is a reference to a data component, i.e. an attribute or a referrer. Table 4.11 gives a few examples to illustrate the difference between queries that constrain values and queries that constrain relations.

Query constraints can be formulated using a special query language or a set of interactive-manipulation controls. In a traditional query language, constraints are expressions built from variables and constants using signs for arithmetic relations ( $=, \neq, >, \geq, <, \leq$ ) and various predicates indicating other types of relations, for example, *In* for set membership or *Touch* for spatial neighbourhood. In such expressions, it is often possible to use arithmetic operations (add, subtract, divide, multiply, etc.), functions (e.g. logarithm or sine), and set operations such as union, intersection, or difference. When several constraints are specified simultaneously, they may be linked by the logical operations AND, OR, and XOR (exclusive OR).

**Table 4.11.** Queries that constrain values versus queries that constrain relations: Some examples

Constraining values	Constraining relations
In which districts of Portugal is the percentage of children greater than 25?	In which districts of Portugal is the percentage of children greater than the percentage of elderly people?
Find all countries where one of the official languages is Italian.	Find all countries that have at least one official language in common with some other country.
Find natural parks situated south of Alps.	Find natural parks situated south of large mountain ranges.

A visual query language represents relations and operations by graphical symbols rather than keywords. Query constraints take the form of a diagram constructed from such graphical symbols; this diagram is an equivalent of an expression in a traditional query language.

Interactive-manipulation controls are much simpler to use than query languages but permit only restricted subsets of possible queries. Typically, they are suitable for constraining values but not for constraining relations. An interactive-manipulation query tool suggests some visual representation of the value set(s) of one or more attributes, such as, the slider bars/lines in Dynamic Query representing value ranges of numeric or ordinal attributes, or the concentric circles in the “Time Wheel”. Sometimes, a query tool does not suggest its own representation of a value set but works on top of a visual data display, such as a map or a graph (see Fig. 4.89). The representation of a value set may be continuous (using a slider bar or a graph area) or discretised (using a “Time Wheel” divided into segments or the territory of Portugal in a map, divided into districts).

Constraining values is done by selecting specific points, segments, or regions in a visual representation of a value set. For this purpose, the user usually applies mouse operations: pointing, clicking, and dragging. A region may be selected by moving delimiters, as in Dynamic Query, or by drawing a geometrical figure, for example a frame or a circle on a map.

The main virtues of query devices based on interactive manipulation are their simplicity and efficiency. The formulation of each query takes so little time that the user can specify many queries during a session of data analysis. If the query tool is capable of providing an immediate answer to any such question, it becomes an extremely powerful instrument for exploratory data analysis. The user obtains quick and easy access to the characteristics of any references and to references with specific characteristics.

He/she may observe the impact of a slight modification of a query and see how constraining the values of one component excludes certain values of other components. These opportunities are much more important for exploratory analysis than are the sophistication and flexibility of query languages, which allow a highly refined specification of almost any information need but are too heavy to encourage “playing” with query constraints and observing the impacts.

We would like to use the term “dynamic query tools” to refer to query tools that

- allow the user to specify and modify queries very easily and quickly, and
- immediately react to any modification of a query by fast provision of the appropriate answer.

Hence, we shall use the term “dynamic query” in a general sense, bearing in mind certain properties of the tool, specifically, ease of use and fast response. The particular query tool widely known as Dynamic Query is one of the tools that possess these properties but not the only one. In our further discussion, we shall refer to this particular tool by its proper name Dynamic Query (where both words start with capitals). When the same words start with lower-case letters, this combination of words will mean the general principle of building various query tools.

Let us now consider what answers to a user’s questions may be expected from a query tool and in what form they may be presented to the user.

## 4.6.2 Answering Questions

As we have pointed out earlier, queries are mostly elementary questions about data and can be classified into lookup, comparison, and relation-seeking.<sup>20</sup> The expected answers to lookup and relation-seeking questions are references and/or characteristics, while the answers to comparison tasks are relations between references or between characteristics.

Let us consider first the case where a query result is a subset of references and/or characteristics. How can we systematically describe the possible ways of presenting such a result to a user?

One of the distinctions that we deem important is how the presentation of query results is related to the information that was shown on the screen before the query was specified. The display of the results may be inde-

---

<sup>20</sup> A little later, we shall describe some query tools suitable for asking a particular sort question related to searching for a behaviour.

pendent of the previous content of the screen, for example it may appear in a new window or simply replace what was previously shown, or it may somehow modify this content. Although complete replacement of the information on the screen by the presentation of query results is theoretically possible, this is rarely done in practice, since radical changes of screen contents may lead to confusion of the user.

For an independent display of query results, any form of data representation may be used, such as text, a table, or graphics. An example of such a display can be seen in Fig. 4.89, where answers to “What’s this?” questions are shown in pop-up windows while the original data representation remains mostly unchanged, except that the selected element of the display is indicated by highlighting.

Modification of the information content of the screen by the results of a query is possible in three basic ways:

- *Filtering*: Screen elements representing data items that do not satisfy the query are removed from the screen. Only the data items satisfying the query are displayed. A variation of this technique is to “mute” the representation of data items that do not satisfy the query. The corresponding graphical elements may appear “bleached” and/or reduced in size. User interaction with such display elements is typically reduced, for example they do not react to pointing or clicking with the mouse.
- *Marking*: Screen elements representing data items that satisfy the query are specially marked, for example by dedicated colouring, so as to be clearly distinguishable from the rest. A variation of the marking technique is sometimes used to represent the results of a query with several constraints or the results of a sequence of queries: visual elements representing data items are shown in different colours or shades depending on how many query constraints they satisfy.
- *Addition*: New graphical elements are added to an existing display to represent data items that satisfy the query and have not been displayed before. If some of the query results have already been shown, the corresponding display elements may either remain unchanged or be marked somehow to attract the user’s attention. Usually, answering a query by adding new display elements requires the user to be able to “clean” the display of the results of all previous queries.

Of these three variants of answering queries, the first two are mostly used in dynamic query tools, which, as we explained before, are extremely valuable in exploratory data analysis. Let us give some examples of these two variants.

#### 4.6.2.1 Filtering

The well-known Dynamic Query (Ahlberg et al. 1992) operates according to the filtering paradigm. As we have described earlier, Dynamic Query allows one to specify questions of the kind “What references correspond to the specified attribute values?” Dynamic Query is combined with a graphical data display (sometimes several displays), which originally represents the whole reference set. Some attributes corresponding to the references may be visualised in such a display, but these attributes and the methods used to represent their values are irrelevant to the functioning of the tool.

When the user of the tool moves any of the sliders of Dynamic Query, the tool interprets this as a specification of a query constraint and immediately responds to this by filtering out the references that do not satisfy this constraint. The corresponding graphical elements are removed from the display or shown in a “muted” manner. All of the user’s subsequent operations on the sliders are interpreted as a modification of a previously specified query constraint (when a slider that has already been moved is moved again) or as adding new constraints (when the user moves a slider that has not previously been moved). After any operation, the display is immediately updated. In some realisations, the display may be dynamically updated even during the process of slider movement.

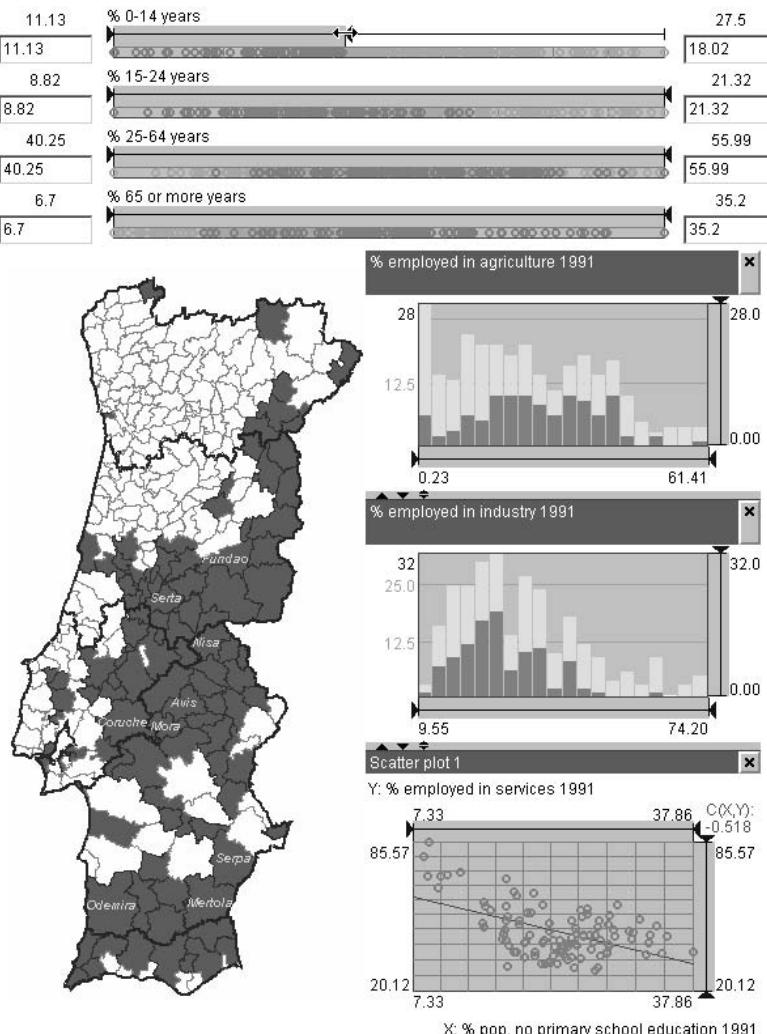
Let us consider a few illustrations. In Fig. 4.94, we see a screenshot of the user interface of one of the existing realisations of Dynamic Query (top) and several data displays:

- a map of Portugal, with the territory divided into administrative districts;
- two frequency histograms, representing the distribution of the values of the attributes “% employed in agriculture 1991” and “% employed in industry 1991” over the set of these administrative districts;
- a scatterplot, representing the combinations of values of the attributes “% employed in services 1991” and “% pop. no primary school education” in the districts of Portugal.

Dynamic Query contains slider bars for four attributes characterising the age structure of the population in the districts of Portugal: “% 0–14 years”, “% 15–24 years”, “% 25–64 years”, and “% 65 or more years”.

Hence, all of the displays, including that of Dynamic Query, represent different attributes of one and the same dataset and have a common reference set, specifically, the set of administrative districts of Portugal. On the map, each district is represented by a figure with a shape and location corresponding to the geographical characteristics of this district. In the scatterplot and the dot plots included in the user interface of Dynamic Query,

the districts are represented by small circles, or dots. The histograms are aggregated data displays and represent the sizes of groups of districts rather than individual districts. The districts have been grouped according to the values of the attributes “% employed in agriculture 1991” and “% employed in industry 1991” by dividing the value ranges of these attributes into 20 intervals of equal length.

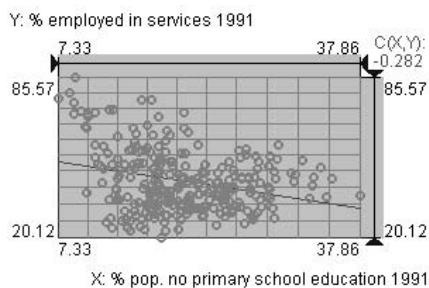


**Fig. 4.94.** The Dynamic Query tool has been applied here to several graphical displays with a common reference set, specifically, the set of districts of Portugal. Specification of query constraints removes or “mutes” the representation of the references that do not satisfy these constraints

Figure 4.94 demonstrates the effect of moving one of the delimiters in Dynamic Query, specifically, the right delimiter on the slider line corresponding to the attribute “% 0–14 years”. Moving this delimiter is interpreted as setting an upper limit upon the values of the corresponding attribute. In the figure, the position of the delimiter corresponds to the attribute value 18.02. For the Dynamic Query tool, this means that the user would like to see which districts of Portugal have a percentage of children (i.e. people aged from 0 to 14 years) less than or equal to 18.02.

In response to the user’s operation, Dynamic Query divides all the districts into two groups: the districts satisfying the constraint (i.e. having not more than 18.02% of children in their population) and the districts not satisfying the constraint (i.e. having more than 18.02% of children). Every display present on the screen reacts to this. On the map, the districts not satisfying the constraint are “muted”: only their contours are drawn, using an inconspicuous light grey colour; the interior is not coloured; and the names of those districts are not shown. Moreover, these contours become insensitive to mouse operations. They are merely placeholders, rather than real, active graphical objects. The real objects are not there any more; they have been filtered out.

A similar metamorphosis has happened in the dot plots of Dynamic Query to the dots symbolising the districts that do not satisfy the current query. These dots have become pale and inactive. At the same time, the scatterplot has reacted in a different manner: only the dots representing the districts that satisfy the query continue to be visible, and the remaining dots have been removed from the display. For comparison, the original appearance of the scatterplot display (i.e. before the constraint was set) is presented in Fig. 4.95.



**Fig. 4.95.** The scatterplot from Fig. 4.94 as it looked before any query constraints were specified

The histograms, which do not represent individual districts but only district counts, have been transformed in their own way. For each bar, the

proportion of the districts that belong to this bar and satisfy the query has been computed. The bar has been divided into two differently coloured segments in accordance with this proportion. The darker segment shows the number of districts that satisfy the query, and the lighter segment the number of districts that do not satisfy the query.

The reaction of all of these displays to the outcome of the Dynamic Query tool allows an analyst to obtain quite an amount of valuable information about the districts of Portugal with a small proportion of children and about those with a high proportion. Thus, the analyst can immediately see where the districts in each group are geographically located. The map exhibits two major clusters of districts with a small proportion of children: in the central inland part of the country and in the south. It can also be concluded that the proportions of children are mostly higher than 18.02% in the north and in the central districts close to the western coast.

From the frequency histogram of the attribute “% employed in agriculture 1991”, the analyst may see that proportions of children over 18.02% prevail in the districts with a small percentage of people employed in agriculture. These districts are represented by the bars on the left of the histogram, and the dark parts of these bars, which show the numbers of districts satisfying the query constraints, are much smaller than the light segments, corresponding to the districts that do not satisfy the query, i.e. have more than 18.02% of children in their population. In the centre of the histogram, where the proportions of people employed in agriculture have medium values, the division of the bars according to the percentage of children demonstrates a slight prevalence of districts with a low percentage of children. On the right, where the proportions of people employed in agriculture are high, there are again more districts with a higher percentage of children. However, the bars in this part of the histogram are quite short, i.e. there are not many districts with a so high employment in agriculture.

From the histogram of the attribute “% employed in industry 1991”, one can see that the bars corresponding to low employment of the population in industry (these bars are on the left) are divided into nearly equal parts with respect to the satisfaction of the query constraints. In the centre, the darker parts become somewhat smaller than the light parts, and on the right, there is only a very small dark segment in one of the bars. This means that higher percentages of children prevail in districts with a high proportion of people employed in industry.

After the query constraint has been set, the scatterplot demonstrates a certain correlation between the attributes “% employed in services 1991” and “% pop. no primary school education”, whereas no clear correlation was visible in the original view (see Fig. 4.95). This means that in the districts with a low percentage of children, the proportion of people without

primary education tends to decrease as the proportion of people employed in services increases.

Some information can also be obtained from the dot plots included in the Dynamic Query device. One can see that excluding high values of the attribute “% 0–14 years” excludes also high values of the attribute “% 15–24 years”: the dots on the right of the corresponding dot plot are “muted”. This means that the districts with a high percentage of children usually have a high percentages of young people (i.e. aged from 15 to 24 years) as well. From the remaining two dot plots, it may be concluded that many of the districts with a high percentage of children (which have been filtered out) have a low percentage of people of working age (i.e. from 24 to 65 years) and of people of retirement age (i.e. 65 or more years). At the same time, in the districts with less than 18.02% of children, high percentages of people aged from 25 to 64 years occur quite rarely as well.

The purpose of this detailed description of what can be seen in various displays after setting a constraint in Dynamic Query was to demonstrate the role of querying in exploratory data analysis. This role does not imply that the user searches for anything specific, as in querying a hotel reservation system. In EDA, the task of finding objects with particular characteristics (for example a hotel in a particular city situated close to the railway station and with a price within a certain price range) is not typical. It is more typical to use queries for answering questions such as:

- How many objects possess/do not possess the given characteristics?
- Where are these objects located in space?
- How are these characteristics related to other characteristics?

Moreover, the analyst is usually not very interested in answering these questions in regard to just one particular subset of characteristics. He/she would rather take various subsets of characteristics and ask these questions again and again. Thus, in our example, it is not actually the goal of the analyst to explore the characteristics of districts that have the percentages of children up to 18.02. Instead, the goal is to investigate how demographic characteristics are distributed over the territory of Portugal and what links exist between different demographic attributes. Repeated selection of various subsets of districts by means of querying is an instrument used for this investigation. This purpose and this method of using a query tool explain why the tool needs to be dynamic: it is important that the analyst is able to change the subset selection easily, and that the properties of the new subset become promptly available for observation.

Let us return to the dot plots included in the Dynamic Query interface. In the following discussion, for the sake of brevity, the references with

corresponding characteristics that satisfy the current query constraints will be referred to as *active references*. References with characteristics that do not satisfy the current constraints will be called *inactive*.

So, the role of the dot plots in Dynamic Query is to provide three categories of information:

1. The distributions of the values of different attributes over the *whole set of references* (in our example, the districts of Portugal).
2. The distributions of attribute values over the *subset of active references*.
3. The distributions of attribute values over the *subset of inactive references*.

The first category of information is shown by the distribution of all dots along a slider line, irrespective of their colouring. The second category of information is shown by the distribution of the “active”, dark grey dots. The lighter, “inactive” dots indicate what characteristics pertain to the inactive references.

Besides showing the distributions and characterising the subsets of active and inactive references, the different colouring of the dots in the dot plots can often help to reveal correlations between characteristics. Thus, in the example in Fig. 4.94, the colouring of the dots in the dot plots indicates that low percentages of children typically co-occur with low to medium percentages of young people and with medium to high percentages of elderly people.

At the same time, the information shown by the dot plots allows the analyst to anticipate what modifications of the current query will cause significant changes to the query results and what modifications will have only a slight effect or no effect at all. Thus, moving the right delimiter on the second from top slider line to the left will not change anything until the rightmost dark grey dot is reached. Analogously, the current set of active districts will be almost insensitive to moving the left delimiter on the third or fourth slider line to the right by about one-fifth or even one-fourth of the total length of the slider line. In contrast, moving the right delimiter on the topmost slider line further to the left will significantly reduce the number of active districts, since this part of the slider line corresponds to a concentration of dark dots. For the same reason, moving this delimiter to the right will notably extend the subset of active districts.

Providing such query-related information directly within the query device is very convenient, since this information can be taken into account in query building. Of course, not only dot plots may be used for this pur-

pose.<sup>21</sup> Moreover, using dot plots is not an ideal solution, because they very often suffer from overplotting, i.e. some symbols covering others. Hence, one dot may actually stand for several or even quite many objects, and it becomes impossible to estimate the sensitivity of the query to a particular slider movement.

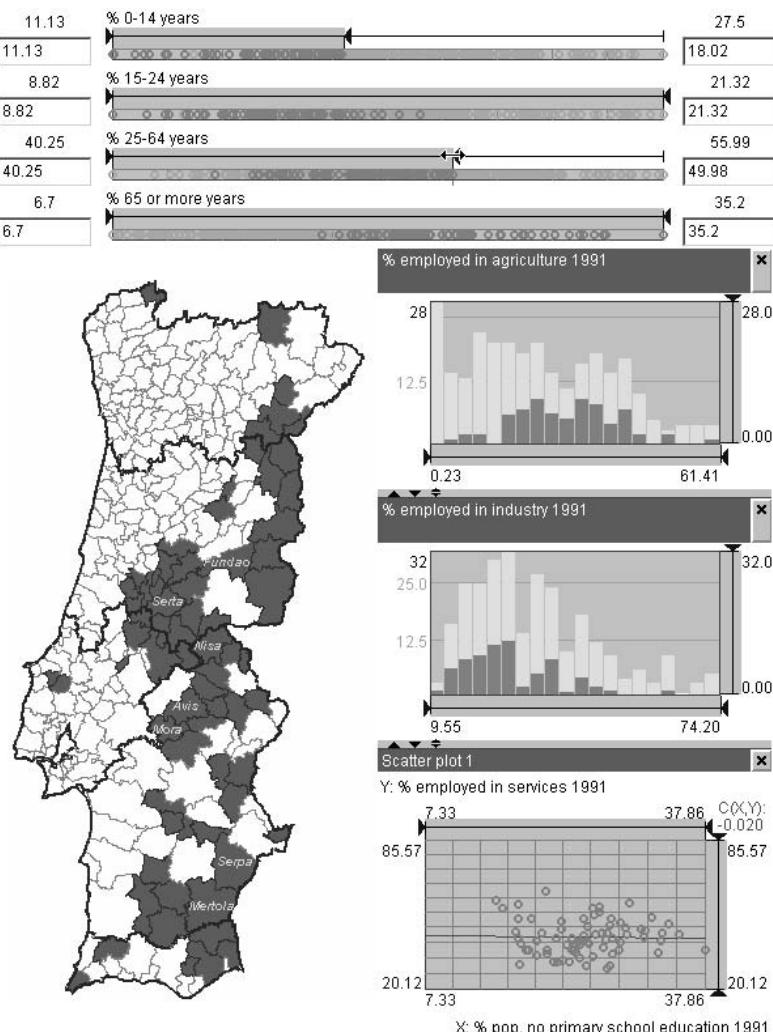
Using frequency histograms instead of dot plots appears to be a suitable alternative although histograms are not problem-free either. Thus, when the distribution of the values of a attribute is very skewed, the corresponding histogram may contain one or a few extremely high bars while the remaining bars are very low. It should be taken into account that the maximum height available for a histogram in a dynamic query device is usually quite limited; hence, some bars in a histogram may be indiscernible.

A typical approach to dealing with skewed distributions in statistics is to transform the values, for example by taking their logarithms. However, such a transformation may cause difficulties in using a dynamic query device: the correspondence between positions on a slider line and the values of the respective attribute becomes non-intuitive. The same distance has different meanings in different parts of a slider line; therefore, the procedure of moving the delimiters involves a significant cognitive effort.

We are aware of a realisation of Dynamic Query where information about value distributions is shown by means of built-in histograms, i.e. each slider bar has a histogram inside (this is described, for example, in Li and North (2003)). The user has an opportunity to switch between linear and logarithmic functions for encoding attribute values by positions. However, the histograms do not change in response to modifications of query constraints. Hence, this query interface provides only the first category of information about the distribution in the list given above. Therefore, it does not facilitate the revealing of correlations between attributes and, in fact, does not properly indicate the sensitivity of the query result to possible modifications of the query. There is another dynamic query tool, called Attribute Explorer (Spence and Tweedy 1998, Spence 2001), where histograms are also used, and here they change dynamically as the query constraints are changed. This tool utilises the idea of marking rather than filtering; and we shall discuss this in more detail in the next subsection.

<sup>21</sup> In the original implementation of the Dynamic Query tool (Ahlberg et al. 1992), dot plots were not used. However, some information to estimate the (in)sensitivity of the query was provided by different colouring of segments of the slider bars: grey segments represented value subintervals for which there were no references satisfying the current query constraints, while yellow colouring of a segment indicated the presence of active references with values in the corresponding subinterval.

We have discussed so far how a single constraint is set in the Dynamic Query tool and what impact this may produce on various data displays. Although it may easily be guessed what outcomes can be expected from further manipulations of the controls in the Dynamic Query device, we shall still say a few words on this topic and give an illustration.

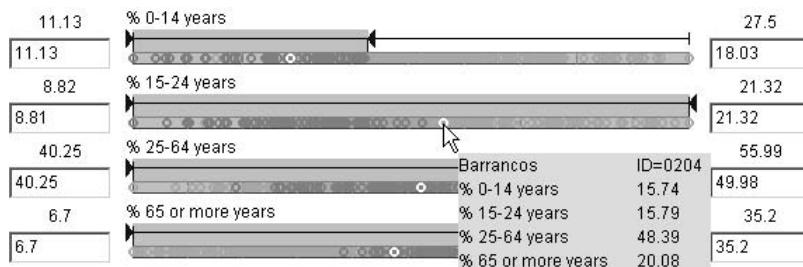


**Fig. 4.96.** Changes in the data displays from Fig. 4.94 after adding one more query constraint in Dynamic Query

Any change of delimiter positions in Dynamic Query can modify the division of the reference set into the references satisfying the query and the

references not satisfying it, or active and inactive references. Adding a new constraint or tightening an existing constraint (by moving a left delimiter to the right or a right delimiter to the left) turns some of the currently active references into inactive ones. Loosening or removing a constraint moves some of the currently inactive references into the class of active references.

Figure 4.96 demonstrates the effect of moving the right delimiter of the slider line corresponding to the attribute “% 25–64 years” to the left, i.e. setting an upper limit on the values of this attribute. Now, the set of active references consists of districts with relatively low proportions of children (up to 18.02%) and low to medium proportions of the population aged from 25 to 64 years (specifically, up to 49.98%). As could quite naturally be expected, these districts are characterised by rather high proportions of elderly people, and this is clearly seen in the dot plot at the bottom of the Dynamic Query window. The range of the values of the attribute “% 15–24 years” corresponding to the active districts has shrunk. Now, the active dot representing the maximum value is located in about the middle of the slider line for this attribute. The “What’s this?” query tool allows us to see this maximum value and the name of the district that this value refers to, as is shown in Fig. 4.97.



**Fig. 4.97.** The maximum value of the attribute “% 15–24 years” available in the subset of the districts of Portugal satisfying the current query constraints is the value 15.79, in the district of Barrancos

The displays that we saw earlier in Fig. 4.94 have changed in response to adding the new query condition. The map in Fig. 4.96 shows us the spatial distribution of the districts satisfying the query: they are mostly located inland, in the east of the country. The upper histogram characterises the active districts as having mostly medium proportions of people employed in agriculture while the lower histogram indicates, as before, mostly low proportions of people employed in industry. In the scatterplot, the group of dots with low values of the attribute “% pop. no primary school education

1991” and high values of the attribute “% employed in services 1991” (in the upper left corner of the scatterplot in Fig. 4.94) has disappeared: it can no longer be seen in Fig. 4.96. This means, first, that such combinations of values of these two attributes were pertinent to districts with relatively high proportions of people aged from 25 to 64 years; and second, that the districts satisfying the query have medium to high percentages of uneducated people and mostly low proportions of people employed in services. The negative correlation, which could be seen in the scatterplot in Fig. 4.94, can no longer be seen in Fig. 4.96.

A shortcoming of Dynamic Query, as well as of any filtering-oriented dynamic query tool, is the difficulty of comparing results of different queries. In particular, it may be not easy to note the changes that occur after a modification of the current query constraints, and after a few such modifications the user can completely forget his/her earlier observations. Of course, it is possible to take screenshots and compare them, as we did with Figs 4.94 and 4.96, but this substantially reduces the dynamism. Another option is to return to the previous states by appropriate modification of the query constraints. This can easily be done, but there is a danger of forgetting the later states.

Daniel Carr and his colleagues (Carr et al. 2000, 2002) have suggested a witty, although partial, solution known as “conditioned maps”. The main idea is to use a matrix composed of  $3 \times 3$  maps (or, in principle, any other type of display) in order to represent simultaneously answers to nine queries. The queries are not arbitrary but are specified through division of the value ranges of two numeric attributes. Each value range is divided into three subintervals. The selection of one subinterval for each attribute defines a two-condition query in terms of these attributes. In accordance to the possible number of different ways of choosing the intervals, the division specifies  $3 \times 3 = 9$  different queries, the answers to which are shown in the  $3 \times 3 = 9$  displays. In each display, only the data items satisfying the respective query are visible. Hence, the filtering technique is applied in the individual displays, whereas the entire collection of displays contains the complete information. The answers to the different queries can easily be compared. The user may dynamically change the intervals that the value ranges of the attributes are divided into. This results in the displays being updated to represent the answers to the modified queries.

In principle, the same approach can be used for querying in terms of a single attribute. In this case, a one-dimensional horizontal or vertical display arrangement is appropriate, instead of a display matrix. Extension of the technique to more than two attributes is hardly possible. The division of the attribute value ranges into exactly three subintervals is not a limita-

tion in principle; there could equally well be two or four subintervals. However, increasing the number of subintervals multiplies the queries and, consequently, the displays needed for the representation of the answers to the queries. This may decrease the legibility of each display and increase the user's cognitive load, since the user will need to view and compare many individual displays.

Further opportunities for the comparison of the results of several queries exist in marking-oriented dynamic query tools.

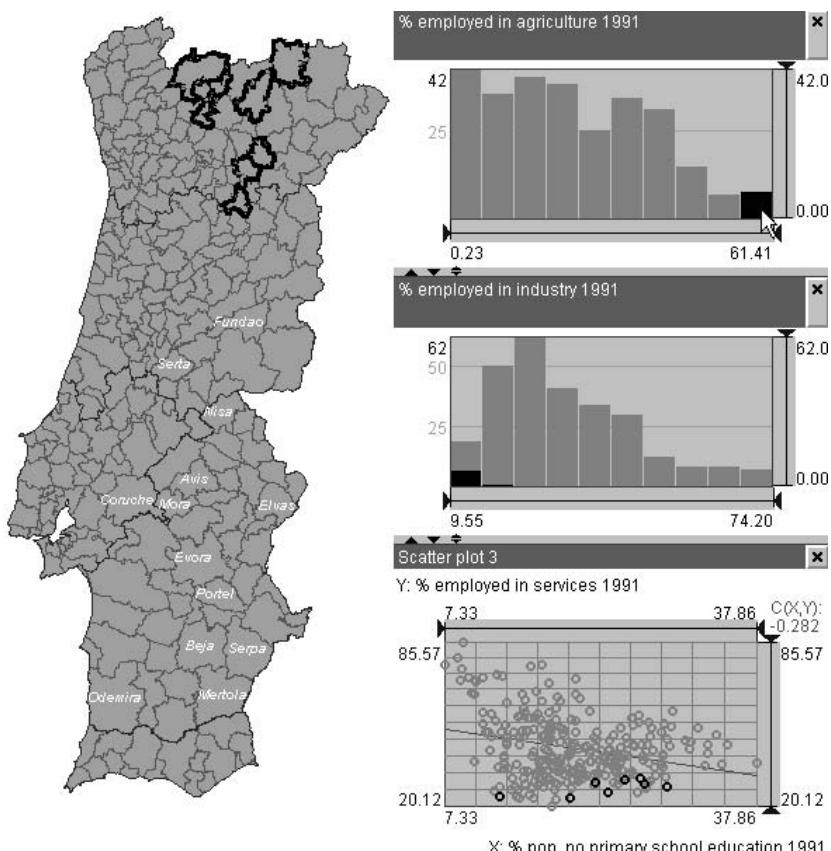
#### 4.6.2.2 *Marking*

Let us start by discussing a simple variant of marking, which assumes that a display element may be in one of two possible states, selected (active) or neutral. The selected state is indicated by a particular visual means, for example a certain colour. In our examples, this will be a black colour. For illustration, we shall use the same map, histogram and scatterplot displays as in Figs 4.94 and 4.96, except that each histogram will have ten bars instead of 20 (i.e. the value ranges of the respective attributes are divided into ten rather than 20 subintervals). The filter will be cancelled; hence, all the displays will portray the full set of districts of Portugal.

Figure 4.98 demonstrates the result of clicking on the rightmost bar of the histogram representing the value distribution of the attribute “% employed in agriculture 1991”. This bar corresponds to the last one-tenth of the attribute’s value range, i.e. to the interval from 55.29 to 61.41. After the clicking, the bar has become “active”, which is indicated by its being shaded in black. However, what is interesting is not the change in the colour of this bar but the changes that have occurred in the other displays. The clicking on the bar has divided the set of districts into two subsets: the districts with more than 55.29% of people working in agriculture, and the districts where this percentage is less than or equal to 55.29. The former subset of districts is treated as selected, or active, and the latter as neutral. In response to this division of the districts into selected and neutral, all the displays have changed so that the selected districts are specially marked using black (for consistency, selections must be shown in a similar way in different displays).

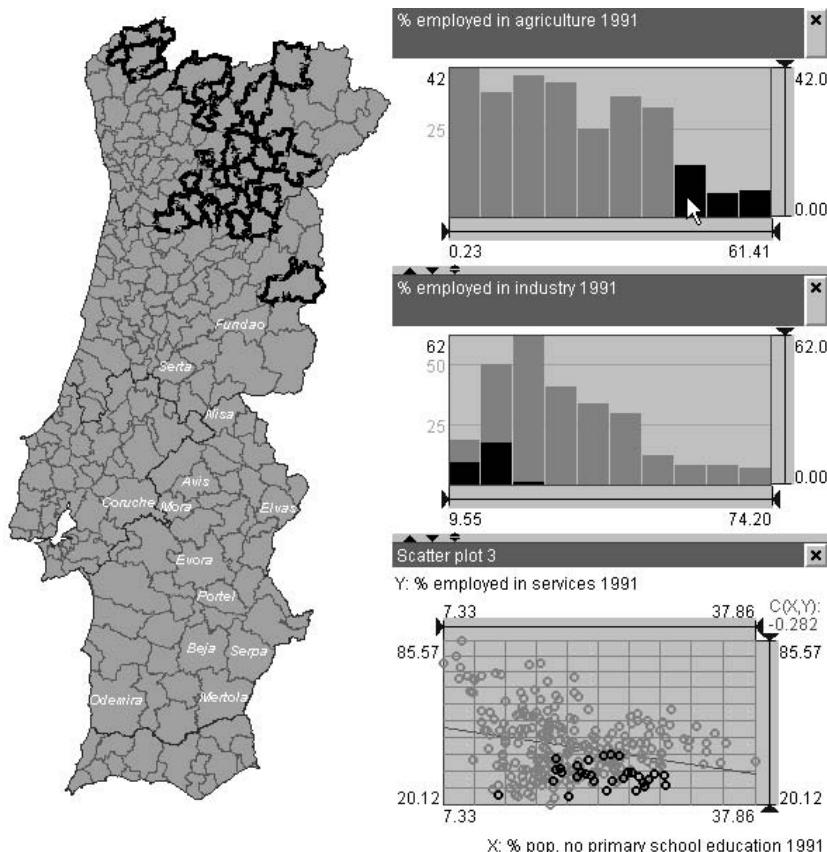
On the map, the selected districts are marked by thick black boundaries. All these districts are in the northern part of the country. The second histogram indicates the position of the selected subset of districts with respect to the value range of the attribute “% employed in industry 1991”. Almost all of these districts have values of this attribute in the first of the ten subintervals represented in the histogram. The number of such districts is indicated in the histogram by shading a segment of the respective bar in

black. The height of this segment is proportional to the number of selected districts with attribute values in this interval. A small black segment can also be noticed in the second bar from the left. This shows that there are some districts (or, most likely, just one district) with values of “% employed in industry 1991” in the second subinterval. Anyway, it is clear that very high proportions of employment in agriculture co-occur with very low percentages of employment in industry. In the scatterplot, the selected districts are shown as black dots. It can be seen that these districts are characterised by low employment in services, while the percentage of uneducated people varies significantly.



**Fig. 4.98.** Clicking on a bar in a histogram selects the districts with attribute values that fit into the corresponding interval. All the displays respond to the selection by special marking of the selected districts. For consistency between different displays, the selection is indicated using the same colour (black) in all cases

Let us now click on the two neighbouring bars of the bar that we have just selected (see Fig. 4.99). This extends the set of selected districts by adding the districts with attribute values that fit into the intervals corresponding to these bars. Hence, the resulting selection consists of the districts with more than 43.06% of people employed in agriculture.



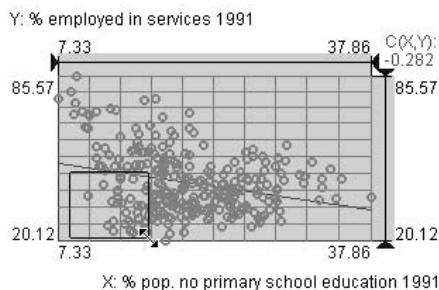
**Fig. 4.99.** Clicking on two other bars in the same histogram extends the set of selected districts by adding the districts with attribute values that fit into the intervals corresponding to these bars. The resulting selection consists of the districts with more than 43.06% of people employed in agriculture

All the displays present on the screen have been updated to represent the result of the new query. In the map, we can see that all but one of the active districts are situated in the northern part of Portugal. The frequency histogram of the attribute “% employed in industry 1991” shows us that the active districts have low values of this attribute. In the scatterplot, it can be observed that the selected districts have low values of the attribute

“% employed in services 1991”. Only one of the selected districts has a low percentage of uneducated people, while all of the others are characterised by medium values of this attribute.

For further investigation, the user may click arbitrarily on graphical elements in any of the displays: bars in any of the histograms, district shapes in the map, or dots in the dot plot. This operation puts the corresponding districts into the active state if the graphical element has not been previously selected, and deselects the corresponding districts if the graphical element has been previously selected (a bar of a histogram is considered as selected if it is completely rather than partly black). Hence, each click sets or cancels a query constraint. Several query constraints are treated in this query tool as being connected by the logical operation “OR” (non-exclusive).

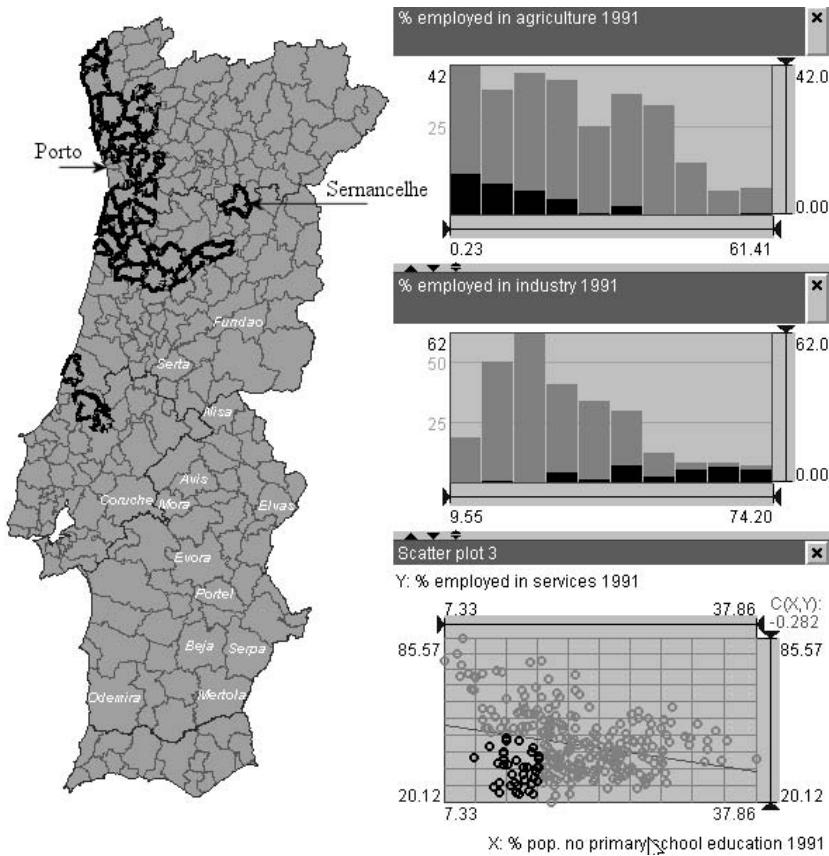
Query constraints may be set not only by clicking but also by outlining graphical elements in a display, for example by drawing a frame around the elements that need to be selected. Figure 4.100 demonstrates how a group of dots in a scatterplot may be selected in this way, and Fig. 4.101 shows the impact of this selection (the previous selection, shown in Fig. 4.99, was cancelled before the new one was made).



**Fig. 4.100.** Selection by drawing a frame around dots on a scatterplot

The frame has been drawn on the scatterplot so as to select districts with relatively low values of both “% employed in services 1991” and “% pop. no primary school education 1991”. The corresponding dots are located in the lower left corner of the display. After the query has been formulated in this way, we can see the spatial positions of the active districts and their characteristics in terms of the attributes “% employed in agriculture 1991” and “% employed in industry 1991”. Thus, we can observe that the active districts are located mostly in the north-west of the country, where they form a cluster with a rather interesting shape: it stretches along the coast in a north-south direction and has a peculiar “tail” at its southern end oriented inland, to the east. However, the city of Porto and a few neighbour-

ing districts do not belong to this cluster. Most probably, these districts have relatively high percentages of people employed in services and hence do not satisfy the query constraints.



**Fig. 4.101.** After the selection of a group of dots in the scatterplot, as shown in Fig. 4.100, the corresponding districts have become active. All the displays represent these districts by marking

The histograms show us that the districts with low employment in services and a low percentage of uneducated people are mostly characterised by quite low employment in agriculture and quite high employment in industry. However, there is an exception, indicated by the narrow black segment in the rightmost bar in the histogram for the attribute “% employed in agriculture 1991”. By applying the “What’s this?” query tool, we can find out that this segment corresponds to the district of Sernancelhe, which has 58.01% of working people employed in agriculture and 18.14%

of people employed in industry. The latter value fits into the second bar from left in the histogram for the attribute “% employed in industry 1991”, where we can also see a small black stripe. We have drawn an arrow on the map to point to the geographical location of this district. It can be seen that this district stands somewhat apart from the main cluster of districts satisfying the query constraints. If we look at the map in Fig. 4.98, where the districts with high employment in agriculture are marked, we can notice that the outline of Sernancelhe is also marked there. Apparently, the active, black dot in the lower left corner of the scatterplot in Fig. 4.98 corresponds to this district.

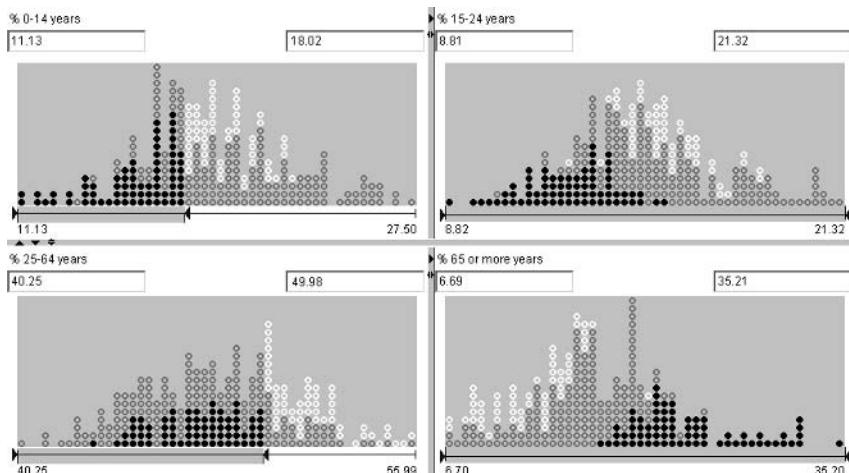
The method of querying demonstrated here, by direct manipulation of graphical data displays with the results being presented by means of marking elements of these displays, is commonly referred to as “brushing” (this term seems to originate from Newton (1978)). Brushing is a very popular technique and can be found almost in all software systems that have been suggested for exploratory data analysis. Some of them implement more sophisticated forms of brushing than what we have just described. One of the enhancements that exists is multicolour brushing: the explorer can use distinct colours for marking different selections. This facilitates the comparison of results of several queries. However, it is unclear with this approach how to mark references that satisfy more than one query; therefore, multicolour brushing is usually applied to non-overlapping selections.

Some researchers apply the term “brushing” to any dynamic query tools where query results are represented by marking rather than filtering, irrespective of whether query constraints are specified through direct manipulation, or by means of sliders or checkboxes or in any other way (a rather comprehensive study on various variants of brushing can be found in Chen (2003)). In this sense, the tool developed by Robert Spence and Lisa Tweedy and widely known as Attribute Explorer (Spence and Tweedy 1998, Spence 2001) is often described as “brushing histograms”. As in Dynamic Query, sliders are used in Attribute Explorer to set query constraints. As we have already mentioned, the slider bars in Attribute Explorer are associated with frequency histograms. Query results are shown in these histograms by means of marking. Different colours are used to denote the fulfilment of all query constraints, one constraint failure, two constraint failures, and so on.

To demonstrate the idea, we have produced some screenshots using for this purpose an analogue of Attribute Explorer, which differs from the original tool in using dispersion graphs rather than histograms to represent attribute value distributions. The dispersion graph is a modification of the dot plot technique, where dots that fit in the same position are put one above the other. In Fig. 4.102, we see the specification of the same query

as in Fig. 4.96 using this dispersion-graph-based query tool. Specifically, the upper limit for the attribute “% 0–14 years” has been set to 18.02, and the upper limit for the attribute “% 25–64 years” has been set to 49.98.

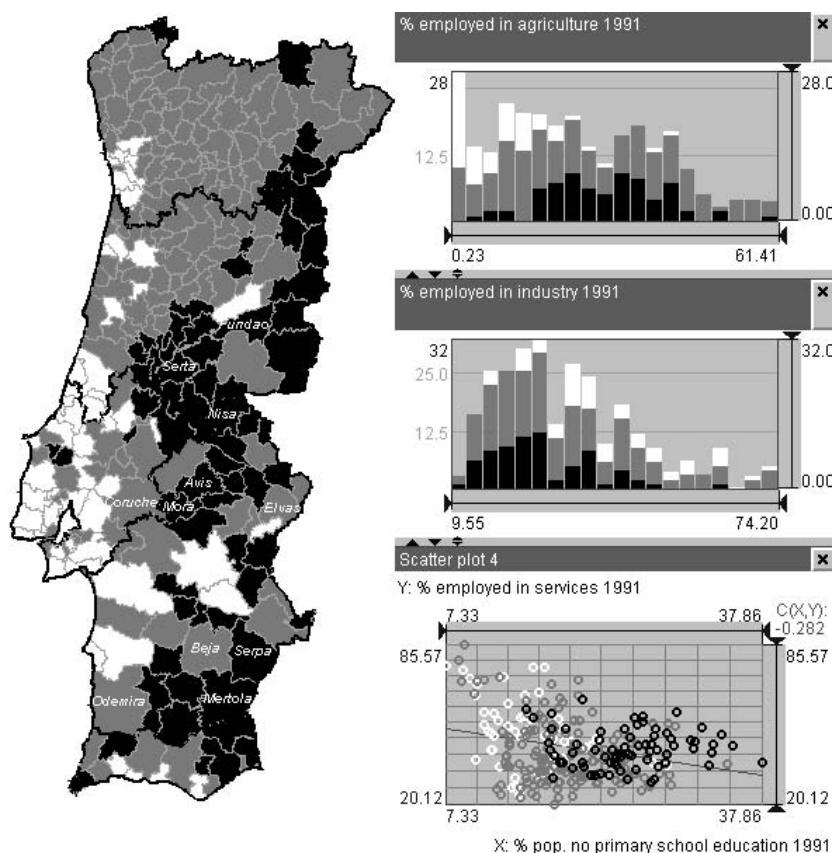
In response to the setting of constraints, the dots on all the dispersion graphs become coloured in various ways depending on how many constraints they satisfy. In the example shown in Fig. 4.102, black is used to denote complete satisfaction of the query, white means that none of the constraints is fulfilled, and grey marks the references that fail to satisfy one of the constraints (no matter which one). When there are more than two constraints, different shades of grey are used, with darker shades corresponding to more constraints being satisfied.



**Fig. 4.102.** A dynamic query tool where marking is used to represent information about constraint satisfaction. Here, two query constraints are specified: the value of “% 0–14 years” must be below 18.02 and the value of “% 25–64 years” must be below 49.98. The black colour marks the references (districts of Portugal) satisfying both constraints, grey indicates that one of the constraints is not fulfilled, and white means that none of the constraints is fulfilled

Not only the graphs included in the user interface of the dynamic query tool, but also other data displays, may apply colour coding to represent constraint satisfaction. Thus, Fig. 4.103 demonstrates how the map, histograms, and scatterplot that were used in our earlier examples represent the result of the query shown in Fig. 4.102.

Multicolour marking can show not only *how many* query constraints are satisfied but also *which* constraints, in the case of multiple constraints, are satisfied. This is illustrated in Fig. 4.104C: for the two query constraints specified in Fig. 4.102, the colours have the following meanings:



**Fig. 4.103.** Various displays react to the query shown in Fig. 4.102 by marking graphical elements according to the degree of satisfaction of the query

- *Green:* Both constraints are fulfilled, i.e. the districts have no more than 18.02% children and no more than 49.98% people aged from 25 to 64 years.
- *Yellow:* Only the first constraint is fulfilled, i.e. the districts have no more than 18.02% children and more than 49.98% people aged from 25 to 64 years.
- *Blue:* Only the second constraint is fulfilled, i.e. the districts have no more than 49.98% people aged from 25 to 64 years and more than 18.02% children.
- *White:* None of the constraints is fulfilled, i.e. the districts have more than 18.02% children and more than 49.98% people aged from 25 to 64 years.

As we can see, such a refined representation of query results requires many different colours to be used. In our example, with just two query constraints, four colours are involved. When more constraints are added, the necessary number of colours increases dramatically: three constraints require eight different colours, four constraints require 16 colours, and so on.<sup>22</sup> The resulting displays become very difficult to understand. In fact, it is already difficult with just two constraints and four different colours, as in Fig. 4.104C: it is necessary to remember the meaning of each colour or repeatedly consult the colour legend.

To reduce the cognitive load, some researchers (see, for example, Spence (2001)) suggest using distinct colours only for particular combinations of query constraints selected interactively by the user. The references satisfying these constraint combinations are marked by dedicated colours, while the remaining references are shown uniformly. Thus, in our example in Fig. 4.104C, the user might wish that the districts belonging to the “white” and “blue” classes were shown in some neutral colour such as light grey. This would facilitate the user’s concentration on the “yellow” and “green” classes. When the number of query constraints increases, the advantages of such “selective marking” become more evident.

#### **4.6.2.3 Marking Versus Filtering**

The question arises: which technique is better, marking or filtering? As in almost all situations, there is no straightforward answer. Each technique has its advantages and its shortcomings. The main advantage of marking is its potential for a more refined representation of query results. A filtering tool always divides the reference set into two classes: one class corresponds to the satisfaction of all query constraints and the other class unites all remaining cases, from the failure of just one query constraint to none of the constraints being satisfied. With marking, these remaining cases can be separated. However, the number of different classes that can be considered concurrently is limited by human cognitive capabilities. This necessitates the use of “selective marking”, which complicates the user interface of the query tool.

An advantage of filtering is the possibility to simplify data displays by reducing their information content when it is necessary to concentrate on a subset of the data. This, however, entails a disadvantage: it is impossible to compare several subsets satisfying different query constraints or constraint combinations.

---

<sup>22</sup> Generally,  $2^N$  colours are needed to represent the satisfaction of all different combinations of  $N$  query constraints.

The approach of marking may reveal further benefits when applied to big datasets which do not allow truly dynamic querying, i.e. immediate response to query modification. With filtering, the user would have to wait each time when a query condition was added or removed. An appropriate solution would a combination of the following approaches:

- Data aggregation and display of the aggregates, for example by means of histograms.
- Formulation of a query with multiple constraints connected by the conjunction “AND”.
- Scanning the database and counting, for each aggregate, the number of references satisfying each combination of query conditions. This process will obviously take some time, and the user will have to wait.
- Selective marking on the displays of aggregates using the counts so obtained.

When the above counts are available, the displays can react immediately to selection by the user of particular combinations of query constraints by marking the corresponding reference subsets or, more precisely, showing the proportions of the references that satisfy these constraint combinations in each of the aggregates displayed. Thus, when histogram displays are used, these proportions are shown by means of bar segmentation as in Figs 4.101, 4.103, and 4.104C.

Hence, after the initial preparation stage, the analyst may use quite dynamic facilities to examine reference subsets with different properties and compare these subsets. As long as the analyst does not need to change the current set of constraints but only selects different combinations of those constraints, no delays are involved. The explorer needs only a convenient user interface for making such selections and choosing colours for marking.

We are not aware of any practical realisation of these ideas. The existing tools for querying large datasets represent query results mostly statically. An interesting example is a tool called InfoCrystal (Spoerri 1999). This allows the user to specify a query with several conditions. In the result, it shows the number of objects satisfying each possible combination of these conditions. The user can then consider any subset of objects in more detail. The most interesting feature of the tool is its particular method of graphical representation of the possible constraint combinations, which utilises the idea of the Venn diagram.<sup>23</sup>

---

<sup>23</sup> As a reminder, a Venn diagram is made up of two or more overlapping circles. It is often used in mathematics to show relationships between sets.

It is hard to say which of the techniques, filtering or marking, is easier or more pleasant to use; this depends strongly on the specifics of the realisation. Nevertheless, some user studies have been conducted, such as a comparison of one of the recent implementations of the Dynamic Query tool, which employs filtering, and a histogram-brushing tool (similar to Attribute Explorer), which uses marking (Li and North 2003). The conclusion from this comparison was that brushing histograms was superior for more complex discovery tasks such as revealing attribute correlations, comparing objects according to multiple criteria (attributes), and evaluating the position of a particular object among others in terms of its characteristics. Dynamic Query was superior for the simpler tasks of finding objects with characteristics that lie within certain ranges.

If we were asked which of the variants of the dynamic query tools is the best to have at the analyst's disposal, our answer would be quite straightforward: all of them. This includes "What's this?" questioning, filtering, simple and multicoloured marking (brushing) through direct manipulation of various data displays and through using specialised devices for setting query constraints, and selective marking to indicate the satisfaction of particular combinations of query constraints. Then, the analyst may choose which of these tools is the most suitable for a particular task or use several tools in combination, as, for example, a "What's this?" interrogation was used together with Dynamic Query in Fig. 4.97.

#### 4.6.2.4 *Relations as Query Results*

We are not aware of any existing query tools that specialise in determining relations between references or between characteristics, except for distances and other metric relations, which are expressed by means of numbers. It is quite logical to assume that an analyst can perform various comparisons effectively and make qualitative judgements concerning relations easily when an appropriate data representation is provided. For example, when values of a numeric attribute are represented by positions along a common coordinate axis, it is a trivial task to determine which of them are greater than others. Topological and directional relations in space are easily perceived from a map display.

Sometimes, however, it may be difficult to determine relations through mere observation of a data display. In this case, display manipulation may be helpful. Thus, in an unclassified choropleth map, two values of an attribute may be represented by rather close colour shades, and it may be hard to say which of the values is greater. Analogous problems may arise when attribute values are represented by symbol sizes: it may be difficult to judge which symbol is larger. In both cases, the problem may easily be

solved by applying the display manipulation technique of visual comparison, as was illustrated in Figs 4.34C and 4.35C. A single mouse click may transform the display in such a way that multiple comparisons are facilitated at the same time. For example, one district of Portugal may be compared simultaneously with all other districts: the districts with greater values than the value in this district and the districts with smaller values will be marked by different colour hues.

A need for a special tool arises when it is insufficient just to ascertain the existence of relations but it is necessary to determine their numeric characteristics. For example, one may need not only to estimate whether two attribute values are the same or different but also to find the precise difference between them. For two spatial objects, it may be insufficient to judge whether they are located close together or far from each other but, instead, it may be necessary to measure the exact distance between them.

Finding differences or ratios between numeric values is usually done by means of appropriate data transformations. For determining spatial distances, specialised tools for measuring distances are typically provided in software systems and packages that deal with spatial data.

When the outcome of measuring or computing a metric relation consists of just a single number, there is usually no need to find any special method to represent it: it may be shown as it is, for example in a special field of the display area or in a pop-up window. A particular visualisation method may become necessary when a tool computes multiple metric relations, for example, pairwise distances between several cities. An appropriate representation for this kind of information could be a distance matrix. Such matrices are often included in street atlases.

What was said above refers to situations where comparison is performed in terms of values of a single attribute (numeric or spatial). However, it is often necessary to compare references with respect to multiple characteristics. It is difficult to do such comparisons merely visually, and some computational tools have been developed to support this task. The basic idea is to somehow aggregate the differences between the values of multiple attributes corresponding to two distinct references into a single number. The resulting number is assumed to indicate the degree of similarity (or dissimilarity) between these two references. This degree of (dis)similarity is often called the distance between the two references in the abstract multi-dimensional space formed by all possible values of all attributes. Here, the term “distance” is a generalisation of the notion of a metric distance in geographical space.

One of the possible distance measures is the Euclidean distance, which is computed, for two references  $X$  and  $Y$  and  $n$  different numeric attributes, according to the formula

$$\sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4.8)$$

where  $x_i$  is the value of the  $i$ th attribute corresponding to the reference  $X$  and  $y_i$  is the value of the same attribute corresponding to the reference  $Y$ . The attributes participating in the computation of the distance must be previously standardised so that their value ranges become comparable. Some methods of standardisation are considered in Sect. 4.5.1.1.

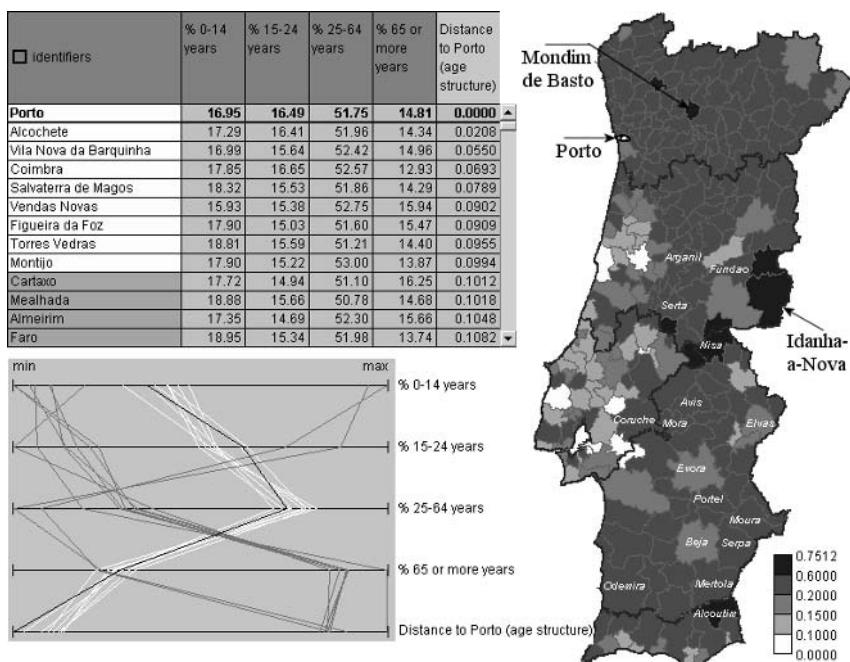
A generalisation of the Euclidean distance formula is the family of distance measures called Minkowski metrics, after a German mathematician Hermann Minkowski (1864–1909). The generalised formula looks as follows:

$$\left[ \sum_{i=1}^n |x_i - y_i|^r \right]^{\frac{1}{r}} \quad (4.9)$$

When  $r = 2$ , this is the typical Euclidean distance. The case when  $r = 1$  is often called the “city block distance” or “Manhattan distance”.

Besides these distance measures, other measures for characterising the degree of similarity between objects or observations in terms of multiple attributes have been suggested. While some of them have been specially invented for measuring dissimilarities with respect to qualitative characteristics, most of these measures deal with numeric attributes. There are also measures specially developed for the comparison of time-series data. Of all of the distance measures, Minkowski metrics are the most often used, and tools for computing them can be found in a number of software systems for data analysis.

Let us give an example of the use of such a tool for computing distances between references in terms of multiple attributes. In our Portuguese census data, we would like to see how similar the age structure of the population in various districts is to that in Porto. For this purpose, we start a distance computation tool, which allows us to choose the reference district, the set of attributes that will participate in the evaluation, and the metrics to be used. So, we choose Porto as the reference district and the attributes “% 0–14 years”, “% 15–24 years”, “% 25–64 years”, and “% 65 or more years” to characterise the age structure of the population. From the metrics available, we choose the Euclidean distance. In response, the tool produces a new attribute with values that are the Euclidean distances of all the districts of Portugal to Porto with respect to the age structure characteristics. The computation results are presented in Fig. 4.105.



**Fig. 4.105.** Results of computing the Euclidean distances of all districts to Porto with respect to the age structure of the population of the district. The rows of the table (top left) are arranged in order of increasing distance. The characteristics of the districts most similar to Porto can be seen in the table fragment visible. On the map, the districts are divided into five classes according to their distances to Porto. White is used to indicate the districts most similar to Porto, and the darkest shade of grey corresponds to the most dissimilar districts. In the parallel-coordinates display (bottom left), profile lines represent the characteristics of the most similar and most dissimilar districts. The black line corresponds to Porto itself

The computed distances range from 0, which corresponds to Porto itself, to 0.7512, which corresponds to the district of Idanha-a-Nova, which is located in the east of Portugal. The numbers as such are hardly interpretable; more meaningful and important are the relative positions between the minimum and the maximum distances.

The names and characteristics of the districts most similar to Porto can be seen in the table display in the upper left corner of Fig. 4.105. The rows of the table are arranged in order of increasing distance from Porto. It can be noted that the percentages of the four age groups in these districts are, indeed, very close to those in Porto.

On the map, the districts are classified according to their distances to Porto. White corresponds to the class of districts closest (i.e. most similar) to Porto; more precisely, with distances below 0.1. Two light shades of

grey are used to indicate the districts that are a little farther from Porto but still not too far. The darkest colour marks the districts most dissimilar to Porto.

In the parallel-coordinates display in the lower left corner of Fig. 4.105, only the districts in the classes with the lowest and highest distances to Porto are represented. The profile lines show the age structures in those districts and their distances to Porto. The white lines correspond to the districts closest to Porto. It can be seen that these lines are quite similar to each other and to the black line, which represents Porto itself. The shapes of the lines show that this group of districts is characterised by medium proportions of children and young people, rather high (but not extremely high) percentages of people aged from 25 to 64 years, and rather low (but not extremely low) percentages of elderly people.

The dark grey lines represent the districts most distant (i.e. most dissimilar) from Porto. From the shapes of the lines, we can detect two different age structure patterns in this group of districts: a “young population” (with a very high percentage of children and young people, a very low percentage of people aged from 25 to 64 years, and a quite low proportion of elderly people) versus an “old population” (with a very low proportion of children and young people and a very high proportion of elderly people). It is interesting that the proportions of the age group from 25 to 64 years are quite low in both subgroups.

This observation prompts us to look for other districts that can be characterised as “young” or “old”. For this purpose, we choose from the districts most dissimilar to Porto two representative districts with “young” and “old” population structures. To represent the “young” districts, we choose the district of Mondim de Basto, situated in the northern part of Portugal (its location is indicated in Fig. 4.105). As a sample of the “old” districts, we choose the district of Alcoutim in the south-east of the country (it is labelled on the map in Fig. 4.105). We prefer Alcoutim to Idanha-a-Nova, which is the most dissimilar to Porto, because the latter district seems to be extremely “old”, and there is a risk that very few districts will have a similar population structure. The profile of the district Idanha-a-Nova in the parallel-coordinates display lies quite apart from the profiles of the other districts (this is the line crossing the two bottom axes at their right ends). The characteristics of the districts of Mondim de Basto, Alcoutim, and Idanha-a-Nova can also be seen in the table fragment shown in Fig. 4.106.

We employed the same distance-computing tool to compute the distances from all districts to Mondim de Basto and to Alcoutim with respect to the four attributes characterising the age structure of the population of the district. After that, we applied multicolour marking to compare three

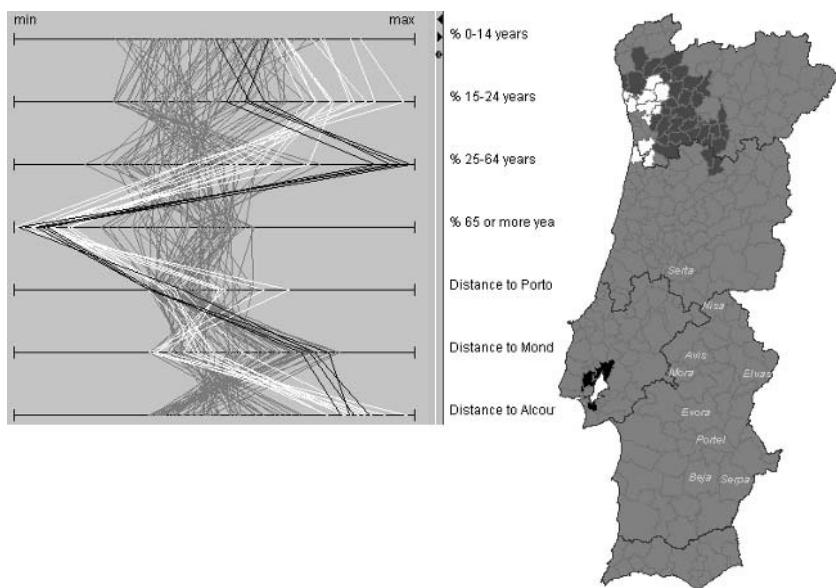
groups of districts: districts similar to Porto, districts similar to Mondim de Basto, and districts similar to Alcoutim. To define the groups, we assumed a district  $X$  to be similar to the model district  $M$  (where  $M$  is one of the districts of Porto, Mondim de Basto, and Alcoutim) if the distance from  $X$  to  $M$  lies in the first one-third of the range of distances from  $M$  to all other districts. Thus, the range of the distances to Porto is from 0 to 0.7512, the distances to Mondim de Basto range from 0 to 0.7212, and the distances to Alcoutim range from 0 to 0.8205. Accordingly, a district was classified as similar to Porto if its distance to Porto was up to 0.25, similar to Mondim de Basto if its distance to Mondim de Basto was up to 0.24, and similar to Alcoutim if its distance to Alcoutim was up to 0.27. If none of these conditions was fulfilled, the district was treated as dissimilar to any of the three model districts.

<input type="checkbox"/> Identifiers	% 0-14 years	% 15-24 years	% 25-64 years	% 65 or more years	Distance to Porto (age structure)
Gaviao	11.13	11.37	45.30	32.20	0.6251 ▲
Vila Velha de Rodao	11.15	11.65	44.96	32.24	0.6256
Vila de Rei	14.24	8.82	46.08	30.87	0.6314
Nisa	11.80	10.57	45.48	32.17	0.6324
Povoa de Lanhoso	27.50	17.94	41.43	13.13	0.6354
<b>Alcoutim</b>	<b>12.75</b>	<b>10.48</b>	<b>44.70</b>	<b>32.07</b>	<b>0.6363</b>
<b>Mondim de Basto</b>	<b>26.66</b>	<b>19.76</b>	<b>40.25</b>	<b>13.33</b>	<b>0.6694</b>
<b>Idanha-a-Nova</b>	<b>12.07</b>	<b>9.58</b>	<b>43.15</b>	<b>35.20</b>	<b>0.7512 ▼</b>

**Fig. 4.106.** In the bottom part (shown here) of the table sorted according to distances to Porto, the characteristics of the districts with age structures most dissimilar to Porto are presented. The top fragment of the table is shown in Fig. 4.105

The resulting groups of districts are presented in a map and parallel coordinates display in Fig. 4.107C. The districts similar to Porto are coloured yellow, those similar to Mondim de Basto are red, and those similar to Alcoutim are green. The districts that are not similar to any of the three model districts are shown in grey.

From the parallel-coordinates display, it can be noted that the age structures in the groups of districts similar to Porto, Mondim de Basto, and Alcoutim are quite consistent. Filtering has been applied to the display in order to hide the lines for the districts that do not belong to any of the three groups. The characteristics of the remaining districts can be seen in another screenshot of the same parallel-coordinates display, which is presented in Fig. 4.108 on the left. Here, the filter conditions have been specified so that only the lines for the districts *not* included in any of the three similarity groups remain visible.



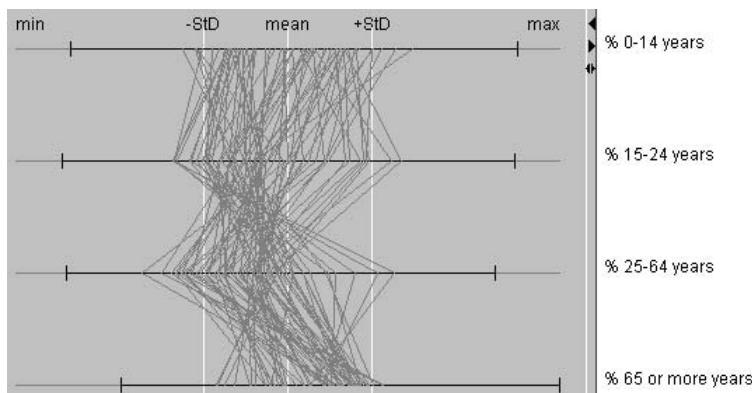
**Fig. 4.108.** The parallel-coordinates display (left) represents the age structures in the districts that were not included in any of the three similarity groups shown in Fig. 4.107C. On the three axes at the bottom, the distances of these districts to Porto, Mondim de Basto, and Alcoutim are portrayed. Two subgroups of districts with the largest distances to Alcoutim have been detected; the corresponding lines are coloured in white and black. On the right, the geographical positions of these subgroups of districts are shown on a map

When we examine this display, we note that while the distances from all the districts to Porto and to Mondim de Basto lie in the middle of the corresponding distance ranges (the second and third axes from the bottom), the distances from some of the districts to Alcoutim are very large. The lines for these districts cross the bottom axis, which represents the distance to Alcoutim, close to its right end. By means of a direct-manipulation query tool, we have marked these lines in the parallel-coordinates display and notice that all the corresponding districts have very low percentages of people aged 65 or more years. This can be seen from the fourth axis of the display. However, the districts split into two subgroups with respect to the other three age groups. In Fig. 4.108, these subgroups are marked in different ways. White marks the profiles of the districts with high proportions of the age groups 0–14 years and 15–24 years and medium proportions of people aged from 25 to 64 years. Black marking is used for the districts with very high proportions of the age group 25–64 years and medium proportions of children and young people. The geographical positions of the two subgroups of districts are indicated on the map on the right in Fig.

4.108. The “white” group forms two compact clusters in the north-west of Portugal. The “black” group is located in the central part of the country, close (geographically) to the capital, Lisbon.

It can be seen in the parallel-coordinates display that the “white” and the “black” group are distinctly separated according to their distances (in the sense of similarity of the age structure) to Mondim de Basto. These distances are represented on the axis second from bottom. The “white” districts are more similar to Mondim de Basto than are the “black” ones. It is interesting that the “white” districts are also geographically close to Mondim de Basto and to the districts classified as similar to this model district. On the map in Fig. 4.108, the districts of the Mondim de Basto group are shown in dark grey.

The remaining districts are characterised by medium values of the four age structure attributes. In Fig. 4.109, their characteristics can be compared with the mean values of these attributes. The axes of the parallel-coordinates display have been scaled and aligned so as to facilitate this comparison.



**Fig. 4.109.** The axes of the parallel-coordinates display have been transformed here to facilitate comparison of the characteristics of the districts not included in any of the groups with the means of the age structure attributes

In this example, we have applied a tool for computing the degree of dissimilarity, or distance, between characteristics of different references in order to discover groups of references with similar characteristics. The major benefit of using this tool is the opportunity to consider multiple attributes simultaneously, which is generally not an easy task.

We have mentioned the existence of many methods for measuring the degree of dissimilarity but have used only one of them in our example, specifically, the Euclidean distance. We are not ready to give any general

recommendations concerning what distance measures to use in different cases. We personally take a pragmatic approach: we apply the measure proposed by the tool as its default option and then look at a suitable graphical display, such as parallel coordinates, to see whether the results of the computation agree with our intuitive conception of what can be considered as similar. In our example investigation, the Euclidean distance served quite well; otherwise, we would try out other measures.

A more or less general recommendation would be not to apply measures specially developed for particular types of data (e.g. time series) to data of other types. On the other hand, if you have a tool capable of computing similarity measure that was specifically designed or optimised for the type of data that you need to analyse, it seems reasonable to try that measure first.

Let us now recall the context for considering this example of exploration of a population structure. We have discussed what answers may be given to comparison (i.e. relation establishment) queries and how they can be presented to the analyst. We have mentioned that special, comparison-oriented query tools are typically used when it is necessary to determine numeric characteristics of relations, for example distances as a numeric measure of a similarity/dissimilarity relation, whereas qualitative judgements concerning the presence of this or that relation are usually made through observing and manipulating appropriate data displays.

A need for visualisation and analysis of the output of a comparison tool arises when this output is not just a single number but consists of multiple numbers. In our example, the output was the set of distances from all districts of Portugal to a selected district. Such an output may be treated as a new attribute and hence visualised and analysed like any other numeric attribute, alone or together with other attributes. We have demonstrated some approaches to analysing this kind of derived data. We have applied visualisation methods such as classified choropleth maps, table displays, and parallel-coordinates displays in combination with dynamic query tools that enable filtering and multicolour marking.

### 4.6.3 Non-Elementary Queries

At the beginning of Section 4.6, we said that, typically, query tools can provide answers only to elementary questions. However, some of our examples of the use of query tools seem to contradict this statement. In these examples, we explored spatial and statistical distributions of various characteristics and tried to discover correlations between attributes. These are undoubtedly synoptic tasks. How can we explain this contradiction?

A close look at the examples presented above may lead one to notice that the query tools were never used alone but were used together with other tools, in particular, visual data displays. It was these displays that allowed us to derive synoptic information from the outcomes of the query tools. We could hardly do anything similar using just a list of references satisfying our query conditions, even though such a list (if not too long) could be quite sufficient for the task of choosing a hotel to stay in for a night.

The role of the query tools in our example analyses was that they were used for *defining subsets of references* with specific characteristics. Then, we looked at these subsets (with known characteristics) on various displays in order to grasp the distribution patterns of these characteristics. In so doing, we considered each subset as a unit rather than attend to its individual members, hence the synoptic level of analysis.

It is this special way of using query tools that makes them suitable not only for finding objects that satisfy somebody's requirements (such as hotels in the city centre with a price within a certain range) or finding particular information about particular objects (e.g. the job position and salary of Mr Smith) but also for sophisticated, synoptic-level tasks of exploratory data analysis. This method of use implies that an explorer defines, concurrently or sequentially, multiple subsets of references with different characteristics, rather than a single subset, as in a traditional search. This explains, in particular, why the query tools used for exploratory analysis need to be dynamic in the sense of allowing easy, quick switching from one subset to another.

There is another, slightly different way of using query tools for synoptic analysis tasks: under certain circumstances, such tools may allow one to search for specific behaviours. One such case arises when there is a specific representation of data consisting of multiple numeric time series, such as the dataset about crime in the states of the USA: for each state, we have a series of attribute values referring to 41 consecutive years.

Every such time series characterises the behaviour of a numeric attribute over a time period. As we have seen earlier, such a behaviour can be represented graphically by a line on a time graph. This allows one to reformulate a task of searching for a particular behaviour as a task of finding lines with a particular shape. The former task is synoptic with respect to time, whereas the latter is elementary: it refers to individual objects represented by lines, for example the states of the USA, and has no regard for time.

Hence, the representation of behaviours by geometrical objects (lines) allows one to replace a synoptic task with an equivalent elementary task. Since the task becomes elementary, a query tool may be designed to support it. The main idea of such a tool is quite straightforward: it should al-

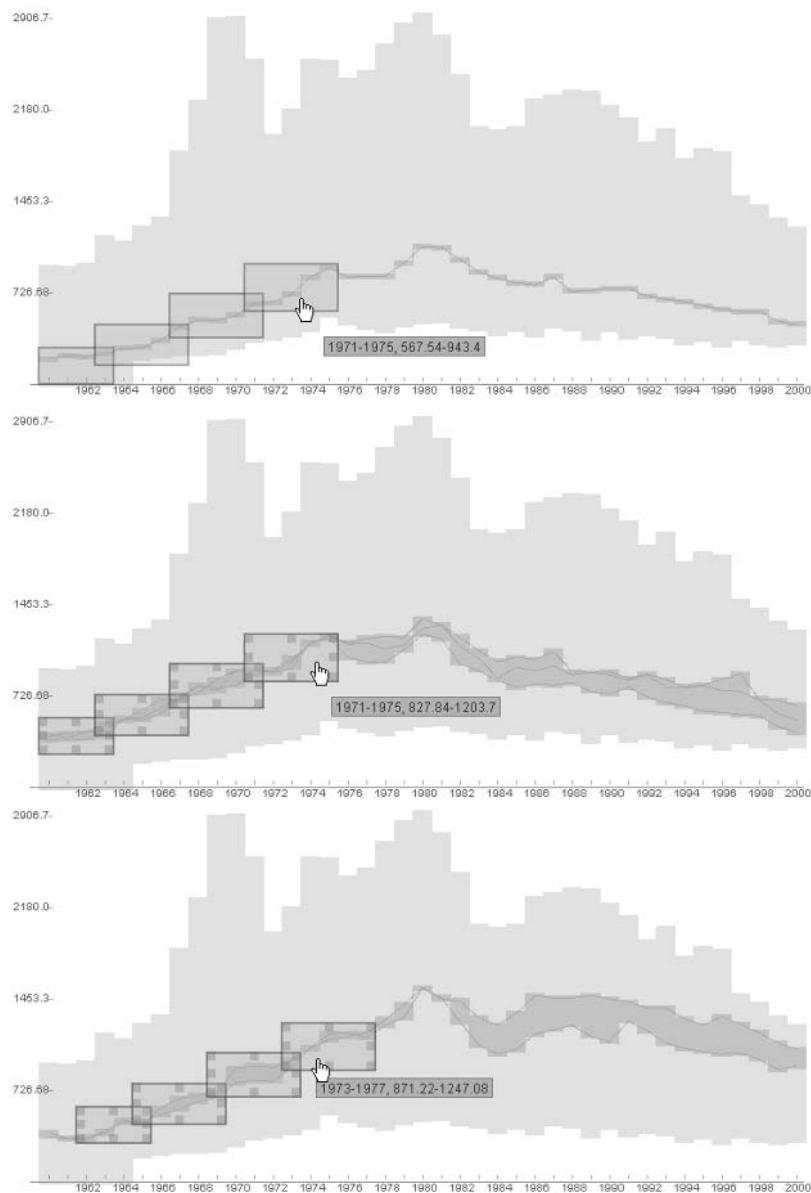
low the user to specify the shape to be searched for and then look for lines with this shape. The problem yet to be solved is how the user specifies the shape that he/she is interested in.

One approach is to allow the user to draw the shape that needs to be found, as is described in Wattenberg (2001). It should be understood, however, that such a drawing is just an approximate expression of the user's idea, and hence trying to find lines with exactly this shape may have no sense. The user, most probably, would like the tool to return him/her lines that have *similar* shapes to what has been specified. But what does the user mean by "similar"? Suppose, for example, that the user has drawn a line that initially goes up and then turns down. The line has numerous geometric characteristics: its length, its maximum height, its slope, the vertical positions of its beginning and end, the horizontal position of the maximum, the curvature at this position – are all of these characteristics important? Or does the user just want to find all lines that first go up and then go down, irrespective of how steeply and how high they rise and at what position they turn down? Depending on what is understood by "similarity", the search may be based simply on computing distances (for example, Euclidean distances) and comparing them with some threshold, or involve stretching and shrinking or other computationally intensive transformations of lines. Algorithms for the latter type of search are being developed in the research area of data mining (interested readers may be referred to the review by Keogh and Kasetty (2003)).

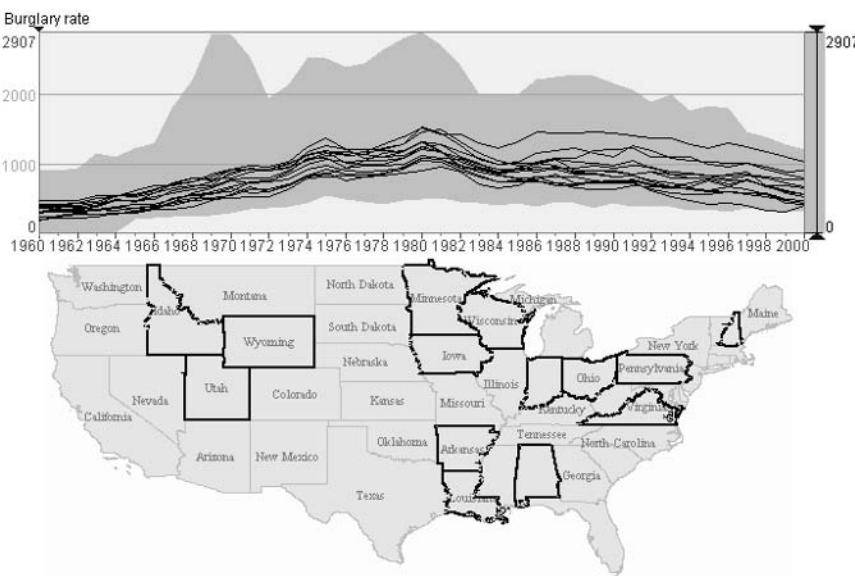
Hence, the seemingly simple and intuitive way of specifying a model shape to be searched for turns out to be quite awkward. Either it is necessary to figure out what the user really means (and hence complicate the user interface) or some simplifying assumptions have to be made.

A different approach is taken in the TimeSearcher tool developed in Ben Shneiderman's laboratory (Hochheiser and Shneiderman 2004). Instead of sketching a line, the user specifies the shape by drawing rectangles, called timeboxes, in various places over the area of a time graph. Each timebox plays the role of a filter: only lines passing through it remain active, and all other lines are hidden. The filtering occurs already during the process of box-drawing. Through timeboxes, the user conveys his/her ideas concerning the distinctive features of the shape that he/she is looking for, and simultaneously sees which of the available lines have those features.

Let us consider the example presented in Fig. 4.110. We have applied TimeSearcher to the USA crime dataset, specifically, to the attribute "Burglary rate". We have drawn four boxes (Fig. 4.110, top) such that each timebox, from left to right, is shifted to a higher vertical position with respect to the previous one. In the result, we have created a mask to find lines with an increasing trend.



**Fig. 4.110.** TimeSearcher has been used to find a 16-year increase pattern among the behaviours of the burglary rates in the states of the USA. The illustration was produced using the demo version of the “TimeSearcher” tool available at <http://www.cs.umd.edu/hcil/timesearcher>. This and further screenshots the TimeSearcher are used with permission of the Human-Computer Interaction Lab, University of Maryland, 2005



**Fig. 4.111.** Top: the states where the burglary rates increased during the period from 1960 to 1975 as specified by the mask in Fig. 4.110. Moving the mask in the vertical direction has allowed us to ignore the actual values of the burglary rates and reveal lines with similar shapes in the interval 1960–1975

The left side of the first box coincides with the beginning of the time period, i.e. the year 1960, and the right side of the last box corresponds to the year 1975; hence, the total length of the mask corresponds to a 16-year time interval, and it will help us to separate behaviours with a period of growth 16 years long. The vertical dimensions of the timeboxes specify the permitted ranges of variation of the values. Thus, in our example, the value at the beginning must be within an interval from 0 to about 290, and at the end between 567 and 943. From the screenshot at the top of Fig. 4.110, it may be seen that there is only one line satisfying our query.

If this were the only way of using timeboxes, we would not count TimeSearcher as an effective tool for detecting specific behaviours. We could achieve the same result with “ordinary” Dynamic Query by simply limiting the value intervals for the burglary rates in the years from 1960 to 1975. However, TimeSearcher allows one to achieve somewhat more than this, owing to the high manipulability and transformability of timeboxes. The user can easily move the mask as a whole or any of its components over the plot area, stretch or shrink it, and even flip it.

For example, in the middle of Fig. 4.110, we have moved the entire mask that we have built upwards. As may be seen from the picture, the rightmost timebox now specifies the value subrange from 827.54 to

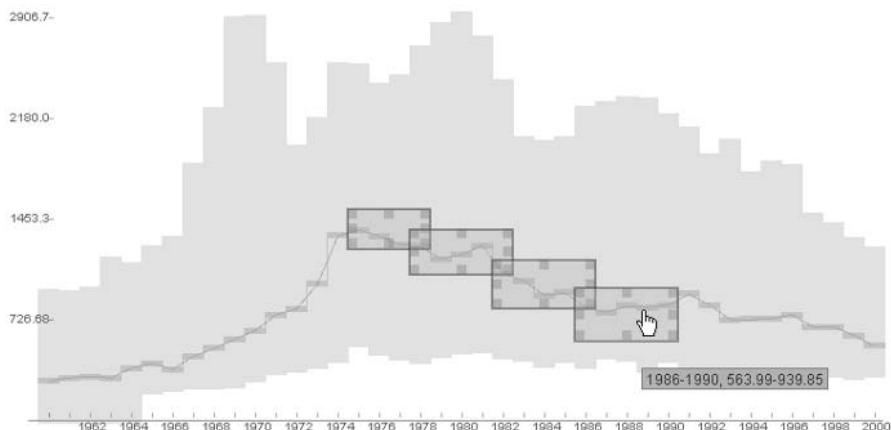
1203.7, instead of the initial 567.54 to 943.4. In response, the tool shows us three lines passing through the repositioned timeboxes. These lines have the same generally increasing trend during the years from 1960 to 1975 as the line found before (see the upper image), but the corresponding attribute values are higher. We can continue moving the mask in the vertical direction and thereby reveal other lines with similar shapes. The whole group of lines found in this way can be seen in Fig. 4.111, top. The map in the lower part of Fig. 4.111 shows the states represented by these lines. Unfortunately, TimeSearcher does not provide the possibility to store or mark findings for further analysis: all previous query results are lost after movement or modification of the mask. Therefore, we used another tool to see all of the results obtained with TimeSearcher together.

Let us return to Fig. 4.110. The image at the bottom demonstrates that a mask can be moved not only in the vertical but also in the horizontal direction. In so doing, we have revealed two lines with a 16-year-long increase starting in the year 1962 rather than 1960. Hence, moving a mask to the right or to the left allows one to detect line fragments with particular shape features irrespective of their relative positions on the time axis.

We can also stretch or shrink the mask in the vertical or horizontal direction. Stretching the mask in the vertical direction allows greater variability of the slopes of the lines but also permits higher fluctuations. Shrinking in the vertical direction, conversely, tightens the constraints and consequently reduces the variation and fluctuation in the query results. Stretching in the horizontal direction makes the mask longer. In our case, we could search for increases over more than 16 years. Conversely, shrinking may be used to search for shorter fragments that have the same general shape features.

Not only the whole mask, but also the individual boxes that it comprises, can be transformed. Thus, if we wished to search for line fragments with steeper growth, we would need to move each box except for the first one to a higher vertical position in relation to its left neighbour.

Flipping a mask inverts the shape specified by it. Thus, if the initial mask was meant to search for increasing trends, the transformed mask will be suitable for detecting decreasing trends. Figure 4.112 shows how the mask that we earlier created looks after flipping it and moving it to the right. The transformed mask has “captured” a line with a generally downward orientation in the interval from 1975 to 1990. It may be noted, however, that the line fluctuates greatly, especially inside the second box from the left. Besides, its behaviour within the rightmost box can hardly be classified as a decrease. If such deviations from the model shape are strongly undesirable, we need to specify a more refined mask using smaller boxes as its components.

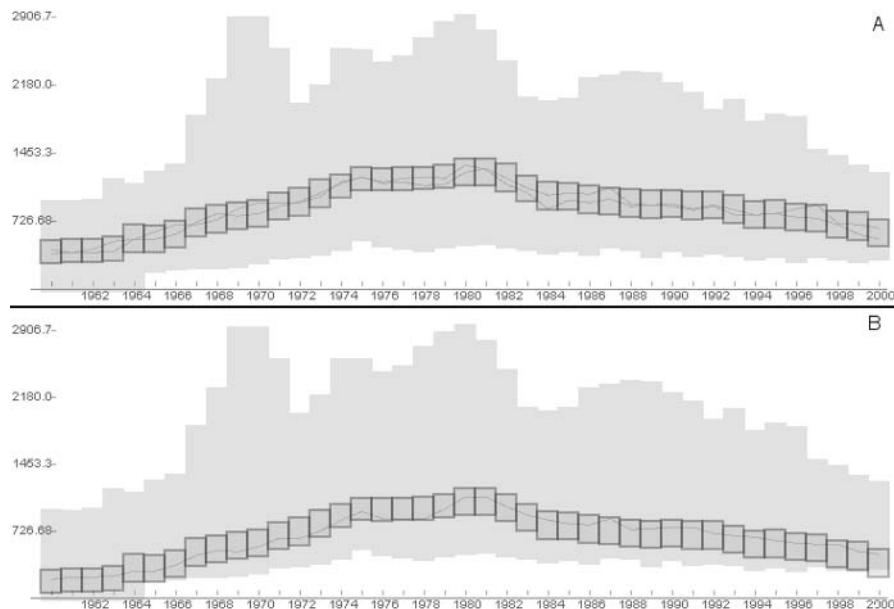


**Fig. 4.112.** After the mask in Fig. 4.110 has been inverted, it can be used to search for decreasing trends

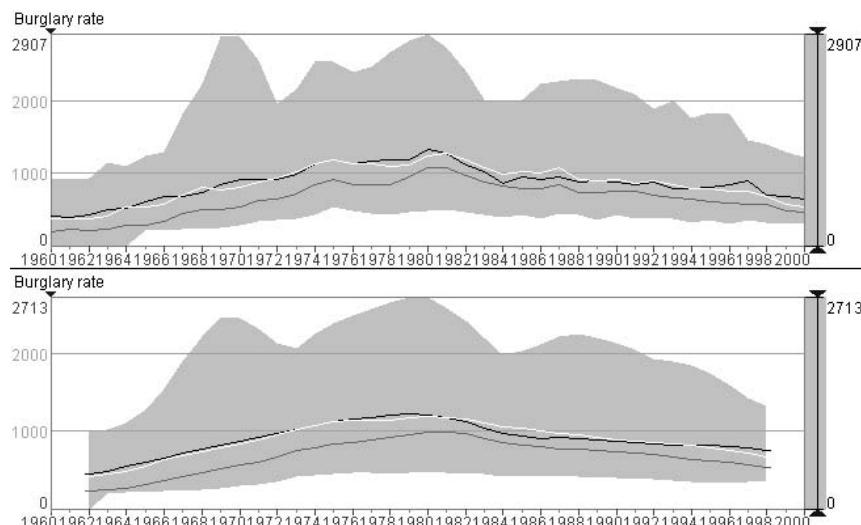
We would like to demonstrate one more opportunity provided by TimeSearcher: one may select any individual line and automatically derive a search mask from it, as is shown in Fig. 4.113. This search mask can be used (after some modification if needed) to search for lines similar in shape to the selected line.

In Fig. 4.113, we have chosen the line numbered 27, which corresponds to the state of Minnesota. The tool has built a mask consisting of as many boxes as there are different time moments, in our case 41. Each box has a certain default height that specifies the permissible range of variation of the values and is positioned vertically so that the value at the respective time moment lies in the middle of it. The top image in Fig. 4.113 shows the mask after a slight manual modification, which has made it a little smoother. The mask thus built has immediately “caught” another line, specifically, the line numbered 49, corresponding to Utah.

After that, we have moved the mask up and down and succeeded in finding one more line with a similar shape. This is line 55, representing Wisconsin. In Fig. 4.114, top, all three lines are shown together. For better distinguishability, the lines are coloured differently: the line for Minnesota is white, the line for Utah is black, and the line for Wisconsin is grey. The lower image presents the same lines after smoothing (the smoothing was done using the technique of the centred moving average over 5-year intervals). It may be seen that the shapes of all three lines are quite similar, especially after smoothing, which mitigates fluctuations. It is a pity that TimeSearcher itself does not provide the opportunity for line smoothing – it would make much sense to apply timeboxes to previously smoothed lines.



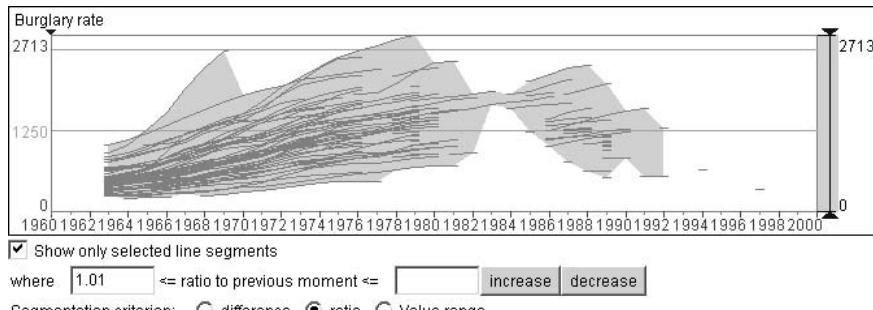
**Fig. 4.113.** A mask has been automatically built on the basis of a selected line. Using the mask, we have found two other lines with shapes similar to that of the selected line



**Fig. 4.114.** Lines with similar shapes revealed using the mask in Fig. 4.113. The upper image shows the original lines, while the lower image presents the same lines after smoothing

Despite some small criticism, we find TimeSearcher to be a quite powerful dynamic query tool for dealing with temporal behaviours of numeric attributes. The opportunities for interactive specification and easy modification of search masks allow the explorer to disregard irrelevant peculiarities of individual behaviours and find groups of behaviours with distinctive features, such as growth followed by a fall. It is important that no assumptions are made concerning what features are relevant and what are irrelevant for the user. The user has full control: he/she performs only those kinds of mask transformations that conform to his/her idea of what is relevant and what is irrelevant.

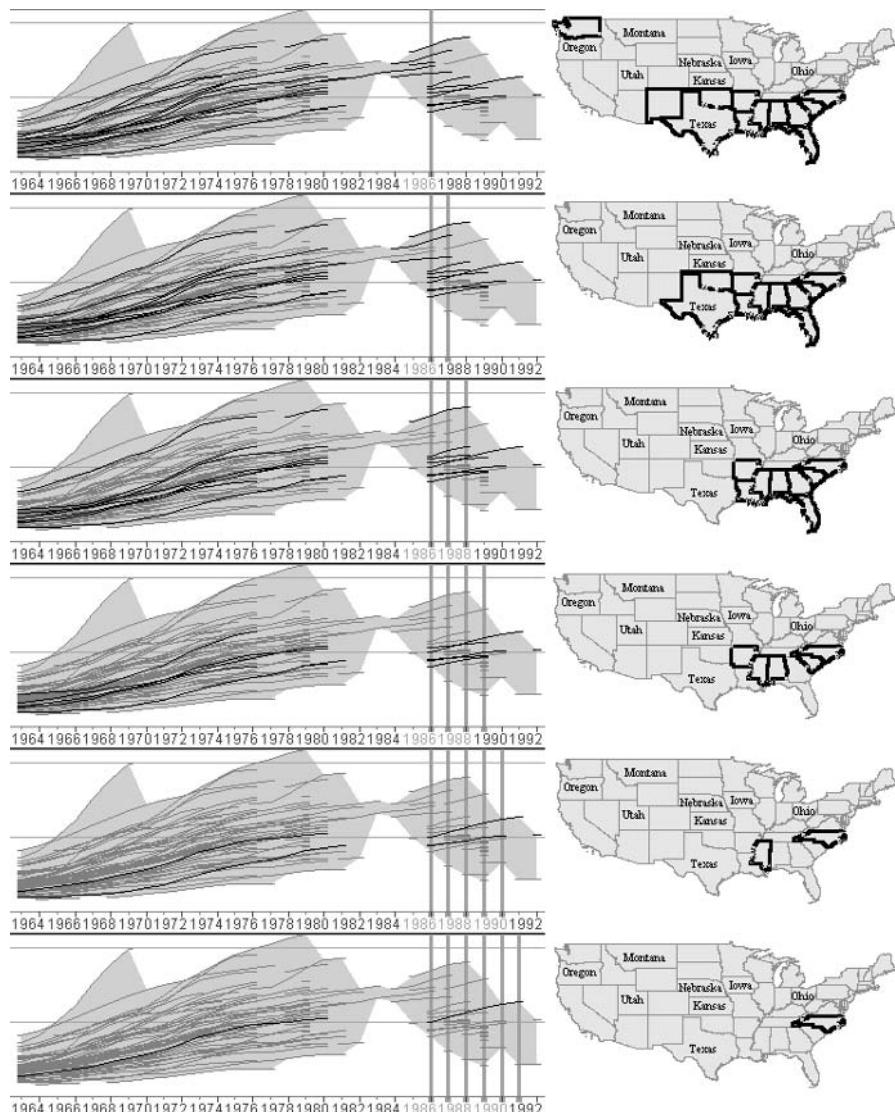
There may also be other approaches to working with geometrical representations of temporal behaviours. One of them is demonstrated in Fig. 4.115. The idea is to show on a time graph only line segments that have a certain inclination, which is specified through setting lower and/or upper limits on the degree of absolute or relative change (i.e. difference or ratio) in comparison with the previous moment. In Fig. 4.115, the user has specified the lower limit on the relative change to be 1.01, which corresponds to an increase of 1% or more in the current year in comparison with the previous year. In response, the tool shows only the line fragments complying with this specification; all other line fragments have been hidden. The time graph represents the same burglary rate data as before. The lines have been previously smoothed using a 5-year centred moving average, as in the lower image in Fig. 4.114.



**Fig. 4.115.** The query tool shows only line fragments with a specified inclination. In this example, the visible line segments correspond to an increase of no less than 1% increase (a ratio of 1.01) in comparison with the previous time moment

Now, it is possible to apply marking in order to find out which states had no less than 1% increase in the burglary rate in particular years. This can be done by clicking on the line segments, but there is also another opportunity: the user can click on the years, that is, on the positions corre-

sponding to different years below the horizontal axis of the graph. In the result, the lines with an increase of no less than 1% in these years become marked, and so do the corresponding graphical elements in other displays, in particular, the outlines of the states on a map display. Selection of two or more years marks the lines that have the specified inclination in all of these years.



**Fig. 4.116.** Looking for states with an increase of 1% or more in the burglary rate in consecutive years starting from 1986

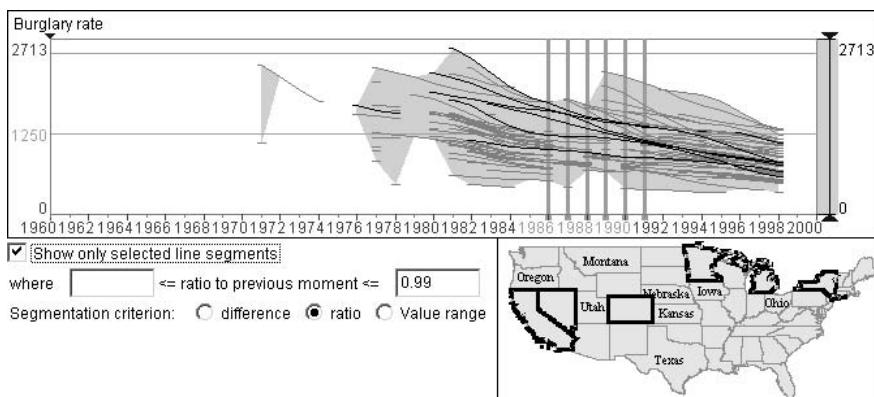
Figure 4.116 demonstrates the effect of a series of successive selections. The vertical lines on the time graph indicate the selected years. The lines with the specified degree of increase in these years are marked in black. To the right of each time graph there is a map with the corresponding states marked by thick black boundaries.

First, we selected the year 1986 and observed on the map which states had the specified increase in the burglary rate in that year. After that, we clicked on the next year. In the result, the marking of two states disappeared, and only the states with an increase in both 1986 and 1987 remained marked. Then, we clicked on the year 1988, and so on. With each subsequent selection, the number of marked states decreased. At the end, when six years from 1986 to 1991 had been selected, only one state remained marked. It may be seen from the graph at the bottom that the corresponding line fragment (shown in black) ends in the year 1991, and hence the selection of the year 1992 would remove the last marking.

Looking at the maps allows us to note some quite prominent spatial patterns formed by the marked states. Thus, the topmost map demonstrates a cluster of marked states in the south and south-east of the country. These are the states where burglary rates increased in 1986 in relation to 1985 by at least 1%. There is also one marked state separated from the others; this is Washington, in the north-west. In the map second from top, which shows the states with an increase in the years 1986 and 1987, Washington is no longer marked, and the cluster in the south-west has lost the state of New Mexico, west of Texas. In the next map (an increase in 1986–1988), the cluster has decreased further on its western side by losing Texas and Oklahoma, north of Texas. In the remaining three maps, the cluster ceases to exist. In the fourth map from top we see two small clusters, one with three and one with two marked neighbouring states, and in the next map only two spatially separated states are marked.

For comparison, we have looked for which states had a persistent value decrease in the same period from 1986 to 1992. For this purpose, we changed the settings of the tool: we specified an upper limit of 0.99 for the ratio between the value in each year and that in the previous year. As a result, we see only the line segments corresponding to a decrease of at least 1%. From Fig. 4.117 it may be seen that the number of descending line fragments in the 1980s and 1990s greatly exceeds the number of ascending ones, which are visible in Figs 4.115 and 4.116. During the years 1986–1992, there were six states with persistently decreasing burglary rates. Moreover, the decreasing trend in all of these states had already begun in 1983, and in some of them even earlier. Furthermore, the decreasing trend in all of these states, except for Nevada continued until 1998 (the line “tails” corresponding to the years 1999 and 2000 have been cut off in the

result of the smoothing). The decrease in the burglary rate in California and Colorado started as early as in 1980. We have discovered all these facts through interaction with the tool, but, to save space, we have not illustrated each of these findings by appropriate screenshots, in the hope that the idea is already sufficiently clear.



**Fig. 4.117.** Finding the states where the value decreased by at least 1% in each of the years from 1986 to 1991 in relation to the previous year

Using the example of TimeSearcher and the line segmentation tool, we have demonstrated the possibility of using query tools to search for specific behaviours in a case where each behaviour can be represented by a single geometrical figure. In this case, the task of behaviour (pattern) search appears as a search for geometrical figures with specific shapes.

Both tools that we have considered deal with data consisting of a numeric attribute and two referrers, one of which is temporal and the other of which is treated as a population referrer. We are not aware of any existing tools where similar ideas have been applied to other types of data. We can try to imagine what tools could be created. Thus, a behaviour over a population-type reference set can be viewed as a certain distribution of value probabilities, which can be represented graphically by a histogram or a probability density function. Hence, to search for such behaviours, one can specify the geometrical properties of the histogram or density function corresponding to the behaviours of interest. Distinctive features of behaviours (distributions) in two-dimensional space could be represented by sketches, such as those in Figs 4.34C and 4.37. However, it may be hard to design an easy and convenient user interface for making such sketches. Besides, an automatic search for spatial patterns complying with a user-provided sketch may be rather computationally intensive.

#### 4.6.4 Recap: Querying

The process of exploratory data analysis consists of finding answers to numerous questions about the data. In Chap. 3, we have suggested a typology of such questions, also called data analysis tasks. To obtain answers to the questions, an analyst uses tools. In most cases, the tools do not provide direct and full answers to the analyst's questions but only supply some convenient devices that can help the analyst to find the answers through observation combined with imagination and reasoning. The analyst typically does not even put any explicit question to a tool but keeps the questions in his/her mind.

However, for some types of questions, specific software tools for finding answers may be created. Such tools are called *query tools*. A query tool allows an analyst to formulate his/her questions explicitly using some language or using interactive graphical widgets. The tool responds by giving direct answers to these questions. The answers are presented so as to be readily perceived and require no additional mental effort.

Most of the existing query tools are intended to give answers to elementary questions, i.e. questions that entail searching for individual objects, locations, time moments, etc. and the corresponding attribute values. This does not mean that a query result cannot consist of a number of objects, locations, or time moments; it means only that the query tool does not deal with this collection of references as an integral entity and does not treat the corresponding collection of attribute values as a behaviour. It is the analyst who can, in the process of further analysis, consider this group of references as a united whole, try to grasp the corresponding behaviours of the attributes, and compare these with behaviours based on other reference subsets and on the entire set. We have demonstrated such analyses of query tool outcomes by examples.

However, there are also query tools that search for certain types of behaviours. This becomes possible when a behaviour can be represented as a single graphical object. For example, a single line on a time graph may represent the behaviour of a numeric attribute over a time period. A line on a map may portray the trajectory of movement of some object such as a stork or a vehicle. There may be many such behaviours, and an analyst may need to search among them for behaviours with particular characteristics. The graphical representation of the behaviours makes it possible to substitute this search task by an equivalent task of searching for graphical objects with particular features such as a particular shape, size, or colour.

Query tools may be characterised and classified according to a number of criteria:

(A) What types of questions they answer:

- (1) direct lookup: Find attribute values corresponding to given references;
- (2) inverse lookup: Find references characterised by given attribute values;
- (3) direct comparison: Measure the degree of similarity/difference between characteristics of given references;
- (4) relation-seeking: Find references with specified relations between their characteristics;
- (5) pattern search: Find behaviours with given characteristics, and the corresponding reference sets;
- (6) behaviour comparison: Measure the degree of similarity/difference between behaviours based on given reference sets.

(B) How the questions are asked:

- (1) query language, formal or visual;
- (2) direct manipulation of graphical displays:
  - (i) mouse pointing (“What’s this?”);
  - (ii) selection of display elements, e.g. for subsequent marking (brushing);
  - (iii) drawing within the graph area, e.g. to specify a mask for filtering;
- (3) special graphical user interface controls such as sliders, switches, selection lists, etc.

(C) In what form the answers are given:

- (1) query results are presented independently of the previous content of the screen (e.g. in additional windows);
- (2) the previous content of the screen changes:
  - (i) new items representing query results are added;
  - (ii) items not satisfying query constraints are removed (filtered out);
- (3) the appearance of the display(s) present on the screen changes:
  - (i) items satisfying the query are marked;
  - (ii) items representing aggregates are segmented, and some segments are then marked.

(D) What mode of query building is assumed:

- (1) the user builds a complete query, and each subsequent query is independent of the previous one(s);
- (2) the user starts with a rough or partial query and then iteratively refines and modifies it, taking into account the feedback received from the tool.

**Table 4.12.** A summary of popular dynamic query tools and techniques

Name	References	Profile	Intended usage
Brushing	(Buja et al. 1991) (Chen 2003)	A1,2; B2ii; C3i,ii; D1,2; E1	Select a subset of references in one display and look for where they are in other displays. Select two or more subsets and compare their positions in different displays.
Dynamic Query	(Ahlberg et al. 1992) (Norman et al. 2003)	A2; B3; C2ii; D2; E1	Specify a combination of intervals of attribute values and look for where the corresponding references are in different displays.
Attribute Explorer	(Spence and Tweedy 1998) (Spence 2001)	A2; B3; C3ii; D2; E1	Specify intervals of attribute values and observe in segmented histograms what values of other attributes co-occur with values from these intervals.
Time Wheel; Temporal brushing	(Edsall and Peuquet 1997) (Harrower et al. 1999) (Monmonier 1990)	A1,2; B3; C2ii; D1; E1	Choose a specific time of day, day of the week, or month of the year and study what happens at this time over many days, weeks, or years (e.g. on an animated map). Helps to disregard cyclic fluctuations and consider longer-term trends.
Temporal focusing	(Harrower et al. 1999) (Monmonier 1990)	A1,2; B3; C2ii; D1; E1	Choose a time interval for consideration, e.g. on an animated map.
TimeSearcher	(Hochheiser and Shneiderman 2004)	A5, B2iii, C2ii, D2, E1	Find lines on a time graph with particular shape features specified by means of a mask, which may be moved over the display area.

(E) How reactive and dynamic the tool is:

- (1) the tool immediately updates its results in response to any user's operation of setting or modifying a query constraint, possibly during the very process of performing the operation;
- (2) the search may be prolonged (because of a large data volume) but the results allow dynamic manipulation to explore the satisfaction of individual query constraints and various constraint combinations;
- (3) dynamic behaviour is not provided.

For exploratory data analysis, highly reactive and dynamic tools are especially valuable, together with the possibility to set and modify query constraints easily and quickly. These requirements arise from the main purposes of using query tools in exploratory analysis, specifically, studying the distribution of characteristics (i.e. the behaviours of attributes) over a reference set and revealing relationships between attributes. These purposes imply two primary modes of use of query tools:

- repeated definition of various reference subsets for subsequent consideration and comparison of the corresponding attribute behaviours;
- repeated specification of various combinations of characteristics for subsequent investigation of their relatedness to other characteristics pertaining to the same references.

It is the repetitive way in which these tools are used that calls for ease of query construction and modification and for fast feedback from the query tool. We call tools satisfying these requirements “dynamic query tools”.

Some widely known dynamic query tools and techniques have been described or mentioned in this section. Table 4.12 gives a brief summary of these and some other tools. The column “Profile” characterises the tools in terms of the criteria enumerated above.

Some query tools involve not only a search in a database but also quite intensive computation, for example when determining the degree of similarity of references with respect to multiple attributes or finding lines with shapes similar to a given sketch. Together with some other computation-based techniques for data analysis, such query tools are part of the field of data mining, which will be discussed briefly in the next section.

## 4.7 Computational Tools

In this section, we are going to overview analysis techniques that rely significantly upon computation. These techniques are different from the data manipulation tools discussed earlier. While data manipulation is used for preparing data for subsequent analysis (whereas the analysis itself is done using other tools), the computational tools that we are going to discuss here are intended to produce something that can already be treated as a result of analysis since it provides a certain kind of generic information about the entire dataset or a substantial part of it. This may take a form of a formula, a logical expression, a classification, or just a single numeric measure. The application of other tools to such a result is, in principle, not required, although it is not excluded.

There are two major groups of computational tools intended for data analysis: statistical and data-mining tools. These groups of tools originate from two big research disciplines, which have made great advances and are continuing to develop. The tools (i.e. methods and algorithms) are very numerous and are described in thick books. Thus, even a book on elementary statistics (Burt and Barber 1996) has 640 pages! We therefore find it completely unfeasible to give a brief but still useful description of statistical and data-mining methods in this book. What is even more important, both groups of methods require deep understanding in order to be used properly. A very brief description cannot give the required level of understanding to inexperienced readers (and would be worthless for experienced readers), but it could encourage them to use these methods, which could lead to totally wrong conclusions. We would like to avoid this.

What we can try to do is to present some general thoughts concerning the possible purposes of using statistical and data-mining methods in exploratory data analysis and the possible benefits from them. If this arouses interest in readers who have not used these methods before, those readers will find opportunities to learn more about them. Many educational materials are available not only in books but also on the Internet. Thus, we can recommend the *Electronic Statistics Textbook* (StatSoft 2004), which explains both statistical and data-mining methods.

We can also present some examples from our practical experience. It happens that we have never used any sophisticated methods of statistical analysis but only some basic computations, such as calculation of the mean and standard deviation, of positional measures, or of the correlation coefficient for values of two attributes. Our experience with data-mining tools may be more interesting for readers, and we shall try to share it.

We shall not try to explain what happens inside the computational tools that we are going to talk about. We shall treat them as “black boxes”, i.e. consider only their inputs and outputs. Accordingly, we shall talk about preparing data for such tools and about interpreting and using the outcomes of these tools. In particular, we shall focus on the ways to represent the results of these tools so as to facilitate their interpretation and their use in further analysis with the application of other tools.

#### 4.7.1 A Few Words About Statistical Analysis

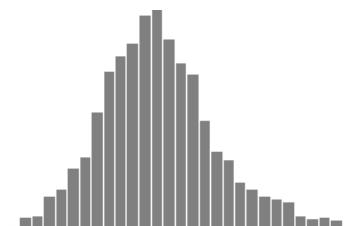
Statistics is the cradle of exploratory data analysis, and many people still view EDA as a branch of statistics. Statistics pays much attention to graphical representation of data, and most of the widely used graphical data displays originate from statistics. In one of the electronic handbooks

on statistics available on the Internet (Dallal 1999), a section entitled “Look At the Data!” precedes the introduction of the basic statistical notions and terminology. The author of that handbook explains how histograms, scatterplots, dot plots, box plots, parallel-coordinates plots, and line plots (which are called time graphs in the present book) are constructed, and what to look for in these data displays before trying to apply any computational methods.

Computational statistics is traditionally divided into *descriptive* and *inferential* statistics. The role of descriptive statistics is to summarise data, i.e. express the most important characteristics of the data in a few numbers. Thus, for a statistically educated person, the mean and the standard deviation of a set of numeric values tells him/her nearly everything that he/she needs to know about these data, *provided that the numbers are normally distributed*. A normal distribution means that the histogram representing the set of numbers has a symmetric shape resembling a bell, as is shown in Fig. 4.118, for example. So, if a set of numbers is normally distributed, the mean can be treated as the most typical value, and the standard deviation shows the degree of variability, or spread, around this most typical value. Moreover, a statistician knows that approximately 68% of all of the data lie within a distance of one standard deviation from the mean, that approximately 95% of the data lie within two standard deviations of the mean, and approximately 99.7% of the data lie within three standard deviations of the mean. Hence, just two numbers provide quite a lot of information and may be used in many situations instead of the whole set, irrespective of its size.

However, it may easily be guessed that data are not always distributed normally. When they are not, the use of the mean and the standard deviation can be misleading. Sometimes, a skewed distribution can be turned into a normal distribution by applying some non-linear transformation to the data. A logarithmic transformation is most often used. Another way to summarise arbitrary data with a non-normal distribution is to use positional measures, i.e. percentiles, in particular, the median and the quartiles. We have discussed the use of positional measures in the Sect. 4.5.4.

Generally, descriptive statistics summarises data by *measures of location* and *measures of variability*. The arithmetic mean and the median are examples of measures of location. One more example is the mode, i.e. the most frequently occurring value. The standard deviation is a measure of



**Fig. 4.118.** A histogram of a normally distributed set of numbers

variability. Some other possible measures of variability are the data range (the difference between the minimum and the maximum) and the coefficient of variation (the standard deviation expressed as a percentage of the mean). It is crucial to understand which statistical measure should be used as a summary index in each particular case.

Measures of location and variability are used to summarise a single attribute. There is also a need to describe the joint behaviour of two or more attributes. The numerical summary includes the mean and standard deviation of each attribute separately, plus a measure of the degree of their relatedness. The most widely used measure is the *correlation coefficient*, a summary of the strength of the linear association between the attributes. If the attributes tend to go up and down together, the correlation coefficient will be positive. If the attributes tend to go up and down in opposition, with low values of one attribute associated with high values of the other, the correlation coefficient will be negative. “Tends to” means that the association holds “on average”, not for any arbitrary pair of attribute values.

It should be noted that statistical practitioners warn against using any numerical measures without looking at appropriate graphical displays of the data: “There are too many ways to be fooled by numerical summaries!” (Dallal 1999). In particular, the correlation coefficient can give misleading results if not used in combination with scatterplots. Thus, this coefficient is greatly affected by outliers. A scatterplot can expose outliers in data, which can be removed before computing the correlation coefficient.

Another problem arises when attributes are related in a non-linear way. For example, data that are symmetrically placed on the curve  $Y = X^2$  will have a correlation coefficient of 0, although  $Y$  can be predicted perfectly from  $X$ . The reason why the correlation is zero is that high values of  $Y$  are associated with both high and low values of  $X$ .<sup>24</sup>

Unfortunately, there is no simple way to measure non-linear relations. If something like a monotonic curve (i.e. a curve that increases or decreases continuously) can be traced on a scatterplot, one can try to transform the values of one or both attributes to remove the curvilinearity and then recalculate the correlation. A typical transformation used in such cases is the logarithmic function. Another approach is to try to identify the specific function that best describes the curve perceived from the scatterplot. After a function has been found, an analyst can test its “goodness-of-fit” to the data.

Besides the generic statistical measures, there are measures specially designed for certain types of data, in particular, spatial and temporal data.

<sup>24</sup> Illustrations of other cases where the correlation coefficient fails can be found, for example, in Dallal (1999), in the section “Correlation Coefficients”.

Concerning spatial data, different measures are used for area, point, and directional data. For time-series data, the field of descriptive statistics includes methods for decomposing time series into several components: trend, cyclical, seasonal, and irregular or random component.

The methods of inferential statistics allow an analyst to generalise the result of a study of a few individuals to some larger group, or, in statistical terms, to generalise from a *sample* to a *population*. The term “population” in statistics means a collection of all possible observations of a specified characteristic of interest. A sample is a subset of a population.

There are two main approaches used in inferential statistics: *estimation* and *hypothesis testing*. In estimation, the information obtained from the sample is used to guess the value of a certain *parameter* of the entire population, where the term “parameter” means one of the statistical measures such as the mean, the standard deviation, the proportion (i.e. fraction of the observations that have a particular property in the entire set of observations), or the correlation. A parameter value is estimated together with a *confidence interval*, that is, a range of values that has a high probability of containing the actual parameter being estimated. For example, a 95% confidence interval contains the actual parameter value with a 95% probability. This percentage is called the level of confidence. There is a trade-off between the amount of confidence that one has in an interval and its length.

In hypothesis testing, one makes a reasonable assumption about the value of a population parameter and then uses the sample information in order to decide whether or not this hypothesis is supported by the data. Both estimation and hypothesis testing are based on statistical relationships between samples and populations; therefore, these two approaches are closely related.

It is hard to say more about statistical analysis without introducing complex definitions and formulae. Since we cannot give a systematic exposition of the whole of statistics in a section of this book, we prefer to stop at this point with our introductory notes. On the basis of the introduction that we have given, we shall now try to present our opinion concerning the use of computational statistical methods in exploratory data analysis.

Let us start with descriptive statistics. One can hardly argue against the assertion that data summarisation by a single number or a few numbers has quite a limited value. First, it cannot replace the consideration of graphical representations of the data. Moreover, such consideration must always precede the computation of any summary measure and must justify the use of this measure. Second, descriptive statistics does not give any additional understanding of the data, as compared with graphical data displays. In fact, it is not intended to give understanding. Its role may be described as

giving an analyst some way to express what he/she sees on a graphical display without using graphics or referring to graphics. This may be necessary, for example, for reporting purposes, including the reporting of findings obtained with the use of graphics.

For judging the role and value of inferential statistics, let us recall that the goal of exploratory data analysis is to gain understanding of the phenomenon behind the data. According to a metaphorical expression from one of the Web courses on EDA (NIST/SEMATECH 2005), the data are used as a “window” to peer into the heart of the process that generated the data. Continuing with this metaphor, we could say that the analyst needs some methods to validate the impressions received by looking through the window. Can these impressions be trusted and used as a basis for conclusions about the phenomenon?

This task is analogous to that of inferential statistics: from a limited sample, generalise to the entire population, i.e. the collection of all possible observations. Therefore, the apparatus of inferential statistics can be used for the validation of findings obtained by means of exploratory techniques, which are predominantly based on the viewing and manipulation of various data displays. The whole analysis process might be constructed in the following way: by means of exploratory analysis, an analyst generates some hypotheses about the phenomenon under study, and then applies inferential statistical methods in order to test these hypotheses.

Finally, we would also like to remind readers that a great many of the data transformation methods that have proved to be useful in exploratory data analysis are based on statistics. Examples considered in this book include various methods of data normalisation and standardisation, smoothing, interpolation, and aggregation. Moreover, statistical techniques are also involved in many data-mining methods.

## 4.7.2 A Few Words About Data Mining

StatSoft (2004) defines data mining as “an analytic process designed to explore data (usually large amounts of data – typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data”. According to the same source, the ultimate goal of data mining is prediction.

It is interesting that the phrase “an analytic process designed to explore data ... in search of consistent patterns and/or systematic relationships between variables” can also be used as a definition of exploratory data analysis. Actually, the authors of StatSoft (2004) see EDA as a part of data min-

ing. The process of data mining consists of three stages: (1) the initial exploration, (2) model-building or pattern identification with validation/verification, and (3) deployment, i.e. the application of the model to new data in order to generate predictions. Exploratory data analysis is used in the first stage, which also includes data preparation. EDA is done using a variety of techniques, both graphical (i.e. various methods of data visualisation and display manipulation) and computational, such as cluster analysis, multidimensional scaling, and building classification trees.

Another view is taken in Fayyad et al. (2002), where data mining is defined as a “mechanised process of identifying or discovering useful structure in data”. This definition emphasises automation of the analysis process as a distinctive feature of data mining, and hence separates data mining from graphics-based data exploration. Graphical displays, however, need to be used for visualisation of the results of data mining, so that a human explorer may perceive and understand them.

In our opinion, neither of these two views is wrong, and we are not going to adhere firmly to either of them. We allow readers to regard our book either as a book on data mining with an emphasis on graphical data-mining techniques, or as a book on exploratory data analysis with a suggestion to combine classical (that is, graphics-based) exploratory techniques with computational techniques from other disciplines, in particular, data mining.

In fact, our position with respect to computational and graphical techniques differs from both of these views. While the first view tends to consider these groups of techniques as alternatives and the second view does not explicitly acknowledge any relation between computational and graphical techniques, we think of computational methods as a complement to graphical methods. From our viewpoint, the purpose of applying computational data-mining techniques is to gain additional knowledge about data which cannot be (easily) gained directly from the viewing and manipulation of graphics. In particular, it may be recommended that one should apply computational techniques to very large datasets, both in terms of the size of the reference set and in terms of the number of attributes.

We do not mean, however, that computations produce ready-to-use knowledge, which needs only to be transferred somehow to the mind of a human analyst. We mean that a human analyst gains knowledge by means of exploration of the output of computational methods, similarly to the exploration of the original data.

Moreover, some researchers admit that gaining knowledge is not the primary role of data mining, and many data-mining techniques are not at all meant for helping an analyst to understand phenomena characterised by data. If we look once more at the definition given by Fayyad et al. (2002), we notice that it talks about “useful structure in data” rather than “knowl-

edge". This emphasis on *usefulness* is a distinctive feature of data mining. The focus is on deriving models that can generate useful predictions (in particular, in business applications) rather than uncover the nature of the underlying phenomena. It should not be thought that the former is impossible without the latter. Thus, data-mining techniques based on neural networks can give valid predictions, but these predictions are not based on any revealed interrelations between data components. A neural-network model has to be accepted as a "black box": even if a human analyst were to try to find out how it makes its predictions, this would not add any understanding of the phenomenon.

Not all data-mining methods, however, are based on neural networks, and not all models that they produce are incomprehensible to humans. As we already noted, many data-mining methods are based heavily on statistics. Data mining also incorporates approaches and techniques from other disciplines, such as information theory, graph theory, and inductive logic. In many cases, the output of data mining when properly interpreted, can really contribute to the understanding of phenomena. While this knowledge increment can be seen as a kind of "by-product" of data mining, whose primary concern is to make useful predictions, there are also researchers who believe that the primary goals of data mining include the description of data, which entails finding patterns within the data that are understandable to humans (Rhodes 2002).

In Chap. 3, which deals with data analysis tasks, we listed the tasks recognised in data mining. Since these tasks are defined in such a way that they give an idea of what types of results are produced by various data-mining techniques, it is useful to list them once again here (Table 4.13).

Practical needs are driving the development of special data-mining methods that are intended to be capable of analysing spatial data, taking into account positions, distances, neighbourhood, and other spatial relations. A basic problem is that computers can only process numbers and letters arranged in chains, tables, or logical expressions. Hence, in order to be automatically "mineable", spatial information needs to be represented by properly arranged numbers or letters. The general approach taken in "spatial data-mining methods" is to transform spatial information into numeric or symbolic form and then to apply standard processing techniques.

Suppose, for example, that we have a dataset containing various census data collected for small enumeration districts, including data concerning the health of the population in these districts. The dataset also contains data about the motorways that cross the territory. On the basis of these data, distances from each district to the nearest motorway may be computed (if the districts are large, and the road network is dense, it may be more reasonable to compute, for example, the average road density per

**Table 4.13.** Data-mining tasks and techniques (from Fayyad et al. (1996) and Miller and Han (2001))

Task	Techniques
<ul style="list-style-type: none"> <li>• <i>Segmentation</i>: Partitioning data into meaningful groupings or classes. This includes two major subtasks:           <ul style="list-style-type: none"> <li>– <i>Clustering</i>: Determining a finite set of implicit classes that describes the data.</li> <li>– <i>Classification</i>: Finding rules to assign data items to pre-existing classes.</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Cluster analysis</li> <li>• Bayesian classification</li> <li>• Decision or classification trees</li> <li>• Artificial neural networks</li> </ul>
<ul style="list-style-type: none"> <li>• <i>Dependency analysis</i>: Finding rules to predict the value of an attribute on the basis of the values of other attributes.</li> </ul>	<ul style="list-style-type: none"> <li>• Bayesian networks</li> <li>• Association rules</li> </ul>
<ul style="list-style-type: none"> <li>• <i>Deviation and outlier analysis</i>: Searching for data items that exhibit unexpected deviations or differences from some norm.</li> </ul>	<ul style="list-style-type: none"> <li>• Clustering</li> <li>• Outlier detection</li> </ul>
<ul style="list-style-type: none"> <li>• <i>Trend detection</i>: Fitting lines and curves to data in order to summarise the database.</li> </ul>	<ul style="list-style-type: none"> <li>• Regression</li> <li>• Sequential pattern extraction</li> </ul>
<ul style="list-style-type: none"> <li>• <i>Generalisation and characterisation</i>: Compact description of the database, e.g. as a relatively small set of logical statements that condense the information in the database.</li> </ul>	<ul style="list-style-type: none"> <li>• Summary rules</li> <li>• Attribute-oriented induction</li> </ul>

district). The new attribute derived in this way may be processed by a computational data analysis tool together with the original census attributes. A possible finding resulting from the application of the tool might be that districts situated close to motorways (or having a denser road network) tend to have higher percentages of ill people in their population.

It should be borne in mind that spatial information is complex and multifaceted, whereas any transformation procedure can encode only a limited part of it, only a specific aspect, in a machine-processable form. Thus, in our simple example, the distances to motorways were suitably represented and therefore could be taken into account in the analysis. However, the computational tool could not tell us anything concerning the pattern of the spatial distribution of illnesses over the territory; in particular, whether districts with a high percentage of ill people form spatial clusters or are scattered over the territory. In order to answer this sort of question, computational tools need the information about the neighbourhood relations between the districts to be appropriately encoded, for example in the form of

a neighbourhood matrix. There are data-mining methods that have been specially devised to work with such matrices and can, for example, detect spatial clusters of districts with similar characteristics.

However, neighbourhood matrices again represent only a limited part of the potentially relevant spatial information. A method that uses such a matrix cannot, for example, tell us that the proportion of ill people tends to increase from north to south or discover any other fact related to spatial directions. To make this possible, one needs to find a suitable representation for directional information, for example by predicates such as SOUTH-OF ( $A, B$ ), where  $A$  and  $B$  are the identifiers of two districts. Then, one can apply a method designed to deal with predicates, and this method will probably detect some regularities related to directions, but will not take account of district sizes, relief, closeness to particular types of industrial enterprise, and so on.

The message we want to convey is that there is no computational tool for spatial analysis that can take account of all potentially relevant aspects of spatial information or detect automatically what aspects are the most important in any particular case, in order to properly encode and analyse those aspects. Human eyes, when supplied with a map representing spatial data, are definitely superior in this respect to any computational tool because they immediately *see* all aspects: neighbourhood, relative distances, sizes, directions, spatial grouping, heterogeneity of the geographical space, various topological relations, symmetry, and so on. Only a human analyst can judge what aspects may be important. Therefore, we do not think that it will ever be possible to devise a computational tool that, once it has been fed with spatial data, will tell us all we need to know about the data.

A realistic scenario for the application of computational tools, in particular, data-mining tools, in the analysis of spatial data is that the analyst first explores the data visually using cartographic representation(s) of spatial information, i.e. map displays. Maps are very important in the exploration of spatial data because they are structurally similar (isomorphic) to two-dimensional space, in particular, geographical space. Therefore, a reasonably faithful representation of locations and outlines is sufficient for a map to properly convey to human eyes all spatial properties and relations that exist in reality. On this basis, an explorer can uncover the significant aspects that require a close look and think about the directions and methods of further investigation, including computational methods. Only then can computational tools be applied effectively. However, to interpret and make use of their results, the explorer will again need visualisation, especially map displays.

This scenario can be recommended for the analysis of not only spatial but also any other type of data. Of course, when the data are non-spatial,

other representations should be used instead of maps, but the general principle remains the same: visualise first, and then, on this basis, choose and apply computational tools. We would like to emphasise this principle and, for this purpose, we have included a special subsection below.

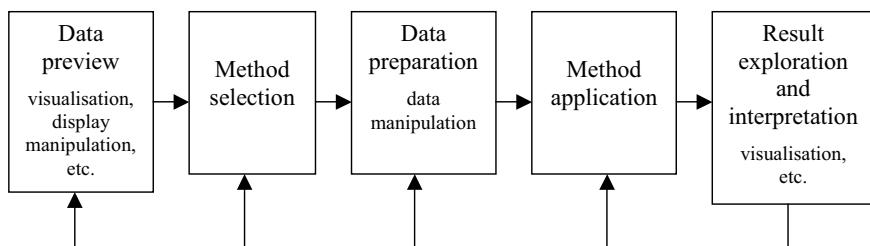
### 4.7.3 The General Paradigm for Using Computational Tools

In our work, we have never used data-mining or statistical computations alone but have always used them in combination with visual and interactive exploratory techniques, i.e. data visualisation, display and data manipulation, and querying. We have developed and applied a paradigm for using computational data analysis techniques. According to this paradigm, data analysis with the use of computational tools consists of the following steps:

1. Look at the data in order to understand what computational tools could be useful to apply. This requires data visualisation with display manipulation.
2. Choose a computational method to apply.
3. Bring the data to a form suitable for applying the method, e.g. transform absolute data into relative data. In this step, data manipulation tools need to be involved.
4. Apply the method and store its outcomes for future use.
5. Explore the outcomes of the method in order to properly interpret them; possibly, compare them with results produced by other methods or by the same method with different parameter settings. This requires using a range of exploratory techniques, at least visualisation and display manipulation tools.
6. Return to one of the earlier steps; try the same method with different settings (step 4), a different data transformation (step 3), or a different method (step 2), possibly after taking a new look at the data (step 1).

This analysis paradigm is represented schematically in Fig. 4.119. We are far from claiming that this paradigm is our own invention, and that no one has used it before us. On the contrary, we are sure that something like this is always used in the practice of data analysis, since this seems to be the most logical way of using computational methods in data analysis. Thus, no one would start any computation without previously looking at the data using appropriate graphical representations, although not every handbook on statistics or data mining emphasises the necessity for data visualisation prior to any computation. Similarly, no one would feel fully satisfied immediately after receiving the results of a computation without

trying to interpret and/or verify those results using appropriate visualisations and other techniques. Moreover, any serious analyst would not be happy with having the results of just a single run of a single method. He/she would at least try to investigate the sensitivity of the results to changes in the parameters of the method, and it is not unusual for several different methods to be applied to the same data in order to gain better understanding. Hence, our scheme should be regarded as a description of how data analysis is actually done, rather than a prescription of how it should be done.



**Fig. 4.119.** The paradigm for applying computational tools in exploratory data analysis

Let us now consider a few examples of the use of computational data analysis tools. For these examples, we have utilised some of the techniques available in an open-source software system for data mining called Weka (see Witten and Frank (1999); the system is available on the Web at <http://www.cs.waikato.ac.nz/ml/weka/>).

#### 4.7.4 Example: Clustering

Section 4.6 contains an example of data analysis where districts of Portugal were grouped according to their similarity to certain selected districts in terms of the age structure of the population, which was represented by four attributes: “% 0–14 years”, “% 15–24 years”, “% 25–64 years”, and “% 65 or more years”. For this purpose, we used a query tool that computed distances between characteristics corresponding to pairs of references.

The data-mining methods for *clustering* are intended to do something similar: they group references by closeness of their characteristics. However, they do not require the previous selection of any “models” so that references could be grouped according to their closeness to those models. The idea is to find “natural” groupings, with small distances between members of a group and large distances between groups.

There are various approaches to doing this and, accordingly, there is a range of clustering methods. We shall not describe the differences between the existing methods, but would like to point out that

- different clustering methods applied to the same data may produce very different groupings (moreover, the results of the same method will differ depending on the parameters chosen); and
- none of the existing clustering methods can explain its results, i.e. why the references have been grouped in this or that way.

Therefore, an analyst needs to visualise clustering results in such a way that he/she can see the distribution of the characteristics across the groups and understand in what way the members of each group are similar and how they differ from the members of the other groups. It may happen that the groups do not seem to have consistent characteristics. In this case, it is necessary to perform another trial, where the parameters of the same clustering method are changed or another method is chosen. This can also be recommended even when the results are satisfactory: why not try to improve them further?

Let us try to apply clustering to our data characterising the age structure in the districts of Portugal. Figure 4.120C demonstrates the result of applying one of the clustering methods available in Weka, which is called “simple  $k$ -means”. We shall not describe how this method works or explain its name; our intention is only to demonstrate its use. From this perspective, it is important to mention that the method requires the user to specify the desired number of groups, or clusters, that must be produced. We have specified that the districts of Portugal must be divided into three groups.

To represent the resulting division, we have used a map display where the districts are shown in three different colours, depending on the cluster in which they have been included. This map can be seen on the left in Fig. 4.120C. It demonstrates a quite distinctive spatial clustering of districts belonging to the same group. However, what do the groups mean? The clustering tool does not provide any description of the clusters produced. We only know that the members of each group must have relatively close characteristics in terms of the age structure of the population. But what are these characteristics? This needs to be investigated.

Since the clustering tool does not give any meaningful names to the clusters but only numbers, we shall refer to them as the “yellow cluster”, “blue cluster”, and “red cluster”. The colour assignment is arbitrary.

In order to see the characteristics of the clusters in terms of the four age structure attributes, we have constructed a frequency histogram display for each attribute. We have then propagated the cluster colours from the map to the histograms as multicolour marking. In the result, the bars of the his-

tograms have been divided into segments according to how many districts from each cluster fit into each bar. The segments are shown in the colours of the clusters. The segmented histograms, which can be seen on the right in Fig. 4.120C, can help us to interpret the meaning of each cluster.

From the positions of the yellow segments in the histograms, it may be seen that the yellow cluster consists of districts with a high proportion of children (i.e. people aged from 0 to 14 years) and young people (aged from 15 to 24 years), and low a proportion of working-age people (aged from 25 to 64 years) and elderly people (i.e. aged 65 years and more). Hence, we can interpret the yellow cluster as a group of “young” districts. From the map, we see that this group is located in the north of the country.

The red cluster seems to have characteristics quite opposite to the yellow one. From the histograms, it may be seen that this cluster is characterised by a low proportion of children and young people and a high proportion of elderly people. The proportion of working-age people is mostly medium. Geographically, this cluster occupies rather a vast territory, mainly inland, in the east and south-east of Portugal.

The blue cluster consists of districts with mostly a medium proportion of children and young people, a quite high proportion of working age people and a relatively low proportion of elderly people. Geographically, this cluster stretches mainly along the western coast in the central part of the country. However, there are several blue spots in other parts of Portugal.

Although we can regard this grouping of districts as quite good (since the characteristics of the groups are rather consistent and understandable), it is useful to check whether a division into more than three clusters can give an even clearer picture. So, we asked the clustering tool to divide the set of districts into four groups. The results are presented in Fig. 4.121C.

We need to explain here that each run of the clustering tool is independent of other runs. Therefore, when a cluster obtained from the second run has the same number as a cluster obtained from the first run, this does not mean that these clusters are related to each other. In order to make the result of the second run more comparable to the results of the first run, we have assigned colours to the clusters obtained after the second run so as to make the resulting map look as similar as possible to the map shown in Fig. 4.120C.

In Fig. 4.121C, we again have yellow, red, and blue clusters, plus an additional green cluster. In both their geographical positions and their characteristics in terms of the age structure of the population, the yellow and blue clusters in Fig. 4.121C are quite similar to the yellow and blue clusters in Fig. 4.120C. The green cluster seems to result from the former red cluster being split so that “extremely old” districts have been separated from “moderately old” ones. As can be seen from the histogram of the attribute

“% 65 or more years” in Fig. 4.121C, the green cluster contains the districts with the highest proportion of elderly people (specifically, from 23.8% to 35.2%, which is the maximum for Portugal). The red cluster consists of districts with a medium proportion of elderly people.

The same histogram shows us clearly that the blue, red, and green clusters are differentiated mostly on the basis of the proportion of people aged 65 or more years: low proportions in the blue cluster, medium in the red cluster, and high in the green cluster. The yellow cluster is characterised by low to medium proportions of elderly people and is close in this respect to the blue and, partly, the red cluster. The yellow cluster differs from the blue and red ones in terms of the proportions of the other age groups, especially children (high proportions in the yellow cluster and low or medium proportions in the other two). There is also a quite clear distinction between the yellow and blue clusters in terms of the proportion of people aged from 25 to 64 years: quite low proportions in the yellow cluster and mostly high proportions in the blue cluster.

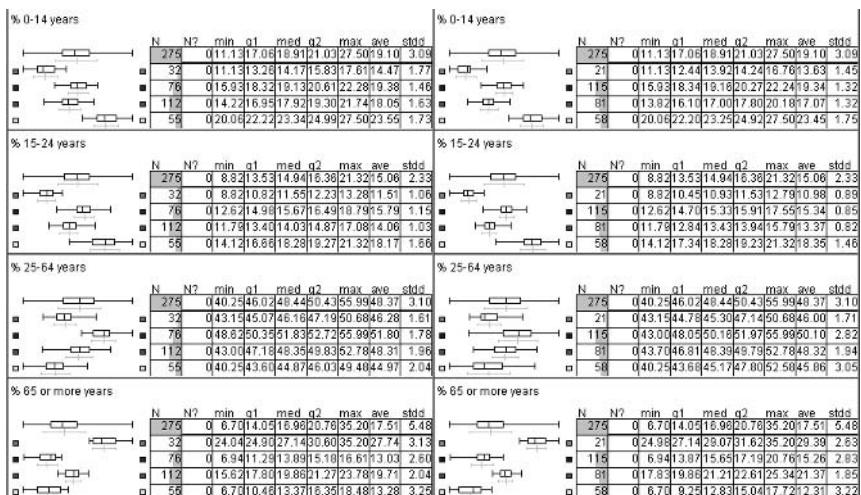
It can be seen that the blue cluster has become significantly smaller than before. More precisely, it has decreased from 108 to 76 districts; the remaining 32 districts have moved to the red cluster. The yellow cluster has lost five of the initial 60 districts; these five districts have also moved to the red cluster. However, the red cluster has increased by only five districts (from 107 to 112), since 32 districts have been grouped into the green cluster. As a result, the yellow and blue clusters have become more coherent, in terms of both their geographical distribution and their age structure. The members of the green cluster also have rather consistent age structure characteristics (i.e. many elderly people and few children and young people) but are more scattered geographically. The red cluster may be regarded as average in all respects.

In general, we cannot say definitely that the division into four groups is much clearer than the division into three groups, in terms of the consistency of characteristics of the districts within the groups. Both divisions are quite interpretable. The representation of the clusters on the map allows us to get a clear idea about the distribution of the age structure of the population over the territory of Portugal. For us, this is the main result of applying the clustering tool, not the clusters by themselves. Therefore, it is not so important how many clusters the districts are finally divided into.

In order to demonstrate that different clustering methods may produce different results, we have applied another clustering method, called “expectation maximisation”, or EM, to the same data. As with the previous method, we have let the method divide the districts first into three and then into four clusters. The results are shown in Figs 4.122C and 4.123C. Again, for a more convenient comparison with the previous groupings, we

have assigned colours to the clusters in such a way that the appearance of the map is maximally preserved, i.e. the yellow cluster is positioned on the north, the blue cluster in the west, and the green cluster in the east.

It may be observed that the two methods do not group the districts in exactly the same manner, although the results are quite similar, especially in the case of four clusters. Interested readers can look at Fig. 4.124 to compare the summary statistics of the divisions into four clusters produced by the two methods. The greatest differences between the results of the two methods can be observed in the parts corresponding to the attribute "% 25–64 years". The groups produced by the EM method have much higher variability with respect to this attribute than the groups resulting from the method of simple  $k$ -means. However, the results of EM have notably lower variability with respect to the three other age structure attributes.



**Fig. 4.124.** Summary statistics of the clusters resulting from the application of two clustering methods: simple  $k$ -means (left) and expectation maximisation (right). The box-and-whiskers plots and the table rows in each section correspond, from top to bottom, to the entire set of districts, to the green cluster, to the blue cluster, to the red cluster, and to the yellow cluster. The table columns show “N”, the number of districts in each set; “N?”, the number of districts with missing values; “min”, the minimum attribute value among the members of the set; “q1”, the lower quartile (25% percentile), “med” – the median, “q2” – the upper quartile (75% percentile); “max”, the maximum attribute value; “ave”, the average (mean); “std”, the standard deviation

We would like to emphasise once again that the goal of our example investigation was not to produce “nice” groups but to understand the distribution (or, in other words, behaviour) of the age structure over the territory

of Portugal. Such a goal is more typical of exploratory data analysis (whereas good grouping might be rather important in a map intended for presentation purposes). From this perspective, both methods produce quite good results. Nevertheless, it might be interesting to look at the characteristics of the districts that the methods have included in distinct clusters.

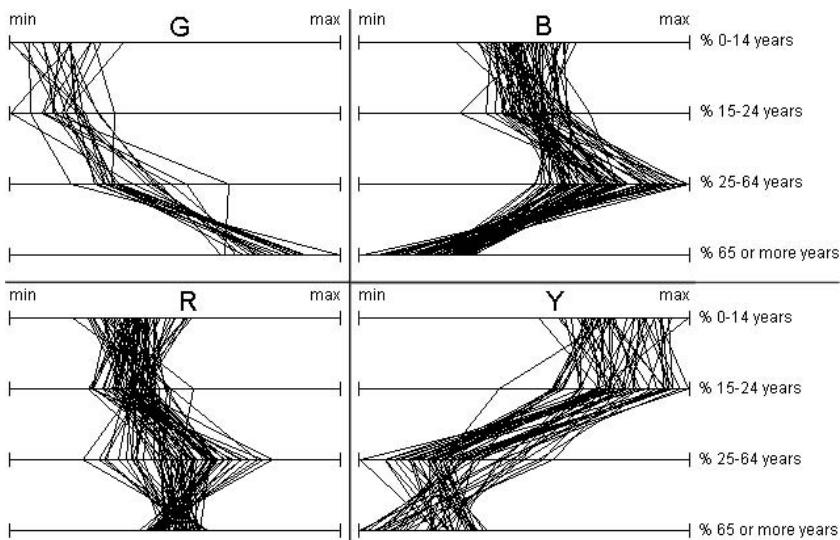
Visualisation, display manipulation, and querying tools can help us to do such an investigation. In Fig. 4.125C, we see a map (on the left) representing the outcomes from both clustering methods together. Green, blue, red, and yellow colours are used to show the districts included in the same clusters by both methods. However, there are four groups of districts in the result of EM that have “moved” to other clusters, in comparison with the results of simple  $k$ -means:

- 11 districts have moved from the green to the red cluster; these are shown in brown.
- 9 districts have moved from the blue to the yellow cluster; these are shown in cyan.
- 42 districts have moved from the red to the blue cluster; these are shown in magenta.
- 6 districts have moved from the yellow to the blue cluster; these are shown in pale green.

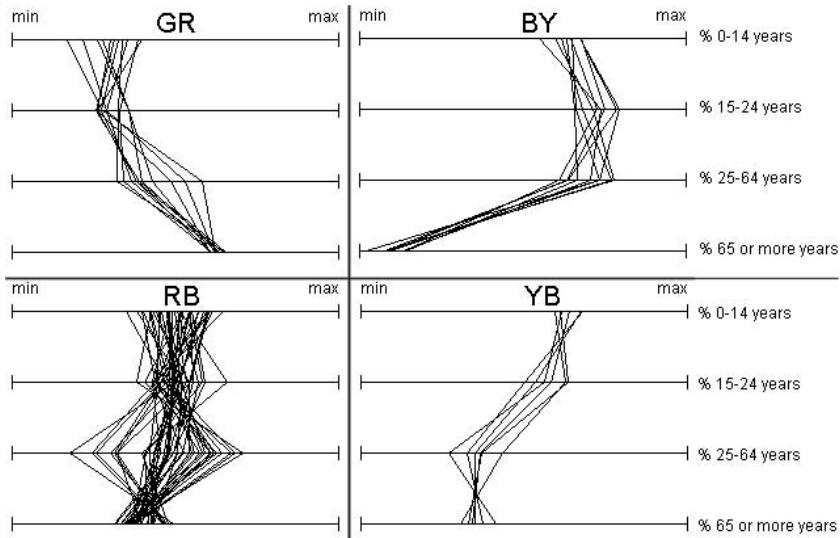
On the right in Fig. 4.125C, the districts that have not changed their cluster membership are shown in grey, so that it is easier to focus on the differences between the two groupings. Between the maps, the absolute and relative sizes of the eight groups of districts are shown as numbers and by a bar chart.

While the map allows us to see the geographical differences between the two groupings of districts, the differences in terms of the age structure can be explored using other displays, for example a parallel-coordinates plot with axes corresponding to the age structure attributes. Using a filtering tool, we can focus on the characteristics of any group of districts. It is interesting to look at the age structures in the districts for which the outcomes of the two methods differ and compare these with the characteristics of the subgroups for which the results of the methods overlap.

Figure 4.126 represents the characteristics of the common parts of the two groupings of districts. In four screenshots from a parallel-coordinates display, we can see the age structure profiles of the districts belonging to the green (upper left), blue (upper right), red (lower left), and yellow (lower right) clusters according to both groupings. The images are labelled G, B, R, and Y, respectively. Analogously, Fig. 4.127 shows the profiles of the subgroups of districts for which the results of the two methods differ.



**Fig. 4.126.** Four screenshots from a parallel-coordinates display representing the age structure profiles in the four subgroups of districts that have been put in the same cluster by both clustering methods. Upper left, green cluster; upper right, blue cluster; lower left, red cluster; lower right, yellow cluster



**Fig. 4.127.** Four screenshots from a parallel-coordinates display representing the age structures in the four subgroups of districts for which the results of the clustering methods differ. Upper left, green-to-red districts; upper right, blue-to-yellow; lower left, red-to-blue; lower right, yellow-to-blue

The upper left image represents the green-to-red group of districts (these districts are coloured brown in the map in Fig. 4.125C), the upper right image represents the blue-to-yellow (cyan) group, the lower left image represents the red-to-blue (magenta) group, and the lower right image represents the yellow-to-blue (pale green) group. These images are labelled GR, BY, RB, and YB, respectively.

It may be seen that the age structure profiles in the coinciding subgroups (Fig. 4.126) are rather clear:

- *green* (G): Low proportions of children and young people, medium proportion of working-age people, and high proportion of elderly people;
- *blue* (B): Medium proportions of children and young people, high proportion of working-age people, and low proportion of elderly people;
- *red* (R): Medium proportions of all age groups, with a slight shift to lower values for the proportions of children and young people and rather high variability for the proportion of working-age people;
- *yellow* (Y): High proportions of children and young people and low proportions of the two other age groups.

The profiles visible in Fig. 4.127 seem to be “boundary cases”. Thus, the lines in the image labelled GR have a certain similarity to the profiles of the green subgroup but also to those of the red subgroup. The corresponding districts have somewhat higher proportions of children and young people and lower proportions of elderly people than in the green subgroup, but the proportions of elderly are slightly higher than in the red subgroup.

Analogously, we can see that the profiles in the BY section look as if they are intermediate between those of the blue and yellow subgroups. The profiles in the RB section could be attached to the red subgroup but have higher proportions of children and young people and therefore are also similar to the profiles of the blue subgroup. The lines in the image labelled YB appear to be similar to the profiles of the yellow and blue subgroups but in a different way than for the lines in the BY section: the YB lines indicate much higher proportions of elderly people than do the BY lines but notably lower proportions of the age group 25–64 years.

Now we can feel reasonably comfortable with the results of the two clustering methods: we have understood quite well how and why they differ and, at the same time, have learned quite a bit about the age structure profiles in the districts of Portugal and their geographical distribution. We know, for example, that “young” districts form a cluster in the north, and that the area around Lisbon and the Atlantic coast north of it are characterised by a high proportion of working-age people and a low proportion of

elderly people. Similar structures occur in a group of districts situated on the southern coast. The inner parts of the country generally have an older population, and some districts in the centre and in the east may be characterised as “extremely old”, with very low proportions of children and young people and also quite a low proportion of working age people.

Let us now leave this analysis and move on to another example.

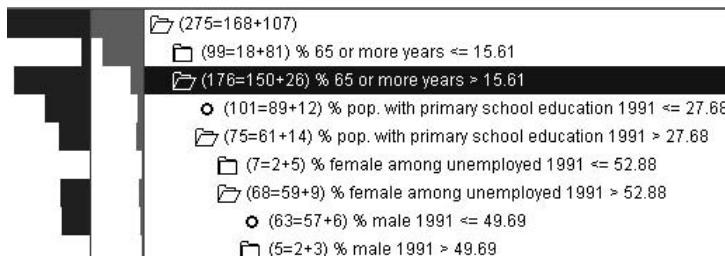
#### 4.7.5 Example: Classification

This time, we shall try to use another data-mining method, which constructs classification trees. The task of classification is defined in data mining as finding rules to assign data items to pre-existing classes. For example, we can group the districts of Portugal in a certain way and let a classification method characterise these groups in terms of the available attributes. We have at our disposal a classification tool called J48, which is available in the data-mining toolkit Weka; so, let us look at how it can be used in analysing the Portuguese dataset.

If we look at the values of the attribute “% pop. change from 1981 to 1991”, which characterises the change in the population of each district in 1991 in relation to 1981, we can notice that there are very many districts where the population decreased. In fact, the number of such districts exceeds the number of districts with a population increase: there are 168 districts where the population decreased by more than 1% and only 88 districts where the population increased by more than 1%. The population in the remaining 19 districts is quite stable: it changed by less than 1%.

We would like to know whether the districts with a population decrease have consistent characteristics in terms of any of the available demographic attributes. So, we divided the districts into two classes, a class of districts with a decreased population and a class of districts with a stable or increased population. We sent these classes of districts to the J48 classification tool, together with the values of all attributes characterising the districts except for the attribute “% pop. change from 1981 to 1991”, which had been used for defining the classes. Then we ran the classification tool and received a result in the form of a classification tree, which can be visualised as is shown in Fig. 4.128, for example.

A classification tree is a hierarchy of nodes, in which the uppermost node (which is called the root node or simply the root) corresponds to the entire set of classified references (in our case, all districts of Portugal), and the other nodes correspond to subsets of that set, which may contain members of one or more classes. The subsets result from successive divisions of the reference set. Each division is made on the basis of one attribute.



**Fig. 4.128.** A classification tree distinguishes the districts with a population decrease from those with a stable population or a population increase

In the display in Fig. 4.128, each line describes one node of the classification tree constructed by the J48 tool for the classes of districts that we have defined. The topmost line corresponds to the root node of the tree. This line shows that the whole set, which consists of 275 districts, includes 168 districts classified as “decrease” and 107 districts classified as “stable or increase”. The bars on the left represent visually the relative sizes of the classes. The darker bar corresponds to the class “decrease” and the lighter bar to “stable or increase”.

The two second-level nodes described in the second and third lines correspond to the division of the whole set of districts into two subsets according to the values of the attribute “% 65 or more years”. One of the subsets, which is represented in the second line, consists of the districts where the values of this attribute are less than or equal to 15.61. The formula  $99 = 18 + 81$  in the description of this tree node means that the subset contains 99 districts in total, of which 18 belong to the “decrease” class and 81 to the “stable or increase” class. Again, the bars on the left represent the relative sizes of these two parts of the subset.

The third line describes the subset of districts where the values of the attribute “% 65 or more years” are more than 15.61. There are 176 such districts, of which 150 belong to the “decrease” class and only 26 to the “stable or increase” class.

Each of these two subsets is subdivided on the basis of other attributes. However, the node corresponding to the first subset is shown in Fig. 4.128 in a “folded” form, so that the lower-level nodes descending from it are hidden. The node corresponding to the second subset has been opened, and we can see the further division of this subset. We see that it is subdivided into two smaller subsets according to the values of the attribute “% pop. with primary school education 1991”, with the break point at 27.68. The subset of districts where the values of this attribute are less than or equal to 27.68 consists of 101 districts, of which 89 belong to the class “decrease” and 12 to the class “stable or increase”. This subset is not subdivided fur-

ther; the corresponding node is final. The other subset consists of 75 districts where more than 27.68% people have primary school education. Of these 75 districts, 61 have a population decrease and 14 have a stable or increased population. This subset is subdivided further into two subsets according to the values of the attribute “% female among unemployed 1991”, with the break point at 52.88, and so on.

In such a tree, every node except the root has a corresponding description in terms of values of attributes. The description of a second-level node consists of a single expression concerning values of one attribute, for example “% 65 or more years > 15.61”. The description of a third-level node consists of two expressions linked by the conjunction (logical operation) “and”, for example “% 65 or more years > 15.61 and % pop. with primary school education 1991 <= 27.68”. The description of a fourth-level node consists of three expressions linked by the conjunction “and”, and so on.

The recursive divisions of the reference set are aimed ultimately at obtaining homogeneous subsets, that is, where each subset consists of members of a single class. Therefore, when a subset resulting from a division contains a mixture of members of different classes, it needs to be further subdivided. For this purpose, an appropriate attribute and an appropriate split of its value set need to be found, so that the subsets are as close to homogeneity as possible. The challenge is to arrive at homogeneous subsets by means of the least possible number of recursive divisions.

If the homogeneity criterion is achieved, the descriptions of the final tree nodes may be used in two different ways. First, they can help an analyst to understand the differences between the classes. Thus, we hope to understand the differences between the districts with a decreased population and those with a stable or increased population. Perhaps we might even be able to identify the possible reasons that make people move to other districts. Analogously, a market analyst may hope to understand the distinctive characteristics of successful stores as compared with unsuccessful ones in order to identify the factors contributing to success or failure. Second, the descriptions can be treated as a set of rules that determine the class membership of any reference according to its characteristics. This allows one to use this set of rules to assign previously unclassified references to appropriate classes. For example, if a market analyst has obtained a set of rules that discriminate between good and bad sites where existing retail stores are situated, he/she may use these rules in order to evaluate the suitability of other sites for opening new stores. Such practical utility is probably the primary aspiration of data mining.

However, classification algorithms often fail to achieve a perfect differentiation between the classes and therefore produce classification trees with “impure” final nodes. Thus, the fourth line in Fig. 4.128 represents a

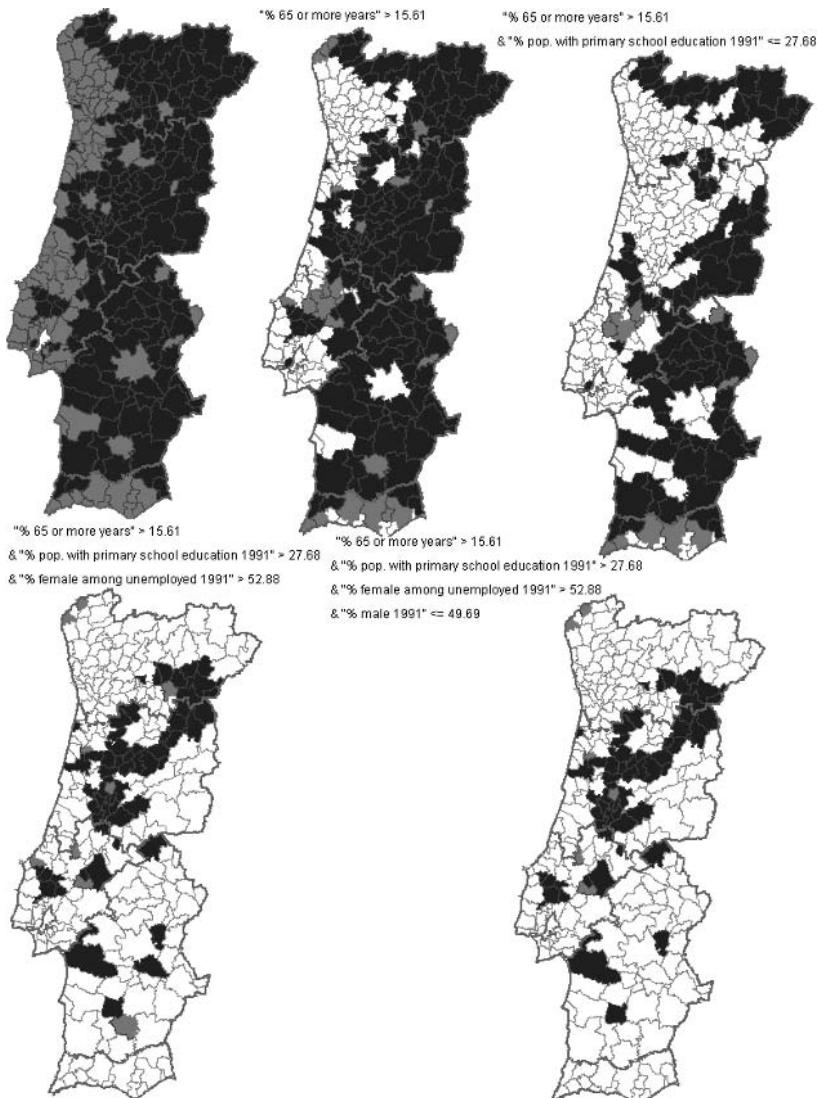
final node that corresponds to a “mixed” subset of districts: 89 districts of the subset belong to the “decrease” class and the remaining 12 to the “stable or increase” class. Evidently, the classification tool could not find suitable attributes to separate effectively these two subgroups of districts on the basis of their characteristics. Analogously, the final node represented by the second line from bottom in Fig. 4.128 stands for a subset containing 57 members of the “decrease” class and six members of the “stable or increase” class. The tool could not find a way to separate the latter six districts from the former 57.

It is unlikely that a tree with such “impurities” could satisfy a market analyst who needed a good instrument for prediction. Our goals are different and, despite the tree being far from perfect, we can gain some useful information from it. Thus, we see clearly that most districts with a decreased population (150 of 168, or 89%) have a medium to high proportion of elderly people, more exactly, over 15.61%, while the whole range of the proportion of elderly over Portugal is from 6.7% to 35.2%. Furthermore, a large fraction of these 150 districts (89, or 59%) is characterised by a not very high percentage of people who have had primary school education – not more than 27.68% (the values of the attribute range from 18.69% to 35.14%). The remaining part, i.e. 61 districts, or 41% of 150, is characterised mostly by a rather high (more than 52.88%) percentage of females among the unemployed and quite a low proportion of males in the total population (up to 49.69%).

Besides looking at the classification tree, we can also explore the results obtained by means of other analysis tools. In particular, we can use visualisation and filtering tools in order to see which districts are contained in the subsets described in the tree nodes and where these districts are geographically located. Thus, in Fig. 4.129, there are five screenshots from a map display. The map at the upper left represents the whole set of districts, i.e. it corresponds to the root node of the tree. A dark shade is used for the districts with a population decrease, and a lighter shade of grey for the districts with a stable or increased population. The next screenshot, i.e. the one in the centre, corresponds to the tree node represented in the third line of the display in Fig. 4.128. Accordingly, this screenshot shows the subset of districts with more than 15.61% of elderly people in their population. The districts that do not correspond to this description are not shaded (i.e. they are white). The next map (upper right) corresponds to the next node (i.e. the fourth line) and shows the subset of districts described as “% 65 or more years > 15.61 and % pop. with primary school education 1991 <= 27.68”.

The map at the bottom left represents the subset described in the third line from bottom in Fig. 4.128: “% 65 or more years > 15.61 and % pop.

with primary school education 1991 > 27.68 and % female among unemployed 1991 > 52.88". The map at the bottom right corresponds to the next node (the second line from bottom in Fig. 4.128). This map shows the subset described by the conjunction of four expression: "% 65 or more years > 15.61 and % pop. with primary school education 1991 > 27.68 and % female among unemployed 1991 > 52.88 and % male 1991 <= 49.69".



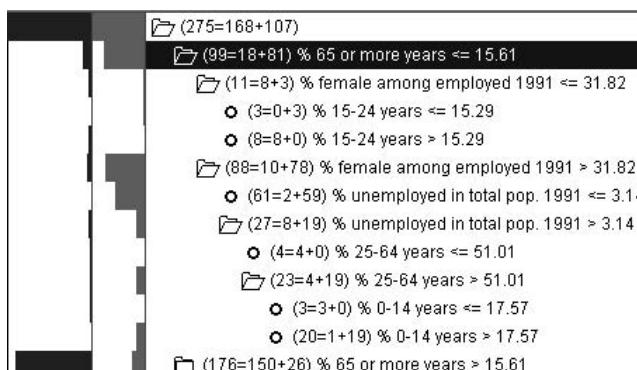
**Fig. 4.129.** The subsets of districts described by different tree nodes are shown here in a map display

It would be convenient for an analyst if he/she could easily obtain such maps, for example by clicking on the nodes of the tree. With a little more effort, it is also possible to obtain such views using a query tool.

It is convenient to observe, by means of visualisation of the subset, the effect of adding a new condition to the description of a subset. Thus, as may be seen from the maps at the top centre and right in Fig. 4.129, adding the condition "% pop. with primary school education 1991 <= 27.68" to the description "% 65 or more years > 15.61" was not very effective: it has removed a great many districts with a decreased population (black) while preserving quite a large number of districts with a stable or increased population (grey). At the bottom of Fig. 4.129, we see that adding the condition "% male 1991" to the description "% 65 or more years > 15.61 and % pop. with primary school education 1991 > 27.68 and % female among unemployed 1991 > 52.88" did not result in any notable changes.

In Fig. 4.130, the classification tree display has been transformed so that the branch corresponding to the subset of districts with up to 15.61% of elderly people in their population is now exposed for viewing. Only 18 districts with a decreased population occur in this branch. To separate them from the districts with a stable or increased population, the classification tool has used the attributes "% female among employed 1991", "% 15–24 years", "% unemployed in total pop. 1991", "% 25–64 years", and "% 0–14 years". Most of the final nodes describe very small subsets of districts, but one of them describes 61 districts, 59 of which have a stable or increased population. We can learn that these districts are characterised by quite high female employment (over 31.82%) and quite low unemployment (up to 3.14% of unemployed in the total population).

Generally, the classification tree suggests that a population decrease may be related to ageing of the population and unemployment (which might be expected) but that it also has something to do with the educa-



**Fig. 4.130.** Another branch of the classification tree is exposed for viewing

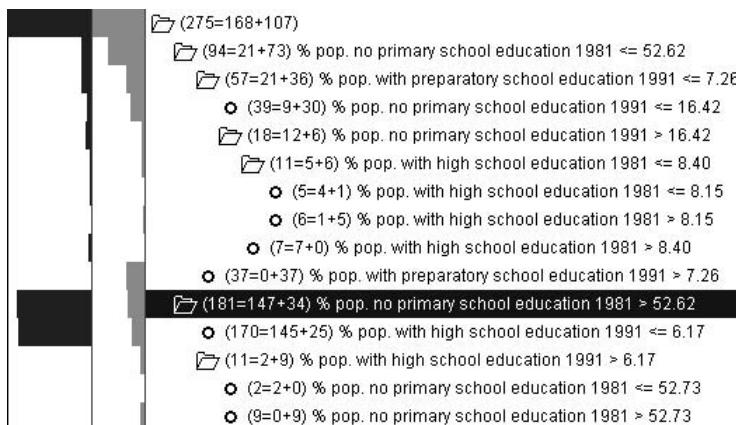
tional level and with the ratio between males and females in the entire population and among working and unemployed people. However, one should always pay attention to the sizes of the subsets represented by tree nodes. Thus, the fifth line from the top in Fig. 4.130 describes only eight of 168 districts with decreased population. These 8 districts are characterised by a low percentage of females among employed people (up to 31.82%, which is less than the country median of 35.2%) and a quite high proportion of young people (more than 15.29%, whereas the range for the whole of Portugal is from 8.82% to 21.32% and the median is 14.94%). It is important to understand that this description applies only to these eight districts, and one should not conclude from it that *all* districts with a population decrease have a low proportion of working women and a high proportion of young people.

What is reasonable to do is to investigate, by means of other exploratory tools, which of the attributes appearing in a classification tree are really related to the specified division of the reference set. For example, we have obtained the statistics of the values of the attributes involved in the classification tree for the entire set of districts of Portugal, the subset of districts with a population decrease, and the subset of districts with a stable or increased population. We have noted that there is indeed some relatedness between a population decrease or increase and the age structure of the population of the district. The districts with a decreased population tend to have fewer children, young people, and working-age people and more elderly people than do the districts with a stable or increased population. The difference in the trends in the proportion of elderly people is especially salient: the lower quartile (i.e. 25% percentile) of the subset of districts with a decreased population is higher than the upper quartile (i.e. 75% percentile) of the subset with a stable or increased population. The differences with respect to the percentage of females among the employed population are also quite notable: in the districts with a population decrease, the values of this attribute tend to be lower than in the districts with a stable or increased population. At the same time, the statistics have not demonstrated any obvious relations between a population decrease and the proportion of unemployed in the population, the percentage of females among the unemployed, the proportion of people with primary school education, or the proportion of males.

In general, the major lesson that we have learned from our experiments on applying the classification tools of data mining in exploratory data analysis is that the outcome of such a method has value mainly as a hint to the explorer concerning what attributes *might* be related to a given division of the references into classes. To make use of such a hint, the explorer

needs to apply other exploratory tools to study whether these attributes are really related to the classes, and how they are related.

It is also important to take account of the fact that a classification tree does not necessarily involve *all* potentially related attributes. Therefore, it makes sense to run a classification tool several times with different subsets of selected attributes. For example, Fig. 4.131 demonstrates a tree constructed by the J48 tool for the same classes of districts where only the attributes characterising the educational level of the population of the districts were allowed to be used. From the tree, we can immediately see that the educational level in the districts with a decreased population was quite low in 1981: in 147 of 168 districts, the percentage of people without primary school education was over 52.62%.



**Fig. 4.131.** Another classification tree has been constructed using only the attributes characterising the educational level of the population of the districts in the years 1981 and 1991

Again, we need to use some other tool(s) in order to see which of the attributes appearing in the tree are really related to a population decrease. Thus, if we use the statistics-computing tool demonstrated in Fig. 4.124, we note a salient shift towards a higher proportion of people without primary school education in 1981 in the districts with a population decrease as compared with those with a stable population or a population increase. An analogous observation can be made concerning the values of the same attribute in the year 1991. In addition, the districts with a population decrease have a tendency towards a lower proportion of people with high school education in both 1981 and 1991 and lower proportions of people with preparatory school education in 1991. In general, we can conclude that there is a link between a population decrease and the educational level.

We are not able to include examples of the use of other data-mining tools. On the one hand, these tools are very numerous, but on the other hand, we have only a limited set of tools at our disposal, and not all of them are suitable for generating good and easily understandable examples of data exploration. We suggest that interested readers should refer to data-mining handbooks, for example Klösgen and Żytkow (2002).

#### 4.7.6 Example: Data Preparation

Earlier we presented a general paradigm for applying computational tools in EDA (see Fig. 4.119). According to this paradigm, the analysis process involves five steps, which are typically performed iteratively. In our examples, however, we have not paid equal attention to all five steps but have mostly focused on the exploration of the outcomes of computational tools. We have also demonstrated returning from the result exploration stage to earlier steps: the application of the same computational method after changing its input settings (where we selected a different set of attributes for the classification tree tool), and choosing another computation method (where we tried two different tools, in the clustering example). Although we did not describe it explicitly in relation to the examples, pre-viewing of data by means of visualisation and display manipulation preceded all other steps in the analysis. In fact, we have used the Portuguese dataset so heavily in our examples throughout the book that there was no need to describe once again how we could visualise the age structure or the attributes characterising the education level.

What was really missing in our examples was the preparation of the data for the application of computational tools. We applied the data-mining methods to the original data present in the Portuguese dataset; there was simply no need for any special preparation of the data. However, this is not always the case. Let us briefly describe an example where data have to be transformed before a data-mining tool can be applied.

Earlier, we discussed the raster format for spatial data and the problems arising in the joint exploration of several attributes. We considered an example concerning the distribution of different types of forest over Europe; see Figs 4.86C and 4.87C. In order to be able to visualise and explore the forest structure, i.e. the proportions of different forest types, we aggregated the raster data by means of the cells of a regular rectangular grid.

The same transformation can be used to make it possible to apply data-mining or other computational tools to data initially provided in a raster format. In fact, we are not aware of any data-mining tools that can work directly with raster data. Most data-mining tools require the data to be in

table format, where each row contains values of various attributes corresponding to one reference.<sup>25</sup> Hence, one needs to transform raster data to this format before trying to do data mining. The raster aggregation tool discussed earlier does exactly what is needed: after having aggregated the data by use of grid cells, it puts the resulting characteristics of the cells in a table, which can then be processed by data-mining tools.

Let us see how we can analyse the European forest data with the use of data mining. The data are specified in five raster files, each containing proportions of some type of forest or land: coniferous forest, broadleaved forest, mixed forest, other wooded land, and non-forest land. We have applied the raster aggregation tool, which generated a rectangular grid with a specified resolution and then computed the mean proportions of each type of forest or other land in the cells of the grid. The means computed in this way have been put in a table, in which each row characterises a certain grid cell. Figure 4.132C demonstrates the result of applying a clustering tool to the transformed data. In accordance with our request, the tool has produced five clusters. We have constructed a map display where the result of the clustering is shown by displaying the grid cells in different colours depending on the clusters in which they are included. The assignment of the colours to the clusters is arbitrary. The map can be seen in the upper left corner of Fig. 4.132C.

In order to understand the meaning of the clusters that have been constructed, we have visualised the characteristics of the grid cells in a parallel-coordinates display. We have applied query tools (multicolour marking and filtering) in order to consider each cluster individually and to compare different clusters. Figure 4.132C contains five screenshots of the parallel-coordinates display; each screenshot represents the characteristics of the members of one cluster. We can see that each cluster consists of cells with quite consistent profiles:

- *green*: A high proportion of coniferous forest, a medium to quite high proportion of mixed forest, and a low proportion of broadleaved forest, other wooded land, and non-forest land;
- *red*: Presence of all forest types in low to medium proportions, and a medium proportion of non-forest land;
- *blue*: Prevalence of mixed or broadleaved forest, while coniferous forest is also present; low proportions of other wooded and non-forest land;

---

<sup>25</sup> There is a group of data-mining methods specifically designed for analysing spatial data. Some of these methods work with other data representation formats instead of or in addition to tables. Other methods require the spatial components of the data (e.g. coordinates or boundaries) to be included in the tables.

- *yellow*: Mainly non-forest land;
- *pink*: Prevalence of other wooded land and a medium proportion of non-forest land.

The map shows us the geographical distribution of these different profiles of forest/land structure. The northern territories are the most wooded, having much coniferous (green), mixed (green and blue), and broadleaved (blue) forest. The red and yellow spots in the north may correspond to mountains or subarctic regions. Central Europe is mainly poor in forest, with much non-forest land (yellow) and rather small proportions of the various forest types (red). The south of Europe is clearly distinguished by the prevalence of other wooded land.

As we described earlier, the raster aggregation tool allows us to change the grid resolution. In response, it automatically recomputes the characteristics of the new grid cells. We were interested to see whether changing the grid resolution would have any influence on the outcome of using the clustering tool. So, we requested the aggregation tool to construct a finer grid than before and reapplied the clustering method to the new aggregated data. The result may be seen in Fig. 4.133C.

Certainly, the map in Fig. 4.133C shows us the geographical distribution of the different forest structures in finer detail than the map in Fig. 4.132C. However, the overall spatial pattern that we have perceived from Fig. 4.132C remains valid. From the parallel-coordinates displays in Fig. 4.133C, we can see that the characteristic profiles of the clusters remain the same; only the line density has increased, since we have now about four times as many grid cells as there were in Fig. 4.132C. Thus, the blue line crossing the “other wooded” axis far apart from the other blue lines in the lower left image in Fig. 4.132C have been replaced in Fig. 4.133C by a bundle of four lines (in both figures, these lines correspond to the small blue spot that can be seen in the north-west of the Iberian peninsula).

This consistency between the results of applying the clustering tool to grids with different resolutions increases our confidence in the validity of the analysis and, in particular, of the data transformation involved in it.

#### 4.7.7 Recap: Computational Tools

The role of computational methods in data analysis is increasing as computer technology provides, on the one hand, more and more computational power, and on the other hand, more and more capacity for storing various types of data, which, in turn, expands the amounts of collected data that need to be analysed. The contemporary computational methods are very

numerous; therefore, we could not comprehensively enumerate and describe them here. We have just slightly touched upon two major classes of computational tools relevant to exploratory data analysis, namely the tools of statistics and data mining.

The main idea that we wanted to convey by our considerations was that, although computational techniques may be very useful in exploratory data analysis, it is never sufficient to use them alone. They should always be combined with other analytical tools, first of all visualisation. The reason lies in the primary goal of EDA, which is to understand phenomena represented by data. Computations may contribute to understanding of the phenomena only if their results are appropriately interpreted. Any figures, formulae, classes, structures, etc. obtained from computations are not yet knowledge nuggets by themselves; but they can be transformed into knowledge nuggets by means of appropriate interpretation. Hence, the results of computations need to be “made perceptible to the mind or imagination”, i.e. *visualised*, according to the definition given in Random House (1996).

Let us recall those few things which we have said concerning computational analysis tools. Speaking about statistical techniques, we have mentioned two major categories, descriptive and inferential techniques. Descriptive statistical techniques are intended for the summarisation of data, i.e. expressing the most important features of a dataset in a few numbers. These techniques are recommended for use only in combination with statistical graphics; otherwise, they may be entirely misleading. Inferential techniques allow an analyst to check whether his/her observations made on the basis of data characterising some part of a phenomenon (a sample) can be taken as valid for the entire phenomenon. Hence, it is assumed that some observations have previously been made and some hypotheses generated. In other words, exploratory data analysis needs to be performed before inferential statistical tools are applied, but these tools themselves are not exploratory in their nature.

Data-mining techniques are more exploratory, but are not always aimed at understanding phenomena. Data mining aims first of all at deriving *useful* models, which do not necessarily explain something but allow one to predict the development of a phenomenon or the consequences of various decisions that can be potentially made. The results of data mining may still be intelligible and contribute to understanding phenomena, but, for this purpose, they need to be explored with the use of appropriate tools.

We have considered examples of the application of two different types of data mining tools, clustering and classification. Clustering is helpful when it is necessary to explore a distribution of characteristics expressed by multiple attributes over a reference set. Clustering techniques group

references with close characteristics together and thereby allow the explorer to consider a relatively small number of distinct characteristic profiles instead of the original multitude of various combinations of attribute values. However, the results of clustering do not explicitly contain these profiles. Only appropriate visualisation in combination with other exploratory techniques allows the explorer to reveal the profiles and to judge the variability of characteristics within the clusters.

Classification techniques are mostly oriented towards prediction. Their value for data exploration may be to suggest to an analyst what attributes may deserve attention as being potentially related to a specified classification of references. These attributes can then be explored using other tools. It is worth mentioning that this manner of using computational techniques is sometimes used in software packages for information visualisation. The idea is to use the results of computations in order to optimise the data display for better perception and to reduce the cognitive load of the viewer. For example, a large collection of documents may be visualised in a compact way on the basis of grouping of the documents by similarity (Wise et al. 1995, Dodge 2000), computation of the degree of relatedness between pairs of attributes may be used to obtain an optimal arrangement of the axes of a parallel-coordinates display or of scatterplots in a scatterplot matrix (Friendly and Kwan 2003), and individual dots on scatterplots in a scatterplot matrix may be replaced by precomputed figures representing the shapes of dot clouds (Friendly 2002).

We have described a general paradigm for applying computational tools in EDA; see Fig. 4.119. The main features are the iterative character of the analysis process and the prominent role of visualisation both before and after the computations. A single run of one computational tool is usually not sufficient. One should at least investigate the sensitivity of the results of the computation to changes in the tool parameters. Often, a data transformation needs to be done before a computational tool can be applied. We have given an example where a transformation of the data format was necessary. This is not the only possible case. Thus, it is often reasonable to transform absolute attribute values into relative values or, in time series data, original values into changes.

In this section we have paid much attention to the combination of computational and visual tools. In the examples given in the previous section, we did not use query tools alone but used them in combination with data displays. Moreover, several data displays were used simultaneously. In general, exploratory data analysis is done with the use of multiple tools and concurrent displays. Let us now look at the ways and mechanisms that exist for tool combination and display coordination.

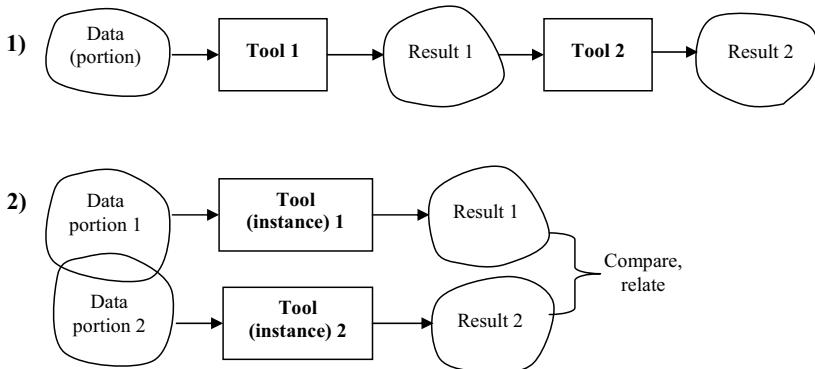
## 4.8 Tool Combination and Coordination

It does not seem necessary to us to go into wordy explanations concerning the importance of using various tools in data exploration. It is clear that each tool has different capabilities, and that there is no tool capable of doing everything. It can also be noticed that the datasets that need to be analysed are often very large and consist of many components. One could hardly find a tool that could process all the data at once and, in the result, tell us everything we would like to know about it. It is usually necessary to process data piecewise and link the fragmentary information thus obtained into a coherent view. Therefore, an explorer needs not only to apply different tools but also to apply one and the same tool several times, sequentially or concurrently, to different portions of the data.

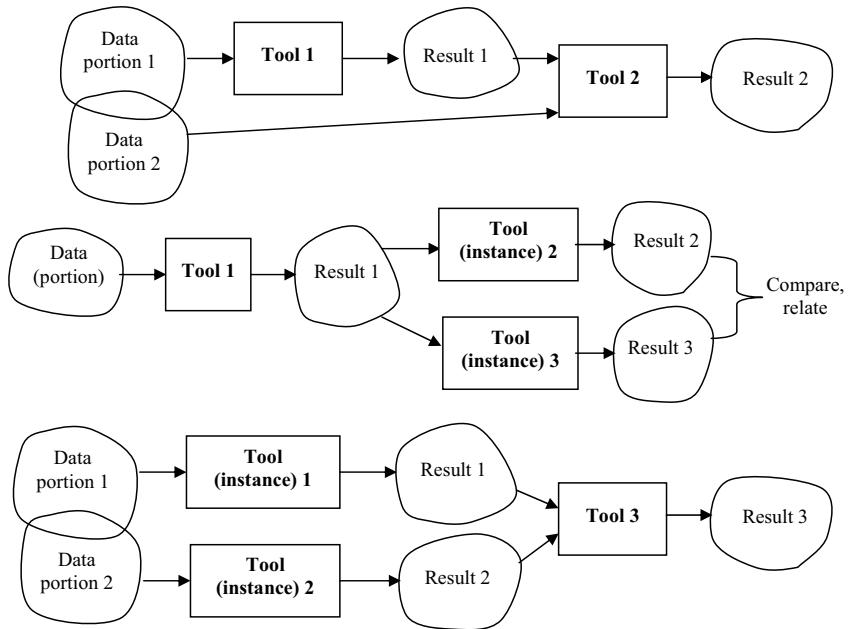
Since there is a necessity to use multiple tools and/or multiple “instances” of the same tool, these tools or instances need to be properly combined and, in the case of concurrent usage, coordinated. The main reason for combination and coordination is to facilitate the process of linking fragmentary observations into a general understanding.

There are two basic modes of combining tools or tool instances in data analysis:

- 1. Sequential mode:** A tool is applied to the outcomes of another tool.
- 2. Concurrent mode:** Two or more tools or tool instances are applied independently, and the analyst needs to compare and relate their results. The tools or tool instances may be applied to the same portion of the data or to (partially) different portions.



**Fig. 4.134.** Two basic modes for tool combination in data analysis. (1) One tool is applied to an output of another tool. (2) Two different tools or instances of the same tool are applied independently to the same portion of the data or different portions. Their results need to be compared and linked into a common picture



**Fig. 4.135.** Various “hybrid” combinations may be derived from the two basic modes of tool combination presented in Fig. 4.134

These two modes are shown schematically in Fig. 4.134. Various “hybrid” modes are also possible. Some variants are shown in Fig. 4.135.

Let us now consider each of the two basic modes in more detail.

### 4.8.1 Sequential Tool Combination

With this mode of tool combination, one tool produces an output, which is used as an input for another tool. We have already considered various categories of tools used for exploratory data analysis. Let us recall what types of outputs they produce and what types of inputs they require for this purpose. Here is a list of the types of tool outputs extracted from our descriptions of various tools:

- *Visual display*, which may be supplied with display manipulation tools. Although display manipulation tools are primarily intended to modify the visual encoding of the information represented in the display, some of them may also produce other types of results, such as ordering or classification. These other results may be considered as intermediate with respect to the main goal, i.e. modification of the visual encoding function. However, nothing prohibits an independent use of these results

as, in particular, inputs of other data analysis tools. Therefore, we include the intermediate, non-visual results of display manipulation tools in our list of the types of tool outputs.

- *Ordering of references*, an output of an ordering tool.
- *Division of the reference set* into subsets as a result of classification, querying, or clustering.
- *Selection of a reference subset*, which may be an output of focusing, zooming, or querying.
- *New attributes*, which may result from attribute transformation or integration. In fact, it is possible to treat the results of reference ordering, classification, or subset selection as new attributes also. The results of ordering may be represented as an attribute with values indicating the positions of the references in the arrangement. For example, a linear ordering could be reflected in an attribute with values that are integer numbers from 1 to the number of existing references. A division of the reference set into subsets (classes or clusters) may be represented as a qualitative attribute with as many values as there are subsets. For each reference, the corresponding value of this attribute indicates which subset this reference belongs to. Analogously, a qualitative attribute may reflect the result of subset selection. In this case, the attribute will have two possible values, “selected” and “not selected”.
- *New references* with corresponding characteristics. New references may result from interpolation and aggregation tools. In the first case, the new references have the same nature as the original references and can simply be added to the reference set, thereby extending it. In the second case, the new references are aggregates, i.e. their nature and level are different from those of the original references. Therefore, these new references should not be mixed with the original ones; they need to be considered separately.
- *Relations between references*, with respect to their characteristics, for example distances, neighbourhood, and similarity. Such relations may be obtained from querying tools.
- *Relations between attributes* (e.g. statistical correlations) and ordering of attributes on the basis of such relations. This type of result can be generated by computational tools, such as statistical or data-mining tools. Attribute ordering may also be an outcome of a display manipulation tool, for example a tool for ordering the axes of a parallel-coordinates display.
- *Summaries* of the dataset or its subsets, such as various descriptive statistics, classification trees, rules, or formulae.

The most common types of inputs for data analysis tools are references and the corresponding values of one or more attributes. Tool outputs in this form can easily be sent to other tools for further analysis. This applies to tools that produce new references and/or new attributes, including tools for ordering, division, and selection, whose results may be represented as new attributes. The results of these tools may be visualised, transformed, aggregated, queried, mined, summarised, etc. just like the original data.

Other types of results may be more restrictive concerning the possibilities for their further processing. The most restrictive of these types is visualisation: in the classes of tools that we have considered, there are no tools that could use visual displays as their inputs. We do not claim that this is impossible in principle, but we have never encountered such analytical techniques.

Relations between references can be visualised or processed by tools specifically designed for this purpose. Thus, binary relations could be visualised in a matrix-like display, where the rows and columns correspond to individual references, and the relations are represented in the cells using some visual encoding, for example by colours. Some information visualisation tools treat similarity relations between references metaphorically as distances in space and, on this basis, arrange the references into “information landscapes”, which are visualised in maps or three-dimensional surface displays; see, for example, Wise et al. (1995) and Dodge (2000). Such an arrangement involves rather intensive computation. As we have already mentioned, there are also specific computational tools for spatial data analysis, which take account of spatial distances or neighbourhood relations between references.

While relations between references require, in general, specialised tools for visualisation and/or further processing, it is often possible and reasonable to visualise and analyse some of these relations, which may be represented by means of attributes of the usual kind. Thus, in one of our examples, we considered similarity relations between districts of Portugal in terms of the age structures of the population; see Figs 4.105–4.109. We did not try to analyse the whole set of similarity relations at once. Instead, we looked at a subset, specifically, the relations of similarity to the district of Porto. This subset was represented by means a new attribute with values expressing the degree of similarity to Porto. The new attribute could be visualised and analysed with the use of various types of displays (maps, tables, parallel-coordinates displays, etc.), as well as many other types of tools capable of dealing with references and attributes. In a similar way, we considered the relations of similarity to a few other districts. However, it would be difficult to analyse the whole set of similarity relations in this way.

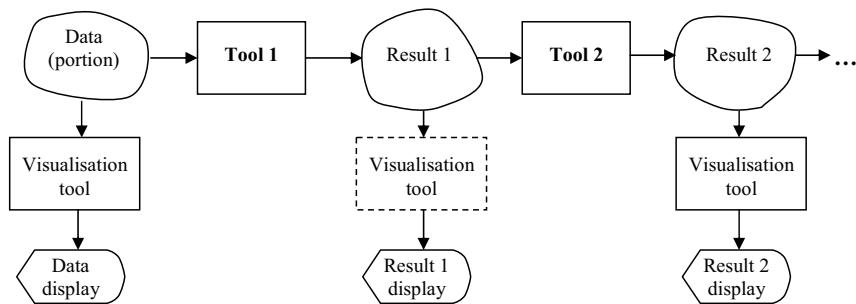
Relations between attributes and ordering of attributes also have quite limited usability as inputs of data analysis tools. This type of information can be used, for example, for an effective arrangement, from the perspective of perception, of the axes in a parallel-coordinates display or of scatterplots in a scatterplot matrix; see, for example, (Friendly and Kwan 2003). Another way of utilising this information is in the selection of attribute combinations for analysis. A dataset may contain many attributes, such that an analyst cannot explore all possible combinations. A computational tool may establish relations between attributes and, on this basis, suggest the most informative combinations (“projections”) for further analysis.

Dataset summaries resulting from computational tools typically have a rather specific form and therefore require dedicated tools for visualisation and analysis. An exception is the summary statistics such as mean, median, and mode, which can be used as characteristics of aggregates. In this case, groups of references are treated as new references, and the summary characteristics of these groups as the corresponding attribute values. The new references and new attributes can be visualised and analysed like “normal” references and attributes. However, specialised tools for the analysis of aggregates also exist and are often more preferable. As we have mentioned in the Sect. 4.5.4, one should be very cautious when using any average characteristics, since they may be totally misleading. To avoid rash and unsound judgements and conclusions, one should examine how the attribute values are distributed within each aggregate. If the distribution is close to normal, the variation is low, and there are no outliers, the data analyst can use averages such as the mean, median, or mode as characteristics of the aggregates in the further analysis, i.e. apply other exploratory tools to these values. Hence, a three-component chain of analytical tools may be recommended in the case of aggregation:

1. A tool to define reference subsets and explore the distribution of attribute values over each subset.
2. A tool to produce appropriate general characteristics of aggregates.
3. A tool applied to the aggregates and their general characteristics for further analysis.

Generally, sequential tool combination is not limited to only two tools, as is shown in the diagram in Fig. 4.134; the chain may consist of any number of tools. Thus, in one of the latest examples we have considered (see Figs 4.132C and 4.133C), a raster aggregation tool was applied to the original data in raster format, then the results of the aggregation were processed by a clustering tool, and then visualisation and display manipulation tools were used to view and interpret the results of the clustering.

It is quite typical that a chain of tool applications ends with visualisation: in any case, the explorer needs to see the ultimate results of the analysis. However, the explorer may also be interested in viewing intermediate results. In fact, it happens very often that the entire chain of data analysis is not planned in advance. A typical scenario is that the analyst applies a tool, looks at the results obtained, and then decides whether these results should be processed further by means of another analytical tool and, if so, what tool to use for this purpose. Hence, the scheme of sequential tool application might be elaborated, as is shown in Fig. 4.136.



**Fig. 4.136.** The involvement of visualisation tools in sequential tool combination

When we say that Tool 2 is applied to the results of Tool 1, we do not mean that these results are necessarily final and static, i.e. they never change after the application of Tool 2. Throughout this chapter, we have considered many examples of dynamic tools, i.e. tools whose results can change, in particular in response to various interactive operations performed by the user. Let us recall some of these examples:

- Dynamic classification according to values of a numeric attribute. The user may change the class breaks and the number of classes (see Figs 4.16 and 4.18).
- Dynamic integration of attributes. The user may change computational parameters such as the attribute weights in computing weighted sums of values of multiple attributes (see Figs 4.52–4.57).
- Dynamic aggregation (e.g. Figs 4.68 and 4.71), in particular, raster aggregation (Figs 4.86C, 4.87C, 4.132C, and 4.133C).
- Dynamic querying (e.g. Figs 4.94 and 4.96).

If Tool 1 is a dynamic tool and Tool 2 is applied to its results, it is necessary that any changes in the results of Tool 1 are reflected in Tool 2 and further along the chain. This means that Tool 2 must be reapplied to the modified results of Tool 1, the tool following Tool 2 in the chain of tools

must be reapplied to the new results of Tool 2, and so on. In modern packages for exploratory data analysis, such a reapplication of tools is often done automatically: when a tool in a chain changes its results, it notifies all other tools using these results about the change that has occurred. In response, these other tools automatically update their own results and, in turn, notify the tools following in the sequence, and so on. An example of such automatic reaction can be seen in Figs 4.86C and 4.87C: after the granularity of the grid used for the raster aggregation was changed, the visualisation tool reapplied the previously used visualisation technique to the new aggregates and their characteristics.

However, it may also happen that a tool that is applied to results of another tool is not capable of such an automatic reaction to changes in those results. This depends not only on the implementation of the tool itself but also on the availability of an appropriate software platform that enables tool linking. In cases where tool reapplication and updating of the results do not occur automatically, the explorer needs to take care about this. For example, the clustering tool that we applied to the aggregated raster data in one of our latest examples (see Figs 4.132C and 4.133C) does not automatically react to changes in the data that it is applied to. Therefore, after we had changed the granularity of the aggregation, which resulted in a new set of aggregates with new characteristics, we had to send the new data to the clustering tool and rerun the tool.

The absence of automatic tool reapplication and updating of results is not always a handicap to analysis. It may be easier to compare the results of several tool runs with different input settings when the results of the previous runs remain unchanged (until the user explicitly performs certain actions to change them) than when all the tools are very reactive, so that all results along a chain change immediately after even a slight interaction of the user with the first tool.

## 4.8.2 Concurrent Tool Combination

Concurrent tool combination means that two or more tools are independently applied to data, and the user may, in principle, view and analyse the results of each tool with no regard for the results of the other tools (this means that visualisation of the results of the tools is necessarily involved). However, there are special mechanisms intended to facilitate comparison of such results and mental integration of the information conveyed by each of them into a coherent general picture of the data and of the phenomenon characterised by the data. In order to clarify from the very beginning what we mean, we would like to refer to one of the examples given earlier.

Figure 4.94 demonstrates the operation of the Dynamic Query tool: when a user specifies a query condition, the tool divides the reference set of the dataset into two subsets: references with characteristics satisfying the query, or active references, and the remainder, or inactive references. In the example, the references are the districts of Portugal. The division occurs independently of the operation of any other tool.

Simultaneously with Dynamic Query, several other tools operate and present their results on the screen: a map display of the districts of Portugal, two histograms showing the results of different aggregations of the districts, and a scatterplot. These other tools are independent of each other, as well as of Dynamic Query; they were applied to the original data before Dynamic Query started its operation. However, these tools are “listening” to everything that happens to the data that they are applied to. Therefore, after the activation of Dynamic Query, they “notice” that the reference set has been divided into active and inactive references. In response, they reflect this event, each tool in its own manner (see the description of Fig. 4.94 in the text).

Any modification of the query conditions in the Dynamic Query tool changes the division of the reference set into active and inactive references. The other tools “notice” the change and reflect it in their appearance, as is shown, for example, in Fig. 4.96.

The dynamic updating of a display according to the results of a query tool allows an analyst to relate the characteristics represented on this display to the characteristics involved in the query. Thus, the map in Fig. 4.94 allows us to relate low or high values of the attribute “% 0–14 years” in the districts of Portugal to the geographical positions of the districts. Looking at the histograms, we can see how low or high proportions of children are related to the percentages of people employed in agriculture and in industry. In the scatterplot, we can detect a negative correlation between the proportion of people employed in services and the proportion of uneducated people in the districts with a low percentage of children. Such fragmentary observations contribute to building our overall understanding of the data.

This is not the only example of concurrent tool combination and coordination that can be found in our book. In fact, almost all of the examples given in the Sect. 4.6 involve dynamic reaction of various visualisation tools to specification and modification of a query. We have discussed two basic modes of such a reaction, filtering and marking, or “brushing”. Marking may result, in particular, from direct manipulation of a data display on the screen, such as clicking on graphical element (see Figs 4.98 and 4.99) or enclosing graphical elements in a frame (see Fig. 4.100). This mechanism allows the user to find easily corresponding parts of different

displays, i.e. parts showing characteristics of the same subset of references. On this basis, the characteristics shown in different displays may be linked mentally into a coherent view.

In all of the examples of the use of dynamic query tools, there were several displays, the outputs of various visualisation tools, and one query tool, which was able to modify the appearance of the displays. In all cases, the query tool influenced the other tools through the division of the reference set into two or more (in the case of multicolour marking) subsets. The visualisation tools do not need to be directly related to the query tool; they need only to be informed about the division of the reference set when such a division occurs or changes.

If a tool is able to “notice” a reference set division and react to it, the division need not necessarily result from a query tool. Thus, any classification tool also divides the set of references into subsets (classes), and hence the classification may be reflected in various displays just as in the case of multi-colour marking.<sup>26</sup> The same applies to the clustering tools discussed in Sect. 4.7.4. In Figs 4.120C–4.123C, the results of the clustering are reflected simultaneously in a map and in four histograms, which change their appearance after the tool is reapplied and produces a different set of clusters. In this example, two modes of tool combination work together. The map display is applied to the results of the clustering tool; hence, this is a sequential combination. The histogram displays exist independently of the results of the clustering; they represent results of data aggregation, using for this purpose the heights of the bars. However, they show additionally the division into the clusters by an appropriate segmentation of the bars and colouring of the segments.

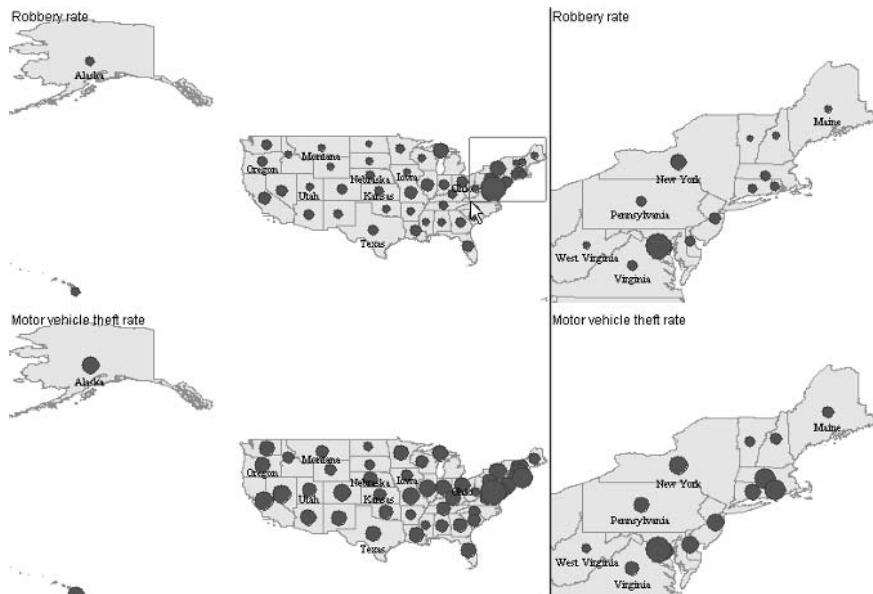
There is another approach to the combination and coordination of different tools operating concurrently. As we have discussed earlier, a tool may be applied to a subset of references rather than the entire set. For example, a map may show only part of a territory, or a time graph may represent only a subinterval of a time period for which time series data are available. Zooming and focusing tools are often used for the selection of reference subsets to be represented. In Figs 4.25–4.30, we demonstrated

---

<sup>26</sup> A problem may arise when a display intended to reflect a classification or multicolour marking already uses different colours for the representation of the primary information. For example, districts in a map may be coloured depending on the values of some attribute(s). This map will not be able to show a classification or marking of districts obtained from another tool without destroying the original representation. Therefore, some tools may have certain “selectivity” with respect to the results of other tools, for example, they may react only to filtering and single-colour marking.

the effect of applying such tools to various displays. In these examples, each tool affected a single display. However, it is also possible that several displays react simultaneously to one and the same zooming or focusing tool, and in this way are coordinated. This coordination mechanism is also based on a division of the reference set, but functions slightly differently from the propagation of marking or of classes of references. In the propagation of marking or classes, the division is shown in a data display *additionally* to the information that was present in this display before the division was made. In coordinated zooming or focusing, the division makes all the displays involved synchronously *replace* the information shown previously by other information (which may overlap partly with the previous information).

An example of coordinated zooming of two map displays is shown in Fig. 4.137. On the left, there are two maps of the states of the USA representing the values of two attributes, the robbery rate and the motor vehicle theft rate, in the year 1970. A zooming frame is drawn over one of the maps with the purpose of enlarging the display of the corresponding part of the territory. On the right, the result of the coordinated zooming operation is demonstrated: both map displays now show only the selected part of the territory, specifically, the north-east of the country.

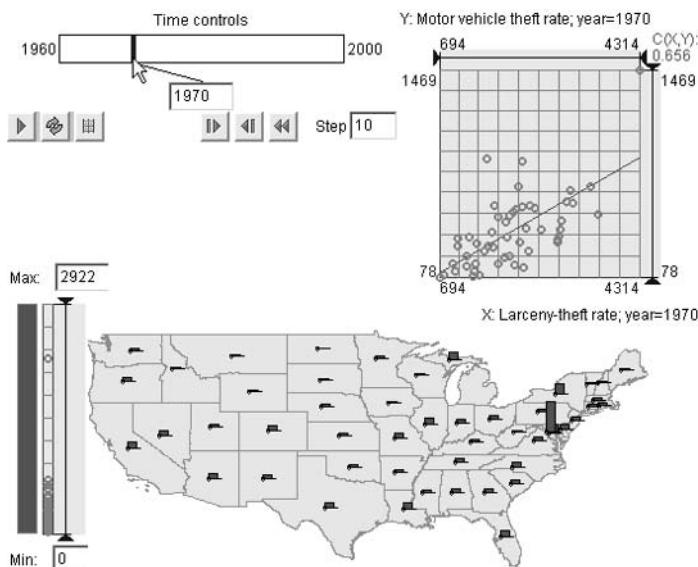


**Fig. 4.137.** Coordinated zooming of two map displays of the same territory representing different attributes

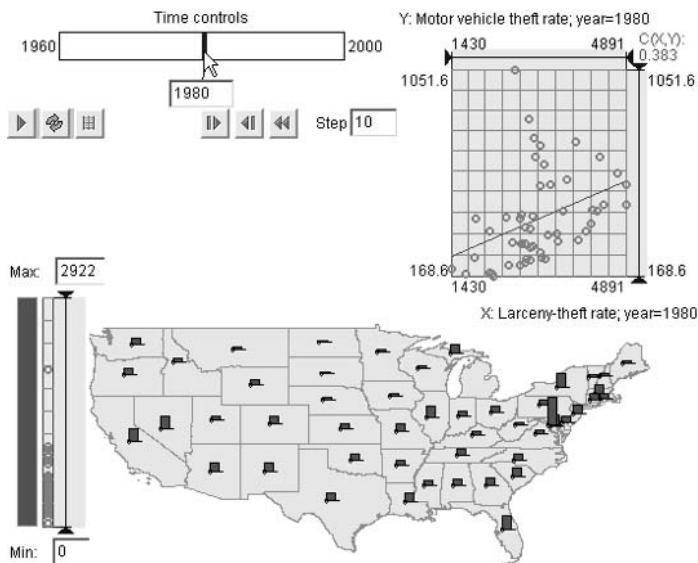
Another example that we would like to discuss deals with the use of the temporal display dimension in data visualisation. When the temporal dimension is used to represent a temporal referrer of a dataset, the display at each moment shows the part of the whole reference set that correspond to a certain selected time moment or interval. When another moment or interval is selected, the content of the display changes so as to represent the appropriate portion of the data. Such displays will be called *animated displays* in what follows. We shall also apply the modifier “animated” to particular display types and talk, for example, about animated maps and animated scatterplots.

Several animated displays simultaneously present on the screen may be controlled through a common user interface device that allows the user to select the current time moment and, in particular, to start automated display animation, where the current moment is periodically incremented without waiting for a command from the user. The displays will react synchronously to changes of the current time moment, either manual or automatic.

An example of synchronous reaction of several animated displays to a change in the selection of the current time moment is demonstrated in Figs 4.138 and 4.139. As before, this example refers to the dataset containing the yearly crime data for the states of the USA over the period 1960–2000.



**Fig. 4.138.** Selected animated displays represent crime data for a selected year, 1970. The displays have a common time control device (shown in the upper left corner) for selecting the year



**Fig. 4.139.** A change of the current time moment in the time control device from 1970 to 1980 makes all animated displays simultaneously update their information content

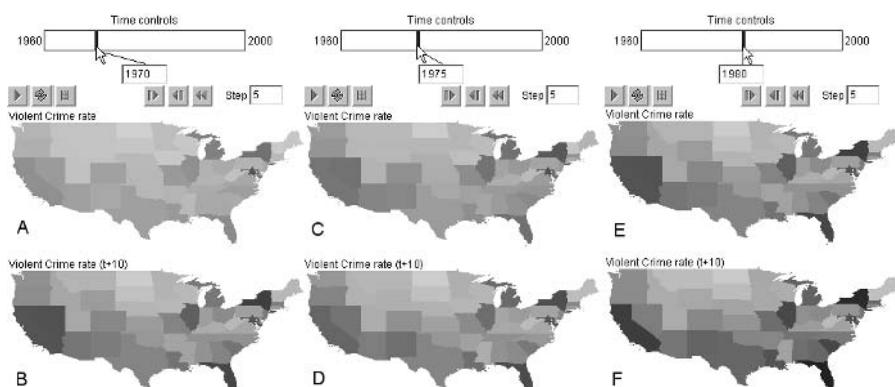
In the upper left corner of each figure, one can see a collection of interactive widgets called “Time controls”, which are used for selection of the current time moment. Upon any change of the current time moment, the device issues a corresponding notification. Any display that deals with the temporal component of the dataset may “listen” to such notifications and modify its information content appropriately. In this example, there are three animated displays: a scatterplot representing the attributes “Larceny-theft rate” and “Motor vehicle theft rate”, a map representing the attribute “Violent crime rate” by heights of bars proportional to the values of this attribute, and a dot plot, which is included in the focusing device supplementing the map.

Figure 4.138 demonstrates the state of all the displays when the year 1970 is chosen as the current time moment. All the displays show the values of their respective attributes referring to the year 1970. In Fig. 4.139, the current time moment has been changed from 1970 to 1980. As a result, all the displays have simultaneously updated their information content and now represent the attribute values referring to 1980.

When we compare Fig. 4.139 with Fig. 4.138, the differences between the situations in 1970 and 1980 are quite noticeable. However, we should not forget that, in reality, the picture shown in Fig. 4.139 replaces the one from Fig. 4.138, and the latter cannot be seen any more. This greatly hin-

ders the comparison of the two situations. A prerequisite for an effective comparison is that the things to be compared are visible simultaneously and, moreover, placed close to each other. As we have discussed earlier, an arrangement of multiple maps corresponding to different time moments is better suited for making comparisons than is an animated single-map display. However, the number of maps or other displays that can be simultaneously present on the screen is very limited. Thus, in the case of the US crime data, we would need 41 maps to represent the value distribution of any crime attribute in each year from 1960 to 2000. This is, in principle, possible but hardly useful.

A possible compromise solution is to have a restricted, manageable number of simultaneously visible displays corresponding to some selected time moments but to be able to change the selection without much effort. One approach is that the time moment represented in each display is controlled individually. Another approach is demonstrated in Fig. 4.140: two animated map displays are controlled through a common time selection interface, but one of the displays is “shifted” in time with respect to the current selection.



**Fig. 4.140.** Two coordinated animated map displays are manipulated here through a common time control device. The upper maps show the situation in the selected year, and the lower maps the situation ten years after the selected year

Here, three screenshots are shown, which correspond to different selected years: 1970, 1975, and 1980 (from left to right). The upper map in each screenshot (i.e. the maps labelled A, C, and E, respectively) represents the distribution of the violent-crime rate over the states of the USA in the selected year using the unclassified-chloropleth-map technique.<sup>27</sup> The

<sup>27</sup> To save space, we have “cut off” the states of Alaska and Hawaii. For a more expressive display, we have used a focusing tool to remove outliers, specifi-

lower map represents the value distribution of the same attribute using the same technique (and, moreover, the same visual encoding function). However, it shows the data referring to the moment ten years after the selected moment. This means that the maps labelled B, D, and F reflect the situations in the years 1980, 1985, and 1990, respectively. The upper and lower map displays are coordinated so that a change of the current time moment results in simultaneous automatic updating of both of them: the upper map represents the selected moment  $t$ , and the lower map the moment  $t + 10$ . Hence, the user can easily choose any two moments with a 10-year interval between them for doing pairwise comparisons. Additionally, the second display may be supplied with a convenient user interface device for changing the time shift with respect to the current selection. This facilitates choosing arbitrary pairs of moments.

In Fig. 4.140, coordination with a time shift has been applied to homogeneous displays, specifically, two maps representing the same attribute in the same way. However, it may also be useful to coordinate animated displays of different attributes in this way; these displays may employ the same or different visualisation techniques. Another useful application is overlaid animated visualisations, in particular, maps with multiple animated layers representing different spatial phenomena. Such visualisations are described, for example, in Edsall and Peuquet (1997) and Blok et al. (1999). The user may run a map animation after specifying an offset for any animated map layer with respect to the current display time. The goal is to help an analyst to detect and investigate cause–effect relationships between phenomena when effects caused by events or changes appear after a delay.

Let us now try to summarise the variants of concurrent tool combination considered thus far. All of these variants involve the simultaneous reaction of several visualisation tools to one of two types of events:

1. Selection of a reference subset (by means of dynamic querying or focusing tools).
2. Division of the entire reference set into two or more subsets (by means of dynamic classification or extended querying, which divides the reference set according to the satisfaction of different combinations of query conditions).

Although reference subset selection can be treated as a particular case of division into subsets, we prefer to consider it separately because this is more convenient for our further discussion.

---

cally, extremely high values in the District of Columbia. The removed outliers are signified by triangles placed at the corresponding locations in the maps.

The following types of reaction of tools to these types of events can be seen in the examples provided in the book:

- Subset selection:
  - *Highlighting* (single-colour marking) of display items corresponding to the selected references. Highlighting is shown on top of the information already present in the display. In aggregate visualisations, such as histograms, highlighting is applied to segments of the visual elements that represent aggregates.
  - *Filtering*, i.e. removing or “muting” display items corresponding to the references that are not selected.
  - *Focusing*, i.e. adjusting the display so as to represent the selected reference subset and the corresponding characteristics with the maximum possible expressiveness and distinctiveness at the cost of skipping the rest.
- Division:
  - *Multicolour marking* of display elements. Like highlighting, multicolour marking is combined with the previous information content of the display and applied in aggregate visualisations to segments of the visual elements that represent aggregates.
  - *Display multiplication*. A display is replaced or supplemented by several displays, each representing one of the reference subsets and the corresponding characteristics.
  - *Rearrangement* of display items so as to group them spatially according to the division of the reference set. Colour marking is typically also used in addition to rearrangement.

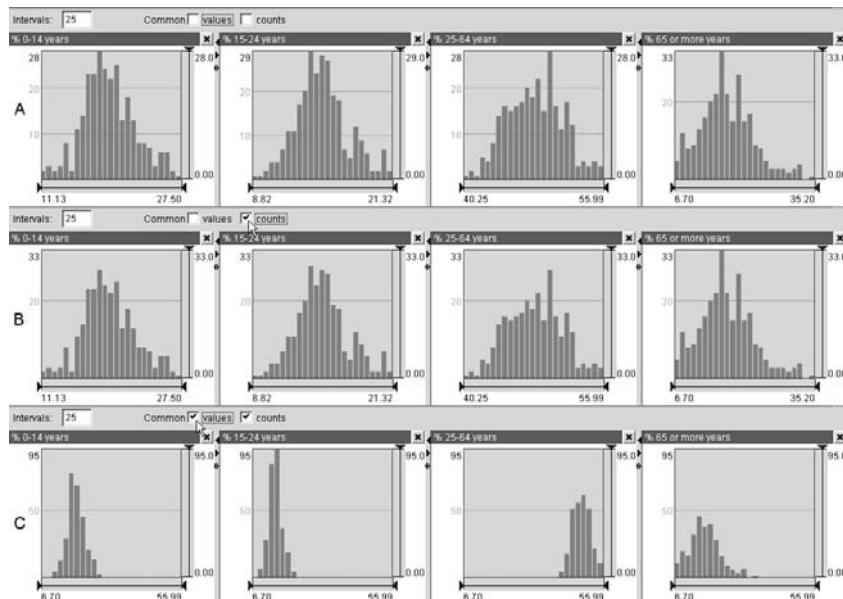
Although we have not mentioned the latter two types of reaction in this section yet, examples of them are present in the book. Examples of display multiplication can be seen in Figs 4.20, 4.132C, and 4.133C (multiplied parallel coordinates displays), 4.24 (multiplied maps), and 4.124 (multiplied box-and-whiskers plots). An example of rearrangement is the grouping of rows in a table display so that rows corresponding to references from the same subset are put together. This can be seen, although not very prominently, in Fig. 4.105.

As may be noticed, different reactions to one and the same type of event are possible. In principle, any particular tool may respond in its own way, i.e. coordinated tools may not necessarily demonstrate the same behaviour with respect to the coordination mechanism used. However, it is desirable that coordinated tools react similarly to each event, since inconsistent be-

haviours can hinder mental integration of the information from different displays and even cause frustration to the user.

We admit that the two-level taxonomy that we have introduced does not encompass all imaginable methods of tool coordination but only the most generic and widely used ones. An example of a less generic coordination technique is the use of a common visual encoding function in different displays. Thus, the map displays in Fig. 4.137 represent the values of two different attributes using the same function for encoding the values by circle sizes. This means that circles with the same size represent the same numeric value irrespective of the display in which these circles occur. The display manipulation tools (such as the visual comparison tool; see Fig. 4.35C) are in this case also common to both map displays.

Another example of the same kind is the coordination of several histogram displays of different attributes, which can have common scales on the vertical and/or the horizontal axis, as is demonstrated in Fig. 4.141.



**Fig. 4.141.** Coordinated histogram displays. Section A: each display has its individual horizontal and vertical scales. Section B: the displays have been brought to a common vertical scale so that bar sizes can be compared between different histograms. Section C: both the vertical and the horizontal scales are now common to all the displays. The relative positions and sizes of the value ranges can now be compared

At the top (section A), we see four histograms for the various age structure attributes characterising the districts of Portugal: “% 0–14 years”, “% 15–24 years”, “% 25–64 years”, and “% 65 or more years”, from left to right. Each histogram has its individual scales on the horizontal and vertical axes. The horizontal scale is determined by the value range of the respective attribute, and the vertical scale by the maximum number of districts fitting into one interval. In each histogram, the horizontal scale is divided into 25 equal-length subintervals.

In the centre (section B), the histograms have been transformed so as to have the same vertical scale. This scale is now determined by the maximum district count per subinterval in all four histograms. In this case, the maximum count is 33; it is reached in the histogram of the attribute “% 65 or more years”. So, the upper edge of each histogram after the transformation corresponds to the number 33, and district counts are encoded by bar heights in the same way in all of the histograms. Hence, the user can now compare sizes of bars in different histograms.

At the bottom (section C), one more transformation has been applied to the histograms. In the result, all histograms now have a common horizontal scale, which is determined by the difference between the maximum and the minimum of the values of all four attributes. The beginning of the horizontal scale in each histogram corresponds to the value 6.70 (the minimum of the attribute “% 65 or more years”), and the end to the value 55.99 (the maximum of the attribute “% 25–64 years”). As in section B, the histograms also have a common vertical scale. Since the histograms now represent a longer value range (from 6.70 to 55.99) while the number of the subintervals remains the same (specifically, 25), the length of each subinterval has increased, and so has the maximum district count per subinterval, which now equals 95. So, the upper end of the vertical axis in each histogram corresponds to 95. As in section B, the user can compare bar sizes between different histograms. Additionally, the histograms show the relative positions and sizes of the value ranges of the attributes.

A limitation of tool combination through common visual encoding is that this method is applicable only to homogeneous displays (e.g. several maps or several histograms) of comparable attributes. One cannot link a map to a histogram in this way, nor displays of numeric and qualitative attributes, or displays of two numeric attributes, one of which has values ranging from 0 to 1 and the other has values from 100 to 1000.

We have thus far considered two groups of tool combination methods:

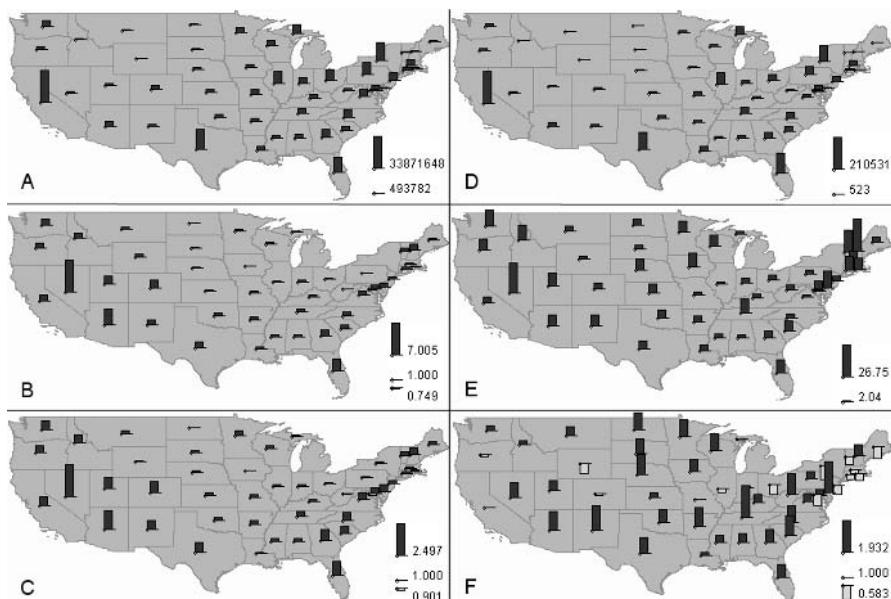
1. Combination on the basis of a common reference set (through division or subset selection).
2. Combination on the basis of a common visual encoding.

It is also possible to imagine two more categories:

3. Combination on the basis of a common set of attributes.
4. Combination on the basis of a common transformation of attribute values.

The third type of combination and coordination might be applied when several tools handle the same attributes in different ways. For example, one and the same subset of attributes may be represented simultaneously in a parallel coordinates display, a table display (as columns), a scatterplot matrix, and a map (e.g. by bar charts or pie charts). In that case, the user could select one of the attributes, and, in response, the corresponding parts on all the displays could be highlighted. Or the user could order the attributes, for instance by rearranging the axes on the parallel-coordinates display, and all the displays could react to this by an appropriate rearrangement of their corresponding parts: columns in the table, scatterplots in the matrix, and bars in the bar charts on the map. However, we have not encountered such a coordination mechanism in practice and are not quite convinced of its utility.

An example of the fourth type of combination is a simultaneous consistent transformation of several temporally referenced attributes, such as change computation or smoothing. Changes in transformation parameters apply to all attributes in this case. This is demonstrated in Fig. 4.142, which shows, from top to bottom, three different states of two map displays representing two different attributes. The upper pair of maps (A and D) portrays the original attribute values, while the other two pairs reflect the results of consistent transformations applied to both attributes in parallel. More specifically, the map images on the left in Fig. 4.142 (i.e. A, B, and C) correspond to the attributes “Population total”, and the images on the right (i.e. D, E, and F) to the attribute “Violent crime total”. The values of both attributes refer to the states of the USA and the years from 1960 to 2000. The upper two maps portray the values of the attributes in the year 2000. The same representation technique, bar symbols, has been used in both displays for more convenience in comparison (different techniques could also be used in this case). However, each display uses its own visual encoding function: on the map labelled A, the highest bar height corresponds to the value 33 871 648 (this is the maximum value of the attribute “Population total” in 2000), while on the map labelled D, the maximum bar height represents the value 210 531 (the maximum of the attribute “Violent crime total” in 2000). A similarity between the spatial distributions of the two attributes may be noticed from a comparison of the maps A and D.



**Fig. 4.142.** The same transformation has been applied simultaneously to two different attributes represented on two map displays. The maps on the left represent the attribute “Population total”, and the maps on the right the attribute “Violent crime total”. At the top, the original values of both attributes in the year 2000 are portrayed. The two maps in the middle represent the results of computing changes, specifically, the ratios between the values for the year 2000 and the year 1960. The maps at the bottom result from changing one of the transformation parameters; specifically, the year 1960 has been replaced by the year 1980

The maps in the middle, labelled B and E, reflect the result of transforming both attributes by computing ratios between the values for 2000 and 1960. Of course, each attribute has been transformed independently of the other, but the same transformation function, with the same parameters, has been used for this. We can see that the ratio for the total population ranges from 0.749 to 7.005 (map B), while the ratio for violent crimes ranges from 2.04 to 26.75 (map E). The spatial distributions of the transformed values are also quite different.

The maps C and F at the bottom of Fig. 4.142 result from a simultaneous change of one of the transformation parameters: the year 1960, chosen before as the base for the computation of changes, has been replaced by the year 1980. This parameter change has been applied to both attributes, which has resulted in a concurrent change in the map displays. Now, the transformed values of the attribute “Population total” portrayed in map C range from 0.91 to 2.497, and the transformed values of the attribute “Vio-

lent crime total" (map F) range from 0.583 to 1.932. While the change that occurs in the map of population numbers after the parameter is changed is not so striking (maps B and C are quite similar, and the difference in the ranges of the transformed values is moderate), the same cannot be said concerning the number of crimes. The range of the transformed values has changed dramatically. In quite many states, the total number of violent crimes decreases in comparison with 1980 (a decrease is signified by transformed values below 1), and the maximum increase is by a factor of 1.9 (see map F). With respect to the year 1960, the number of crimes in each state increased by a factor of 2 at a minimum, and the maximum increase was by a factor more than 26 (see map E). The spatial patterns are also quite different in maps E and F.

A certain commonality exists between tool combination on the basis of a common visual encoding and on the basis of a common transformation of attribute values. In both cases, attribute values are involved in some way. In the first case, they are transformed into visual items (positions within a display, colours, sizes, etc.), and in the second case they are transformed into non-visual things such as numbers or characters, which, however, need to be visualised in order to make the results of the transformation perceivable by a human. Coordinated transformation of attribute values is less restrictive in its applicability in comparison with common visual encoding. It can, in principle, be applied to heterogeneous displays such as maps involving different visualisation techniques, or a map and a time graph. Still, a big diversity of concurrent tools that apply a common data transformation is not necessarily useful even when achievable.

We do not claim that we have enumerated all possible ways of linking tools. One can, in principle, imagine various linking mechanisms that cannot be subsumed under our typology, and one can even find software implementations of these mechanisms. For example, changing weights in a tool for attribute integration (see Figs 4.52 and 4.56) may change the lengths of axes in a parallel-coordinates display, as is described in Andrienko and Andrienko (2001). However, these methods of linking are specific to certain types of tools and not easily generalisable. As for more generic methods, we believe that our inventory is sufficiently complete.

### 4.8.3 Recap: Tool Combination

There are two basic modes of combining tools:

- Outputs of one tool are used as inputs for another tool. The second tool starts its operation and produces results only after receiving the output of the first tool. This mode of tool combination is called *sequential*.

- Two or more tools operate independently, and their results can be seen simultaneously on the screen. However, there are certain mechanisms that facilitate the work of the analyst in comparing these results and linking the pieces of information provided by the separate tools into a coherent mental image of the features of the data under analysis, and of the underlying phenomena. This mode of combined use of several tools is called *concurrent*.

In sequential tool combination, it depends on the type of results a tool produces, what tools can be used after it to process these results further. As a rule, any non-visual tool requires visualisation of its results. We have considered various types of results, some of which require very specific tools for visualisation or further processing. However, some types of results, such as new attributes or new references, have the same form and meaning as the original attributes and references present in the data; hence, general data visualisation and analysis tools can be applied to these results.

It is important to remember that some tools can change their results, in particular, in response to interactive modification of the tool parameters by the user. If other tools use these results, they must take these changes into account. Visualisation tools must update their display so that the new results are properly represented, computational tools must rerun their computations, and query tools must once more check the satisfaction of the query conditions. Not all tools do this automatically; sometimes the analyst needs to reapply a tool when its input changes.

In concurrent tool combination, the analyst deals with several displays present simultaneously on the screen. Accordingly, the outputs of several concurrently operating tools are most often linked by means of display coordination, i.e. simultaneous consistent reaction of the displays to certain events. Another approach is a kind of “static” linking, where two or more displays apply the same visual expressive means to the pieces of information that need to be compared and/or related. Such displays may be controlled through common display manipulation tools. This intensifies the effect of the static visual linking. Different approaches to tool combination can be used together.

We have suggested the following taxonomy of mechanisms for linking concurrently running tools:

- Coordination on the basis of *selection of a subset of the references*. Several displays take special measures to show a selected subset of references (resulting from a querying or focusing tool, for example) and the corresponding characteristics prominently, so that the user can readily see the information relevant to the selected subset in each display. Various methods may be used to achieve this:

- *Highlighting* (special marking, e.g. by changing the colour or increasing the size) of the display items corresponding to the selected references so that they can be easily discerned from the remaining items.
  - *Filtering*, i.e. removing the display items that do not correspond to the selected references or “muting” their visual appearance and restricting user interaction with them.
  - *Focusing or zooming*. A display is adjusted so that the information relevant to the selected subset is shown with the maximum possible expressiveness and legibility at the cost of the rest of the information being omitted or reduced in its conspicuousness.
- Coordination on the basis of a *division of the reference set*. The reference set is divided into two or more non-overlapping subsets (for example, using a classification tool), and, in response, several displays take special measures to make the information relevant to each subset easily recognisable and distinguishable from the information related to the other subsets. This can be achieved in various ways:
    - *Multicolour marking*. Each subset receives a unique earmark (typically a colour), which is used for marking the display elements corresponding to this subset in all coordinated displays.
    - *Display multiplication*. A display is replaced or supplemented by several displays of the same type so that each display represents one of the subsets and the corresponding characteristics.
    - *Rearrangement of display items*. Display items corresponding to the same subset are put close to each other within the display space.
  - Display linking on the basis of a *common set of attributes*. The values and value combinations of the attributes may be represented in different ways, but special techniques help the user to discern the visual elements corresponding to any particular attribute in all the displays. Examples of such techniques are:
    - *Identical arrangement* of display items corresponding to different attributes (such as the order of the axes in a parallel-coordinates display or the order of the bars in a bar chart map).
    - *Colour marking* of the display items (such as bars in bar charts or sectors in pie charts) so that a unique colour corresponds to each attribute. The marking must be consistent between the displays.
  - Linking on the basis of a *common visual encoding* of attribute values in several displays, such as:

- *Common scales* along the display dimensions.
- *Common meanings* of colours, sizes, symbols, etc.

When the encoding is changed by means of display manipulation tools, the changes must affect all the linked displays.

- Linking on the basis of a *common transformation of attribute values* when the same transformation function (tool) is applied to values of attributes represented in different displays (the attributes in the displays may also be different, or the same attribute(s) may be represented in different ways). When the user changes any transformation parameters, the changes affect all the linked displays.

Besides these general methods of tool linking, there are also methods specific to particular display types.

Sequential and concurrent modes of tool combination are often used together. Moreover, different mechanisms of sequential and concurrent tool combination can be used simultaneously, for example filtering can be used together with multicolour marking and common transformation of attribute values. In any tool combination, at least one visual display needs to be present so that the user can perceive the results of the operation of the tools.

As a final note, we would like to mention that tool combination and coordination is currently a hot topic in the research areas related to exploratory data analysis, such as information visualisation, geographic visualisation, and statistical graphics. Dedicated international conferences have been convened (see (CMV 2003) and (CMV 2004)) and special journal issues published (InfoVis 2003). The papers describing various specific cases of the combined use of multiple exploratory tools are innumerable. A few papers of a more general kind could be recommended to interested readers, specifically Buja et al. (1991), North and Shneiderman (1997), and Roberts (1998). For technically oriented readers, we can also recommend some papers that suggest models and software architectures for tool combination, for example North and Shneiderman (1999), North et al. (2002), and Boukhelifa and Rodgers (2003).

## 4.9 Exploratory Tools and Technological Progress

All the tools discussed in this chapter (perhaps with the exception of the visualisation tools) rely significantly upon modern computer technology. They did not exist and could not have existed at the time when John Tukey wrote his pioneering work, which launched the term “exploratory data analysis” (Tukey 1977). Technological development has and will continue

to have a strong influence on the variety, capabilities and characteristics of the exploratory tools that are being designed and implemented. Computers are increasing in capacity – this means that more extensive, more comprehensive, and more detailed datasets can be stored and processed. Computers are increasing in speed – this means that data transformation, querying, computational analysis, and other data processing can be quick and responsive to any modification of data or parameters. Computers are increasing in sophistication and are developing new capabilities – this means that they are becoming able to release human analysts from various routine operations and thereby make more of their time available for imaginative perception and creative thinking. Computers are increasing in user-friendliness – this means that analysts may benefit from new, more convenient, and more effective possibilities for interaction. Computers are increasing in portability – this means that data analysis can be done whenever and wherever needed, possibly in tight combination with data collection and observation of real things and processes.

However, despite great progress, computers cannot (yet?) substitute for human analysts who possess capabilities to link, comprehend, generalise, and abstract, which are indispensable for any exploration. Computers can act as technical assistants that provide the results of their work to higher-level staff for summary, synoptic processing, and the drawing of conclusions or implications. The results provided to an analyst must have such a form that the analyst can use them in his/her further work. Since this work is done mostly in the analyst's mind, the results must be presented in a form perceptible by the mind, that is, they must be *visualised*.

We have emphasised many times the prominent role of visualisation in exploratory data analysis. We started this chapter with a hymn to visualisation and are expressing our appreciation of it again at the end. Visualisation is involved in every example given in this chapter, be it an example of data transformation, querying, or computational analysis. Visualisation provides food to our brain in a form that can be digested much better than numbers or even words (recall the Chinese proverb that a picture is worth a thousand words).

So, what does technological development mean for visualisation? The benefits are quite numerous:

- Quick display generation, which allows an analyst to make displays on demand and discard them when they are no longer needed, and to have purpose-oriented displays that show only relevant information, rather than overloaded multipurpose presentations.

- Modifiability and interactivity of displays; this enables dynamic presentation (e.g. of time-referenced data), display manipulation, and display coordination.
- High resolution (including colour resolution), which allows representation of larger data volumes and finer detail.
- The possibility to use a three-dimensional display space for the representation of data with a complex structure, for example spatio-temporal data or other multidimensional data.
- The possibility to combine visual displays with computation, querying, and data transformations, which enhances the analytical potential of visualisation and allows genuine tool synergism to be achieved.

It may be noted that we have not included in this list such things as highly realistic images or immersive environments. Although we admit that these technologies may be very useful in some application domains (for example, city or landscape planning), we have doubts about their utility in exploratory data analysis. The basic problem is the high realism of the representation: the things look so concrete that the process of abstraction, which is necessary in exploration, is inhibited, and the mind is confused by the multitude of realistic details.

We would like to refer again to the book by Rudolf Arnheim that we adore so much. He says that highly realistic images do not, by themselves, guide understanding:

Paradoxically, they may even make identification difficult, because to identify an object means to recognize some of its salient structural features. A mechanically produced replica may hide or distort these features. One of the reasons why persons brought up in cultures that are unacquainted with photography have trouble with our snapshots is that the realistic and accidental detail and partial shapelessness of such images do not help perception. (Arnheim 1997, p. 140)

So, technological progress provides new, exciting opportunities, but the designers and developers of tools for EDA need to think carefully about how these opportunities can be better utilised. There is yet another implication of technological progress for exploratory data analysis: besides new opportunities, it also creates new challenges. The challenges arise because more and more data are being collected and stored, and hence need to be analysed and understood. Not only volumes but also the complexity of data is increasing as more and more aspects and components of various things and phenomena are being sensed and measured. The variety of types of data is also increasing: modern analysts need to deal with satellite images and DNA structures, behaviours of individuals (humans or animals) and social groups, sales dynamics and historical events, and many

other things. Heterogeneous data types need to be analysed together. All this calls for new analytical tools.

We shall not try to predict what future analytical tools will look like. However, we are sure that the key feature of their further development will be a synergy of techniques and approaches, in which visualisation will continue to play a leading role. Researchers in data-analysis-related areas and tool developers should look (but look critically!) for new opportunities offered by technological progress and, at the same time, be ready to face new challenges.

## Summary

In this very long chapter, we have reviewed the existing tools that can be used to support exploratory data analysis. We have tried to keep a general level of discussion and consider the tools as “pure ideas”, without regard to the details of specific implementations, although we had to use specific software to produce the illustrations. While we have not achieved an absolute “purification”, we consider the resulting level of generality to be quite satisfactory.

We have considered several major classes of exploratory tools according to their primary functions:

- visualisation;
- display manipulation;
- data manipulation;
- querying;
- computational analysis.

We have tried to give a systematic review of each class, introducing, when possible, intra-class taxonomies. We have demonstrated with numerous examples how various types of tool can be used in data analysis. The examples also demonstrate that any non-visual tool needs to be used together with visualisation, and show how various types of tool outputs can be visualised.

Not only the combination of non-visual tools with tools that visualise their results is important in exploratory data analysis, but also the use of multiple tools in various combinations. On the one hand, different tools have different capabilities and therefore can aptly complement each other and jointly produce synergistic effects. On the other hand, data are often very abundant and/or very complex, multidimensional and multifaceted, and therefore cannot be adequately analysed using any single tool.

We have considered two basic modes of tool combination, sequential and concurrent, and enumerated the general mechanisms applied for tool combination. Like the tools themselves, these mechanisms can also be combined. Modern software packages for EDA typically provide a variety of tools and of methods for linking them.

Generally, we did not intend in this chapter to relate tools to the types of analysis tasks that they could support, although we did this from time to time, especially when discussing the examples. Now it is time to do this in a more systematic way. This will be the content of the next chapter.

## References

- (Ahlberg et al. 1992) Ahlberg, C., Williamson, C., Shneiderman, B.: Dynamic queries for information exploration: an implementation and evaluation. In: *Proceedings of ACM CHI'92* (ACM Press, New York 1992) pp. 619–626
- (Allen 1983) Allen, J.F.: Maintaining knowledge about temporal intervals. *Communications of the ACM* **26**(11), 123–154 (1983)
- (Andrienko and Andrienko 1999) Andrienko, G., Andrienko, N.: Interactive maps for visual data exploration. *International Journal of Geographical Information Science* **13**(4), 355–374 (1999)
- (Andrienko and Andrienko 2001) Andrienko, G., Andrienko, N.: Constructing parallel coordinates plot for problem solving. In: *1st International Symposium on Smart Graphics*, ed. by Butz, A., Krüger, A., Oliver, P., Zhou, M., Hawthorne NY, March 2001 (ACM Press, New York 2001) pp. 9–14
- (Andrienko and Andrienko 2004) Andrienko, N., Andrienko, G.: Cumulative curves for exploration of demographic data: a case study of northwest england. *Computational Statistics* **19**(1), 9–28 (2004)
- (Andrienko et al. 2001) Andrienko, G., Andrienko, N., Savinov, A.: Choropleth maps: classification revisited. In: *Proceedings of ICA 2001*, Beijing, vol. 2, pp. 1209–1219 (2001)
- (Arnheim 1997) Arnheim, R.: *Visual Thinking* (University of California Press, Berkeley 1969, renewed 1997)
- (Bertin 1967/1983) Bertin, J.: *Semiology of Graphics. Diagrams, Networks, Maps* (University of Wisconsin Press, Madison 1983). Translated from Bertin, J.: *Sémiologie graphique* (Gauthier-Villars, Paris 1967)
- (Blok et al. 1999) Blok, C., Koebben, B., Cheng, T., Kuterema, A.A.: Visualization of relationships between spatial patterns in time by cartographic animation. *Cartography and Geographic Information Science* **26**(2), 139–151 (1999)
- (Boukhelifa and Rodgers 2003) Boukhelifa, N., Rodgers, P.J.: A model and software system for coordinated and multiple views in exploratory visualization, *Information Visualization*, **2**(4), 258–269 (2003)

- (Brewer 1994) Brewer, C.A.: Color use guidelines for mapping and visualization. In: *Visualization in Modern Cartography*, ed. by MacEachren, A.M., Fraser Taylor, D.R. (Elsevier, New York 1994) pp 123–147
- (Buja et al. 1991) Buja, A., McDonald, J.A., Michalak, J., Stuetzle, W.: Interactive data visualization using focusing and linking. In: *Proceedings of IEEE Visualization '91* (IEEE Computer Society Press, Washington 1991) pp. 156–163
- (Burt and Barber 1996) Burt, J.E., Barber, G.M.: *Elementary Statistics for Geographers*, 2nd edn (Guilford, New York 1996)
- (Carr et al. 1992) Carr, D.B., Olsen, A.R., White, D.: Hexagon mosaic maps for display of univariate and bivariate geographical data. *Cartography and Geographic Information Systems* **19**(4), 228–236 (1992)
- (Carr et al. 2000) Carr, D.B., Wallin, J.F., Carr, D.A.: Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. *Statistics in Medicine* **19**, 2521–2538 (2000)
- (Carr et al. 2002) Carr, D.B., Zhang, Y., Li, Y.: Dynamically conditioned choropleth maps: shareware for hypothesis generation and education. *Statistical Computing & Statistical Graphics Newsletter* **13**(2), 2–7 (2002)
- (Casner 1991) Casner, S.M.: A task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics* **10**, 111–151 (1991)
- (Catarci et al. 1997) Catarci, T., Costabile, M.F., Levialdi, S., Batini, C.: Visual query systems for databases: a survey. *Journal of Visual Languages and Computing* **8**(2), 215–260 (1997)
- (Chen 2003) Chen, H.: Compound brushing. In: *IEEE Symposium on Information Visualization*, Seattle, October 2003, ed. by Munzner, T., North, S. (IEEE Computer Society Press, Washington 2003) pp. 181–188
- (Cleveland and McGill 1984) Cleveland W.S., McGill, R.: Graphical perception: theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* **79**(387), 531–554 (1984)
- (Cleveland and McGill 1986) Cleveland W.S., McGill, R.: An experiment in graphical perception. *International Journal of Man–Machine Studies* **25**(5), 491–500 (1986)
- (CMV 2003) Roberts, J.C. (ed.): *Proceedings of the First International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'03)*, July 2003, London (IEEE Computer Society, Los Alamitos 2003)
- (CMV 2004) Roberts, J.C. (ed.): *Proceedings of the Second International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'04)*, July 2004, London (IEEE Computer Society, Los Alamitos 2004)
- (Chrisman 1997) Chrisman, N.: *Exploring Geographic Information Systems* (Wiley, New York 1997)
- (Cressie 1991) Cressie, N.A.C.: *Statistics for Spatial Data* (Wiley, New York 1991)
- (Dallal 1999) Dallal, G.E.: *The Little Handbook of Statistical Practice*. <http://www.StatisticalPractice.com>. Accessed 28 Mar 2005

- (Dodge 2000) Dodge, M.: *NewsMaps: Topographic Mapping of information*, (2000), [http://mappa.mundi.net/maps/maps\\_015/](http://mappa.mundi.net/maps/maps_015/). Accessed 28 Mar 2005
- (Dorling 1992) Dorling, D.: Visualising people in time and space. Environment and Planning B: Planning and Design **19**, 613–647 (1992)
- (Edsall and Peuquet 1997) Edsall, R., Peuquet, D.: A graphical user interface for the integration of time into GIS. In: *Proceedings of the 1997 American Congress of Surveying and Mapping Annual Convention and Exhibition*, Seattle (1997) pp. 182–189
- (Egbert and Slocum 1992) Egbert, S.L., Slocum, T.A.: EXPLOREMAP: an exploration system for choropleth maps. Annals of the Association of American Geographers **82**, 275–288 (1992)
- (Fayyad et al. 2002) Fayyad, U., Grinstein, G.G., Wierse, A. (eds): *Information Visualisation in Data Mining and Knowledge Discovery* (Morgan Kaufmann, San Francisco 2002)
- (Fotheringham and Rogerson 1994) Fotheringham S., Rogerson P. (eds): *Spatial Analysis and GIS* (Taylor & Francis, London 1994)
- (Friendly 1994) Friendly, M.: Mosaic displays for multi-way contingency tables. Journal of the American Statistical Association **89**, 190–200 (1994)
- (Friendly 2002) Friendly, M.: Corgrams: exploratory displays for correlation matrices. American Statistician **56**(4), 316–325 (2002)
- (Friendly and Kwan 2003) Friendly, M., Kwan, E.: Effect ordering for data displays, Computational Statistics & Data Analysis **43**, 509–539 (2003)
- (Furnas 1986) Furnas, G.W.: Generalized fisheye views. In: *Proceedings of CHI'86* (ACM, New York 1986) pp. 16–23
- (Green 1998) Green, M.: *Toward a perceptual science of multidimensional data Visualisation: Bertin and Beyond* (1998), <http://www.ergogero.com/dataviz/dvis0.html>. Accessed 28 Mar 2005
- (Harrower et al. 1999) Harrower, M., Griffin, A.L., MacEachren, A.M.: Temporal focusing and temporal brushing: assessing their impact in geographic visualization. In: *Proceedings of the 19th International Cartographic Conference*, Vol. 1 (1999) pp. 729–738
- (Hernandez et al. 2005) Hernandez, V., Göring, W., Voß, A., Hopmann, C.: Sustainable decision support by the use of multi-level and multi-criteria spatial analysis on the Nicaragua development gateway. In: *8th International Conference on Global Spatial Data Infrastructure, GSDI-8*, Cairo, April 2005
- (Hochheiser and Shneiderman 2004) Hochheiser, H., Shneiderman, B.: Dynamic query tools for time series data sets: timebox widgets for interactive exploration. Information Visualization **3**(1), 1–18 (2004)
- (InfoVis 2003) Roberts, J.C. (ed.): *Special Issue on Coordinated and Multiple Views in Exploratory Visualization*. Information Visualization **2**(4), (2003)
- (Jenks 1977) Jenks, G.F.: *Optimal data classification for choropleth maps*, Occasional Paper No. 2 (Department of Geography, University of Kansas 1977)
- (Keim and Kriegel 1994) Keim D., Kriegel, H.-P.: VisDB: database exploration using multidimensional visualization. IEEE Computer Graphics and Applications **14**(5), 40–49 (1994)

- (Keogh and Kasetty 2003) Keogh, E.J., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery* 7(4), 349–371 (2003)
- (Klösgen and Żytkow 2002) Klösgen, W., Żytkow, J.M. (eds.): *Handbook of Data Mining and Knowledge Discovery*, (Oxford University Press, New York 2002)
- (Kosslyn 1994) Kosslyn, S.M.: *Elements of Graph Design* (Freeman, New York 1994)
- (Leung and Apperley 1994) Leung, Y.K., Apperley, M.D.: A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction* 1(2), 126–160 (1994)
- (Li and North 2003) Li, Q., North, C.: Empirical comparison of dynamic query sliders and brushing histograms. In: *Proceeding of IEEE Information Visualization 2003*, Seattle (2003)
- (Lyutyy 1986) Lyutyy, A.A.: *The Language of Map: Essence, System, Functions* (Institute of Geography of the Russian Academy of Sciences, Moscow 1986) (in Russian)
- (MacEachren 1994) MacEachren, A.M.: *Some Truth with Maps: a Primer on Symbolization and Design* (Association of American Geographers, Washington, DC 1994)
- (MacEachren 1995) MacEachren, A.M.: *How Maps Work: Representation, Visualization, and Design* (Guilford, New York 1995)
- (Mackinlay 1986) Mackinlay, J.: Automating the design of graphical presentation of relational information. *ACM Transactions on Graphics* 5(2), 110–141 (1986)
- (Miller and Han 2001) Miller, H.J., Han, J.: Geographic data mining and knowledge discovery: an overview. In: *Geographic Data Mining and Knowledge Discovery*, ed. by Miller, H.J., Han, J. (Taylor & Francis, London 2001) pp. 3–32
- (Monmonier 1990) Monmonier, M.: Strategies for the visualization of geographic time-series data. *Cartographica* 27(1), 30–45 (1990)
- (Newton 1978) Newton, C.M.: Graphics: from alpha to omega in data analysis. In: *Graphical Representation of Multivariate Data*, ed. by Wang, P.C.C. (Academic Press, New York 1978) pp. 59–92
- (NIST/SEMATECH 2005) NIST/SEMATECH e-*Handbook of Statistical Methods. Chapter 1: Exploratory Data Analysis*, <http://www.itl.nist.gov/div898/handbook/>. Accessed 29 Mar 2005
- (Norman et al. 2003) Norman, K.L., Zhao, H., Shneiderman, B., Golub, E.: Dynamic query choropleth maps for information seeking and decision making. In: *Proceedings of Human-Computer Interaction International 2003. Vol. 2: Theory and Practice* (Lawrence Erlbaum Associates, 2003) pp. 1263–1267
- (North and Shneiderman 1997) North, C., Shneiderman, B.: *A Taxonomy of Multiple-Window Coordinations*, Technical Report CS-TR-3854 (University of Maryland Computer Science Department, College Park 1997)

- (North and Schneiderman 1999) North, C., Schneiderman, B.: *Snap-Together Visualization: Coordinating Multiple Views to Explore Information*, Technical Report CS-TR-4020 (University of Maryland Computer Science Department, College Park 1999)
- (North et al. 2002) North, C., Conklin, N., Indukuri, K., Saini, V.: Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases. *Information Visualization* **1**(3–4), 211–228 (2002)
- (Peuquet 2002) Peuquet, D.J.: *Representations of Space and Time* (Guilford, New York 2002)
- (Rana and Dykes 2003) Rana, S., Dykes, J.: A framework for augmenting the visualization of dynamic raster surfaces. *Information Visualization* **2**, 126–139 (2003)
- (Random House 1996) *Random House Webster's Unabridged Electronic Dictionary* (Random House, Broadway, NY 1996)
- (Rhodes 2002) Rhodes, P.J.: Discovering New Relationships: A brief overview of data mining and knowledge discovery. In: *Information Visualisation in Data Mining and Knowledge Discovery*, ed. by Fayyad, U., Grinstein, G.G., Wierse, A. (Morgan Kaufmann, San Francisco 2002)
- (Roberts 1998) Roberts, J.C.: On encouraging multiple views for visualisation. In: *Information Visualisation IV'98*, ed by Banissi, E., Khosrowshahi, F., Safraz, M., July 1998 (IEEE Computer Society Press, Washington 1998) pp. 8–14
- (Roth and Mattis 1990) Roth, S.M., Mattis, J.: Data characterization for intelligent graphics presentation. In: *Proc. SIGCHI'90: Human Factors in Computing Systems*, Seattle, 1990 (ACM Press, New York 1990) pp. 193–200
- (Rousseeuw et al. 1999) Rousseeuw, P.J., Ruts, I., Tukey, J.W.: The Bagplot: a bivariate boxplot. *The American Statistician* **53**(4), 382–387 (1999)
- (Sadahiro 2002) Sadahiro, Y.: A graphical method for exploring spatiotemporal point distributions. *Cartography and Geographic Information Science* **29**(2), 67–84 (2002)
- (Salichtchev 1982) Salichtchev, K.A.: *Cartography: a Textbook for Geographical Specialities of Universities*, 3rd edn (Vysshaya Shkola, Moscow 1982) (in Russian)
- (Senay and Ignatius 1994) Senay, H., Ignatius, E.: A knowledge-based system for visualization design. *IEEE Computer Graphics and Applications* **14**(6), 36–47 (1994)
- (Shekhar and Chawla 2003) Shekhar, S., Chawla, A.: *Spatial Databases: a Tour* (Pearson Education, Upper Saddle River 2003)
- (Shneiderman 1992) Shneiderman, B.: Tree visualization with treemaps: a 2-D space-filling approach. *ACM Transactions on Graphics* **11**(1), 92–99 (1992)
- (Slocum 1999) Slocum, T.A.: *Thematic Cartography and Visualization* (Prentice Hall, Upper Saddle River 1999)
- (Spence 2001) Spence, R.: *Information Visualisation* (Addison-Wesley, Harlow 2001)

- (Spence and Tweedy 1998) Spence, R., Tweedy, L.: The Attribute Explorer: information synthesis via exploration. *Interacting with Computers* **11**, 137–146 (1998)
- (Spoerri 1999) Spoerri, A.: InfoCrystal: a visual tool for information retrieval. In: *Readings in Information Visualization: Using Vision to Think*, ed. by Card, S.K., Mackinlay, J.D., Shneiderman, B. (Morgan Kaufmann, San Francisco 1999) pp 140–147
- (StatSoft 2004) StatSoft, Inc.: *Electronic Statistics Textbook* (StatSoft, Tulsa 2004), <http://www.statsoft.com/textbook/stathome.html>. Accessed 28 Mar 2005
- (Stolte et al. 2002) Stolte, C., Tang, D., Hanrahan, P.: Multiscale visualization using data cubes. In: *Proceedings of the IEEE Symposium on Information Visualization 2002, InfoVis '02*, Boston, USA, October 2002, ed. by Wong, P.C., Andrews, K. (IEEE Computer Society, Piscataway 2002) pp. 7–14
- (Tufte 1983) Tufte, E.R.: *The Visual Display of Quantitative Information* (Graphics Press, Cheshire CT, 1983)
- (Tufte 1990) Tufte, E.R.: *Envisioning Information* (Graphics Press, Cheshire, CT 1990)
- (Tukey 1977) Tukey, J.W.: *Exploratory Data Analysis* (Addison-Wesley, Reading, MA 1977)
- (Unwin and Hofmann 1998) Unwin, A.R., Hofmann, H.: New interactive graphics tools for exploratory analysis of spatial data. In: *Innovations in GIS*, Vol. 5, ed. by Carver, S. (Taylor & Francis, London 1998) pp. 46–55
- (Wattenberg 2001) Wattenberg, M.: Sketching a graph to query a time-series database. In: *Extended Abstracts of CHI '01*, Seattle, March–April 2001, (ACM Press, New York 2001) pp. 379–380
- (Wilkinson 1999) Wilkinson, L.: *The Grammar of Graphics* (Springer-Verlag, New York 1999)
- (Wise et al. 1995) Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information from text documents. In: *Proceedings of IEEE 1995 Symposium on Information Visualization*, Atlanta (1995) pp. 51–58
- (Witten and Frank 1999) Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (Morgan Kaufmann, San Francisco 1999)
- (Zenkin 1990) Zenkin, A.A.: Waring's problem from the standpoint of the cognitive interactive computer graphics. *Mathematical and Computer Modelling* **13**(11), 9–37 (1990)

# 5 Principles

## Abstract

In this chapter, we describe the major principles that are used in exploring data and in choosing tools for this purpose. We have extracted these principles from our experience, by inspecting our usual approaches and choices when we receive new data that need to be analysed. However, these principles correspond very well to ideas expressed by other researchers in the areas of visualisation, data analysis, systems analysis, and cognitive psychology. This certifies the principles as generic, relevant not only to our particular way of handling data but also to some fundamental processes involved in exploration, reasoning, and understanding.

To show where the principles come from, we present a view of the process of data exploration as a combination of top-down and bottom-up procedures, i.e. analysis and synthesis. At the beginning, an explorer has the most general task: to characterise and explain the overall behaviour of the characteristics over the entire reference set. In the course of the exploration, this general task is decomposed into subtasks of various types. We illustrate this view by several examples, in which datasets with different structures are considered and the exploration procedures outlined. These examples demonstrate that the major instrument of exploratory analysis is the human mind, equipped with appropriate visual displays of the data, which provide an object for the explorer's observations and food for his/her thought.

The great role of visualisation is also pronounced in our presentation of the principles. We introduce ten general principles of EDA:

1. *See the whole.* Represent the data so that the overall behaviour can be perceived by means of vision. This requires that first, nothing essential is omitted (ideally, all data items are present in the display); second, all aspects are reflected; and third, the visual elements representing the data can be perceived all at once as a unified whole.
2. *Simplify and abstract.* Disregard excessive detail, fluctuations, and occasional peculiarities, which obstruct one from seeing the essential features of the behaviour.

3. *Divide and group.* When it is seen or expected that the overall behaviour is not the same throughout the reference set, divide the reference set so that the behaviour within each subset can be regarded as sufficiently homogeneous. Then, the overall behaviour can be characterised as a combination of the partial behaviours.
4. *See in relation.* For a proper characterisation of a behaviour divided into parts, reveal the substantial differences as well as the similarities between the parts. It is also important to compare the behaviours of different attributes or groups of attributes.
5. *Look for recognisable.* Represent the data so that specific sorts of features, or subpatterns, can be easily detected. The features to look for depend on the structure and nature of the data.
6. *Zoom and focus.* In exploring partial behaviours over reference subsets, apply tools that help in concentrating on the part currently being analysed and in representing this part with the maximum possible expressiveness. However, it is important to position this part with respect to the entire behaviour, i.e. to see it in context.
7. *Attend to particulars.* Detect, thoroughly examine, and try to explain various cases of unusual characteristics.
8. *Establish linkages.* Integrate the observations and partial patterns derived from the investigation of various parts and aspects of the overall behaviour into a coherent view.
9. *Establish structure.* When the overall behaviour is suspected to result from an interplay of several structural components, such as linear and cyclic processes in time-related phenomena, explore each component and its interactions with the other(s) by splitting relevant referrers into several referrers or introducing additional referrers.
10. *Involve domain knowledge.* Whenever possible, make use of what you know concerning the nature and properties of the phenomenon underlying the data, or even make use of your common sense. This may take the form of anticipation of general tendencies in the behaviour, of substantial distinctions between certain parts of the data, of the sort of features (subpatterns) that can occur, etc. Such anticipations influence the choice of tools and their parameters and the focus of attention.

In our presentation of these principles, we indicate what types of exploratory tasks they are relevant to and what categories of tools can support their implementation. By using examples of various kinds of data, we demonstrate how the principles can be implemented and what can be gained from this.

At the end of the chapter, we put all the principles into the overall context of exploratory data analysis, viewed as the systematic decomposition

and performance of the top-level task “characterise the behaviour of the characteristics over the reference set”. We consider four general cases of analysis, depending on the peculiarities of the data:

1. The basic case: A one-dimensional reference set and a single attribute or a group of attributes that can be visualised in a way that supports perceptual unification.
2. Multidimensional data, i.e. data with multiple referential components.
3. Multiple attributes that need to be analysed jointly.
4. Data characterised by a large volume, i.e. a large size of the reference set.

Each successive case refers to the cases previously described, as the original task is decomposed and turned into a sequence of simpler operations dealing with subsets and slices of the data.

The cases are summarised compactly in the form of tables, which list the actions performed and specify the types of exploratory subtasks involved, the appropriate tool categories, and the relevant principles. We regard this as a summary of the major results of our study. We indicate the ways in which these results may be used by data explorers and by designers and developers of instruments for EDA. We also present an example of an application of the suggested generic scheme of data analysis to the exploration of a particular dataset.

## 5.1 Motivation

The general goal of our study is to understand the nature of the tasks arising in exploratory data analysis and their influence on the choice of tools and the ways in which these tools are used. On this basis, we would like to formulate guidelines that could help data explorers to choose the right tools, as well as help tool designers and developers to anticipate and satisfy the demands of explorers. In brief, we would like to relate the tools to the tasks, or, more precisely, find the principles by which we can relate tools to tasks.

In Chap. 3, we identified the types of tasks that exist, and in Chap. 4 we described and classified the tools available. So, why don’t we simply cross-reference these two lists, i.e. specify for each type of task which tool(s) can support it, and for each tool which task(s) it supports? Why do we want instead to find principles?

The fundamental reason is that the tasks arising in data exploration are too specific (they are always formulated in the terms of data components),

whereas the task categories that we identified are too generic. It is impossible to link each specific task to the appropriate tool(s) because the specific tasks are countless. Linking the tools to the generic task categories is also problematic, but for a different reason: the categories are so generic that no tool can perform all tasks belonging to the same category.

Let us take, for example, the task category “behaviour characterisation”. Depending on the nature and dimensionality of the reference set and on the number and properties of the attributes, there may be numerous possible types of behaviours. The spatial variation of the amount of forest and its structure, the dynamics of the crime rate in a country, the movements of storks or vehicles, the spatial and temporal distribution of earthquakes – these are just a few examples of possible behaviour types. It is clear that one cannot find or design a tool that would be equally appropriate for characterising any of these behaviours. Hence, if we decided to identify the tools capable of supporting the task category “behaviour characterisation”, we would need to relate each tool to the specific type(s) of behaviour that it is suitable for. We would also need to enumerate all possible types of behaviours in order to relate each type to the appropriate tool.

Even on the elementary level, the task “On a given date, what is the price of stock X?” is different from the task “What was the population of Loures in 1981?”, even though both tasks are classified as direct lookup tasks (see Table 3.5 in Sect. 3.4.8). In principle, the situation with elementary tasks is easier: there are general query tools capable of answering a wide range of elementary questions. However, as we discussed in the previous chapter, to get an answer to a question from a query tool, one needs to formulate the question in a form understandable by the tool. Unfortunately, all general query languages, although they allow one to formulate almost any question and thereby give full access to the power and flexibility of general query tools, are difficult to learn and inconvenient to use. Specific query tools, applicable to certain types of data and restrictive as to the range of possible questions, may be much more helpful for data exploration, at the cost of ease of use and dynamic response (see Sect. 4.6, especially the discussion of dynamic query tools in Sect. 4.6.1.3).

So, it seems that relating tools to generic task categories is either unfeasible or unhelpful. For an appropriate association, it is necessary that the structure and properties of the data to be analysed are taken into account, but the generic categories are too abstract for this, they stand too far away from the data. Hence, the association has to be done on a much more specific level, i.e. either for a specific dataset or for a class of datasets with a common structure and common properties. The first possibility is more appropriate for an explorer, and the second for a tool designer. However, to design an appropriate tool or tool combination for a class of datasets, the

tool designer needs to have (or at least have in mind) some specific examples of datasets from this class.

Although the level of dataset classes is more general than the level of specific datasets, this does not significantly help us in achieving our goals: the classes are still too numerous to be comprehensively described and linked to the right tools. The classes differ according to their data structure, i.e. the number and types of referrers and attributes, and the properties of the components, such as ordering, the existence of distances, continuity, smoothness, etc. There are so many different combinations of numbers, types, and properties that it is unfeasible to consider all of them.

How, then, can we help explorers and tool designers? We can try to set out some general principles for choosing and designing exploratory tools. We can also demonstrate, with various examples, how these principles can be applied in practice. Then, we can try to describe, on a very general level, the overall procedure of exploratory data analysis: we consider the major cases, specify the steps and the possible options, and refer to the appropriate principles. We believe that this should provide reasonably good guidance. One may pick the most suitable case, look through the suggested steps of analysis, note the recommended tool categories, and then, by applying the corresponding general principles, try to choose or devise particular tools and approaches suited to the data at hand.

We shall start our search for the general principles from an attempt to uncover some generic components and features of the process of exploratory data analysis

## 5.2 Components of the Exploratory Process

At the end of the Chap. 3, we characterised the process of data exploration as an interplay of two major subprocesses, top-down and bottom-up, or analysis and synthesis. At the beginning, the explorer has a very general goal: to grasp the distinctive features of a dataset and the underlying phenomenon, that is, to build a compact representation of this data/phenomenon in his/her mind (a mental model) reflecting these distinctive features. In terms of our task typology, this is a synoptic task of behaviour characterisation, or pattern definition.<sup>28</sup> Furthermore, the explorer may also have the goal of finding the reason for the existence of these par-

<sup>28</sup> As a reminder, we use the word “pattern” to denote a compact representation or description of a behaviour, be it an internal representation in the explorer’s mind (i.e. a mental model), a description in some language, a formula, or a drawing.

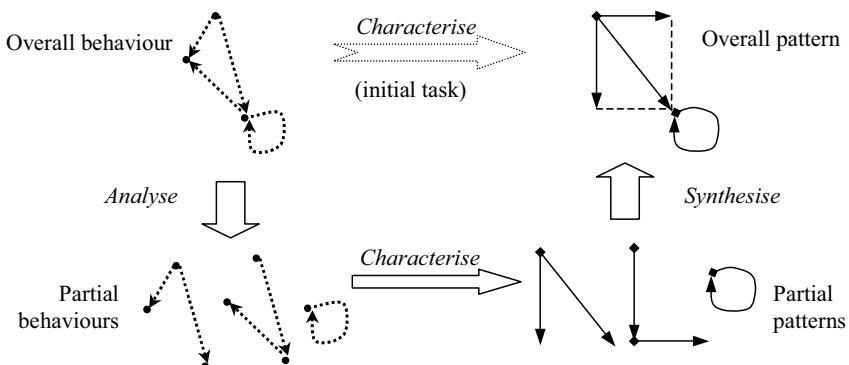
ticular features, i.e. of explaining the features as resulting from structural or causal links between inherent parts of the phenomenon and/or between this phenomenon and other phenomena in its environment. According to our framework, this is a synoptic task of connection discovery. The expected result is some representation of the essential internal and/or external links, which is called a connection pattern.

Typically, such a general goal cannot be achieved immediately, and the explorer has to decompose the initial task into smaller and more easily accomplishable subtasks. For this purpose, the explorer divides the overall behaviour into parts that will be easier to study and describe. For example, the explorer may consider individual attributes or small groups of attributes rather than all attributes simultaneously; he/she may divide the reference set into subsets and characterise the behaviour on each subset; or, if the dataset is multidimensional, he/she may look at various slices and projections. Depending on the complexity of the data, this process of decomposition, or analysis, may go down further. As a result of this process, the explorer derives a number of fragmentary patterns representing certain parts or aspects of the overall behaviour. The explorer cannot be satisfied with these fragmentary patterns; he/she needs to integrate them into a complete descriptive and/or connectional pattern for the overall behaviour. This process of integration, or synthesis, is another intrinsic part of data exploration, which complements the process of analysis.

It should not be thought that the process of synthesis always starts only after the process of analysis is fully accomplished. The explorer may switch from analytic to synthetic and from synthetic to analytic activities many times throughout the process of exploration. For example, the explorer may characterise a partial behaviour and immediately try to establish its position with respect to the overall behaviour or relate it to another partial behaviour characterised earlier. At this stage, some partial behaviours may not yet have been characterised, but this is not an obstacle to the integration of the fragments of the general model that have already been built.

Regardless of the actual sequence of the analytic and synthetic activities, we can represent the process of data exploration by the abstract scheme drawn in Fig. 5.1. The initial task, i.e. to characterise the overall behaviour by an appropriate pattern, is accomplished by means of three major groups of activities:

1. *Analyse*, i.e. divide the overall behaviour into partial behaviours.
2. *Characterise*, i.e. derive (partial) patterns in order to approximate (partial) behaviours.
3. *Synthesise*, i.e. integrate the partial patterns into an overall pattern.



**Fig. 5.1.** A schematic representation of the process of data exploration as a combination of three major groups of activities: analyse (divide the overall behaviour into partial behaviours), characterise (derive patterns to approximate behaviours), and synthesise (integrate the partial patterns into an overall pattern)

How does this scheme relate to our task typology? It is clear that the activities labelled as “characterise” correspond to the synoptic tasks of behaviour characterisation. Note that “characterise” appears on two levels in the scheme: on the level of the overall behaviour and on the level of the partial behaviours. This means that the initial task of characterisation of the overall behaviour involves tasks of characterising the partial behaviours as its subtasks. However, subtasks of other types are also involved. To determine their positions with respect to the scheme, let us consider some examples.

### 5.3 Some Examples of Exploration

Suppose that our task is to characterise the behaviour of the burglary rate in the state of California over the period from 1960 to 2000. Probably the best way to do this is by looking at a time graph, such as the one in Fig. 3.12, that represents this behaviour visually. As we mentioned earlier (see Sect. 3.8), we do not simply *look at* the time graph but *look for* something relevant to our task. In this case, certain types of patterns that can be expected in this type of data are relevant, specifically increases, decreases, stability, and fluctuations, which show up as particular shapes of the line. In terms of our task typology, we perform pattern search tasks. Since the line as a whole does not correspond to any single pattern type, we divide it into fragments interpretable as an increase, a decrease, etc. The presence of small fluctuations complicates the recognition of patterns. We need somehow to abstract from these fluctuations, to disregard them.

As we detect interpretable patterns, we also note the points where one pattern type turns into another one, or, in other words, where substantially dissimilar or opposite behaviours take place in contiguous time intervals. This corresponds to synoptic relation-seeking tasks in our task typology. Noting the turning points also involves pattern comparison tasks, since we are comparing the behaviours in the adjacent intervals.

Hence, in the process of dividing the overall behaviour of the burglary rate in California into parts, pattern search tasks play the primary role, and are accompanied by comparison and relation-seeking tasks. Another observation is that tasks of different types may occur in parallel; for example, we may recognise a line fragment as having an increasing trend and simultaneously note its difference from a neighbouring fragment.

During the pattern search and behaviour division process, we already start to characterise the partial behaviours extracted: we label them as “increase”, “decrease”, etc. Sometimes this is a sufficient characterisation of the partial behaviours. However, higher precision of the approximation of the behaviour is often desired. For a more specific and detailed characterisation, we need to determine the beginning and ending times of each partial behaviour, the attribute values at the beginning and at the end, the minimum and maximum values and when they were attained, etc. These are elementary lookup tasks, either direct or inverse. We can use the values found to compute certain summary characteristics of the behaviour, such as the average rate of increase or decrease. To reflect the fluctuations that exist in the behaviour characterisation, we can look for major differences between values at adjacent time moments and measure these differences, i.e. perform elementary relation-seeking and comparison tasks. However, as we said before, elementary tasks mostly play only a subordinate role in exploratory data analysis.

Now, when we have extracted the partial behaviours and characterised them, i.e. approximated them by suitable partial patterns, we need to integrate these partial patterns into an overall pattern approximating the overall behaviour. While a simple enumeration of the partial patterns may sometimes be sufficient, it is usually not the case. Synthetic activities imply that the partial patterns are appropriately linked, that a kind of order and/or structure is established among them. Ideally, an overall pattern should appear as a cogent, well-substantiated structure, with a clear position and role for each partial pattern. To achieve this ultimate goal, we need to accomplish connection discovery tasks, i.e. reveal essential relations between the partial behaviours. Such tasks are very complex; they require creative thinking and often insight, and success is never guaranteed. In many cases, the data available are simply insufficient for revealing essential relations.

What is easier but still worthwhile to do is to put the partial patterns in chronological order and compare the lengths of the time intervals corresponding to the patterns and other characteristics of the patterns, for example the rate of increase of the burglary rate in the first half of the time period and the rate of decrease in the second half, or the degrees of fluctuation in those intervals. These are pattern comparison tasks, both direct and inverse. By means of pattern comparison, it is sometimes even possible to discover a certain order in the arrangement of the partial patterns, such as periodicity or a recurrent appearance of one pattern type after another, i.e. pattern comparison may eventually lead to connection discovery (this does not apply, however, to the behaviour of the burglary rate in California).

In this example, the analytical activities are mostly driven by pattern search tasks, which are supported by pattern comparison and relation-seeking tasks, and the synthetic activities are mostly driven by connection discovery and pattern comparison tasks.

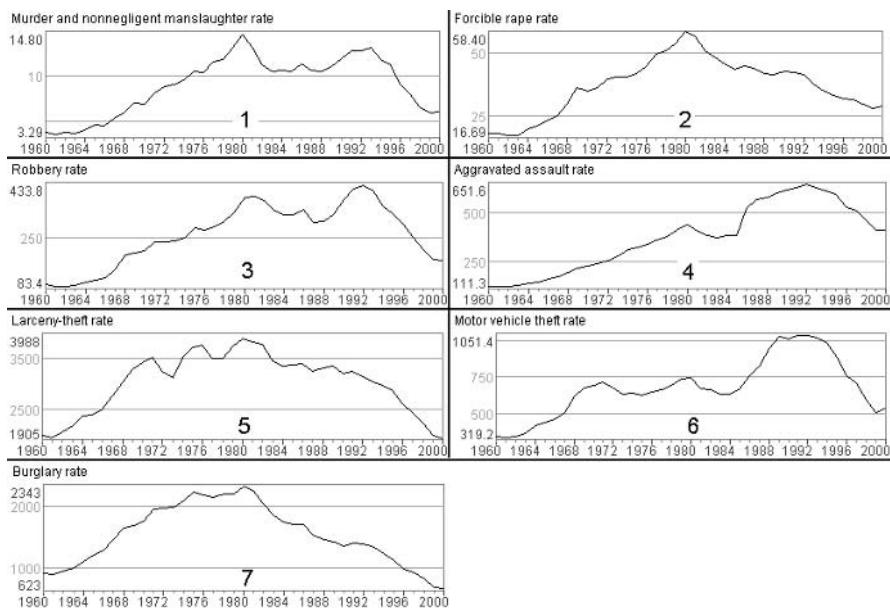
This example supports the statement made earlier (see Sect. 3.8) that tasks of different types and different generality levels may intermingle in exploratory data analysis, and therefore that the approach of building a specific tool that would optimally serve the needs of a given task is unfeasible and counterproductive. An explorer needs instruments that allow him/her to do a variety of tasks and have sufficient freedom to switch from one task type to another and to change the generality level. Accordingly, most tools for EDA are designed so as to be able to support a range of tasks and to connect to other tools complementing their capabilities.

The tool that we used for the exploration of the burglary rate in California was a time graph. It was sufficient almost for all tasks, including the synoptic tasks that had the primary importance: pattern search, pattern comparison, and relation-seeking (more specifically, looking for major changes in the trend). It was very beneficial that we could see the whole behaviour as a single image.

The discussion of this example is, actually, a hint as to where our general principles come from: they result from contemplating various examples and trying to generalise from them. Thus, readers should now be prepared to encounter such formulations as “see the whole”, “look for recognisable”, and “divide and group”.

To prepare ourselves even better for the following material, let us consider a few other, more complex examples. One of them is the exploration of not only the burglary rate but also the rates of other types of crime in California. A possible approach is to visualise the behaviour of each attribute on a time graph, as is shown in Fig. 5.2, and compare the behaviours. Note that the comparison of the behaviours starts from the very be-

ginning of the exploratory process and not after each individual behaviour has been fully characterised.

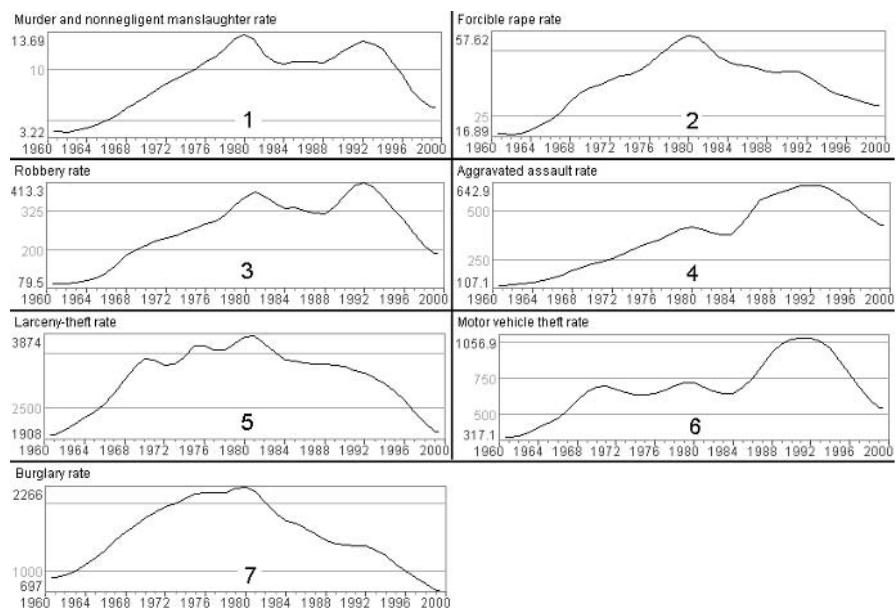


**Fig. 5.2.** Visualisation of the behaviours of seven attributes characterising the rates of different types of crime in the state of California during the period from 1960 to 2000. Each behaviour is represented on a separate time graph

As in the previous case, we need to concentrate on the general shapes of the lines and disregard small details and fluctuations. This can be done better when the lines are smoothed, as in Fig. 5.3.

By means of pattern comparison, we can note some features common to all or many behaviours. Thus, almost all of the crime rates mostly increased in the first half of the time period (from 1960 to 1980), and all of the lines have peaks in the year 1980. Four out of seven attributes have peaks in 1992, and all attributes have a decreasing trend after 1992. Four attributes reach their maximum in 1980, and the remaining three do so in 1992.

In this joint characterisation of multiple behaviours, we not only noted the distinctive features of the behaviours such as peaks, maxima, and increasing and decreasing trends (such observations belong to the class of behaviour characterisation, or pattern definition, tasks), but also associated these features with corresponding time moments and intervals by means of lookup tasks. In principle, it could be easier to check whether similar features of several behaviours occur at the same time moments or in the same



**Fig. 5.3.** The time graphs in Fig. 5.2 have been smoothed here so that minor fluctuations will not hinder pattern search and comparison

time intervals if the behaviours were represented in a common display rather than separately. For example, the lines representing the behaviours could be overlaid within a single time graph. However, this would require a previous transformation of the attributes to make them comparable, since the original attributes have quite different value ranges.

The more detailed characterisation of the detected characteristic features of the behaviours is done in basically the same way as in the characterisation of a single behaviour, i.e. we extract and characterise interpretable partial patterns. At the same time, we detect and measure the similarities and differences between the behaviours.

The synthetic process towards an overall pattern approximating the joint behaviour of the seven crime rates is based on the common features of the behaviours noted. The overall pattern must reflect the peculiarity of the years 1980 and 1992, the common trends, the similarities of the behaviours of some attributes, and the major differences. It would be also very nice to explain why the years 1980 and 1992 are so peculiar and why the behaviours 1 and 3 are so similar. However, the data and domain knowledge that we have are insufficient for those connection discovery tasks.

One major difference of this example from the previous one needs to be discussed. Whereas in the previous example we began with dividing, the exploration in the second example started with linking the behaviours by

similarity, which is more a synthetic than an analytical activity. How can this difference be explained?

Actually, it is not quite right to say that we started with dividing in the first case and with linking in the second case. In both cases, we started with visualising the data that we needed to analyse. In the first case, we managed to represent all the data in just one line, which is a highly holistic portrayal. As a result, the entire behaviour could be grasped in one sight, and no integration effort was needed. In the second case, the visualisation divided the overall behaviour into seven partial behaviours, which could not be perceived together. This is not a unique case: a holistic representation of all data in a single image is a rare possibility. So, the tools that we apply often divide the overall behaviour into pieces, and we need to spend considerable effort to bring these pieces together.

From the comparison of these two examples, we see that:

- synthesis may sometimes be prompted by an exploratory tool, such as a holistic visualisation, which allows us to perceive a behaviour as a whole in one instance of vision;
- besides the deliberate and voluntary division of a behaviour into parts, which is done to obtain a more precise behaviour characterisation, there may be a division that we are compelled to make, which is caused by the impossibility of representing a behaviour in a single image;
- when a piecewise rather than a holistic representation of the overall behaviour has to be used, synthetic activities such as linking or grouping are involved in the exploratory process from the very beginning.

The second example was more complex than the first one because we had to analyse multiple attributes instead of a single attribute. However, the data in both examples had a single referrer, specifically time. Let us now briefly discuss what we can do with multidimensional data, i.e. data that has two or more referrers. An example of such data is the time-series crime data for all states of the USA. The complexity of this dataset is such that there is no way to view the overall behaviour of even a single attribute, and we have to deal with aspectual behaviours (see Sect. 3.4.4): the behaviour (distribution) of the local temporal behaviours in space, and the behaviour (evolution) of the yearly spatial behaviour in time. Possible visualisations of these aspectual behaviours are shown in Figs 3.13 and 3.16, respectively. Unlike the previous case, where the behaviours of all attributes were visualised uniformly, and it was quite easy to do comparison and grouping, the representations of the aspectual behaviours in the current example are very different and cannot easily be joined into a coherent picture. The difference between the visualisations is not accidental;

it is a consequence of the intrinsic difference between the aspectual behaviours.

Another major distinction from the previous case is that the displays of the multiple attributes showed us different and even non-overlapping parts of the data, while now we have different views of the same data. Therefore, comparison and grouping of patterns perceived from different displays is not only difficult but also meaningless.

However, in the exploration of each of the two aspectual behaviours, comparison and grouping play an important role. Thus, using the visualisation in Fig. 3.13, we can compare and group the local temporal behaviours as we did previously with the behaviours of the different attributes. There is a difference from the previous case: while the behaviours of the attributes could be grouped arbitrarily, we must now take into account the spatial neighbourhood (see Fig. 3.14 for an example of grouping). Using the visualisation in Fig. 3.16 (or, rather, a more complete visualisation of the same kind, in which all time moments are represented), we can compare the spatial patterns, group similar patterns that occurred in consecutive years (i.e. take into account the temporal neighbourhood), and note the years when major changes of the spatial pattern took place.

In general, the process of characterising the aspectual behaviours is similar to what we did for the overall behaviours in the previous examples. However, the aspectual patterns so derived need to be integrated into a pattern characterising the overall behaviour.

Let us look once again at the representation of the local behaviours. In Fig. 5.4C, we have marked two prominent spatial clusters of similar local behaviours. One cluster, which is outlined in blue, is situated in the western and south-western part of the country. The other cluster, outlined in green, is in the north-central part. While the green cluster differs from the remaining territory by having quite low burglary rates, there are two states within the cluster that have even lower values than their neighbours. These are North Dakota and South Dakota, which are encircled in yellow.

It is true that we cannot compare the local behaviours shown in Fig. 5.4C with the spatial behaviour over the whole territory and its evolution over time. But why not to try to put the outlines that we have drawn on a representation of a yearly spatial behaviour?

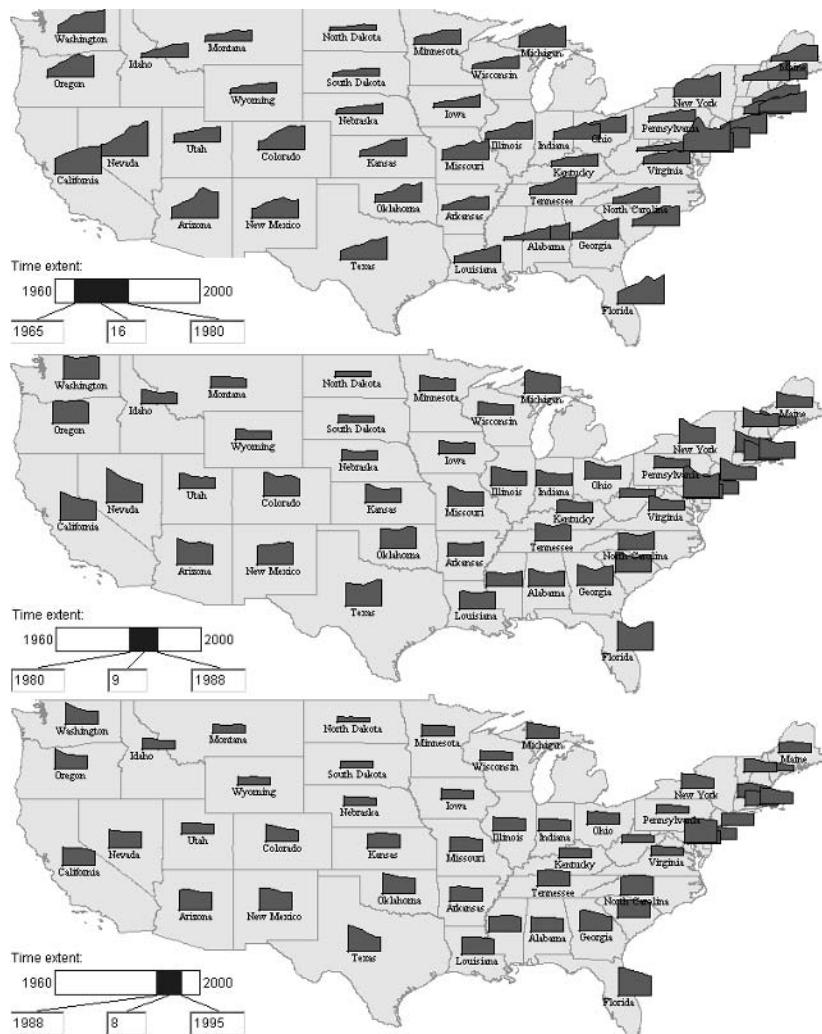
In Fig. 5.5C, the cluster outlines defined in Fig. 5.4C are superimposed on several maps showing the yearly spatial distribution of the burglary rates in selected years. Note that, unlike the case for Fig. 3.16, an individual colour encoding of the crime rate values is applied in each map, and hence the same degree of darkness corresponds to different values for different years. This has been done intentionally to achieve the maximum possible expressiveness for each individual map.

Now we can compare the cluster outlines with the spatial patterns for different years that may be perceived from the maps. We can note that the blue outline corresponds fairly well to the cluster of high values in 1985. In fact, the same cluster can be also seen in 1980, but it does not look so homogeneous as in 1985, owing to an extreme value in Nevada, which is represented by a dark brown colour. Moreover, from viewing the complete sequence of yearly maps (which cannot be reproduced here for reasons of space), we have made an observation that this cluster formed in about 1965 and mostly preserved its unity since then until the mid-1990s; however, in the 1990s the southern part of the zone of high values spread to the eastern coast. At the end of the 1990s, a zone of high values is observed in the south (see the map for the year 2000) and does not correspond any more to the blue outline.

The green outline corresponds to a zone of low values, which was also quite stable throughout the whole period. Analogously, comparisons between the two representations can be done for other groups of states with similar behaviours. In this way, we can link our patterns of aspectual behaviours.

We can also move in another direction. From observation of the evolution of the spatial behaviour over time, we have noted several time intervals where there were distinct development trends. Thus, for example, the period from 1965 to 1980 was a period of overall increase of the crime rate throughout the country, while the character of the spatial distribution remained mostly the same. From 1980 to about 1988, the spatial distribution did not change significantly, whereas the rates and the range of their variation decreased and the contrasts diminished. From 1988 to the mid-1990s, the spatial behaviour gradually changed; in particular, the distinction between the north and north-east, on the one hand, and the west and southwest, on the other hand, gave way to a distinction mostly between the north and the south. Figure 5.6C demonstrates the spatial distributions in several representative years. As in Fig. 5.5C, each map uses its individual colour encoding. Additionally, an operation of visual comparison with the median value for the whole country in the respective year has been applied in each map. The values higher than the median are shown in brown, and the values lower than the median in blue.

How can we link our observations of the changes of the spatial behaviour to our observations of the local behaviours? A possible approach is to look at fragments of the local behaviours corresponding to the different intervals revealed in the course of the study of the evolution of the spatial behaviour. Thus, the three maps in Fig. 5.7 show us the local behaviours in the intervals from 1965 to 1980 (top), from 1980 to 1988 (middle), and from 1988 to 1995 (bottom).



**Fig. 5.7.** By means of focusing, we can concentrate on fragments of the local behaviours corresponding to selected time intervals

In the first map, we see a quite consistent growth in most states and can easily detect some behavioural outliers, such as Oregon, Arizona, and Florida. In the second map, we can distinguish states with decreasing trends in the south-west and in the east from states in the north with nearly constant values. In the south-east, we see many M-shapes indicating decrease followed by increase. Again, behaviour outliers (e.g. New Mexico) are visible. In the third map, we see a decrease in the north-west, in the east, and in the south, and value stability almost everywhere else.

It can be noted that, if we group the states according to the similarity of their partial local behaviours, the grouping will be different for each interval (and different from the initial grouping as shown in Figs 3.14 and 5.4C). So, on the basis of the perceived pattern of the evolution of the spatial behaviour, we can refine the pattern of the distribution of the local behaviours.

We have demonstrated an example of linking patterns that approximate different aspectual behaviours. This process is rather non-trivial; it is not based on mere comparison in the sense of noting similarities and differences. Instead, we tried to discover connections between the aspectual behaviours: how a change in the spatial pattern is related to the local behaviours, and vice versa. We did this by propagating observations made with one visualisation to the other visualisation.

An even more difficult task is to characterise the behaviour of the different crime rates in space and time. We shall not describe how to analyse such data, but only mention that we would do the following in this case:

- divide the overall behaviour into the aspectual behaviours of the individual attributes;
- investigate these aspectual behaviours for similarities and differences;
- group the attributes according to the similarity of their aspectual behaviours;
- characterise the aspectual behaviours of each group and the differences between the groups;
- as in the previous example, try to relate the aspectual behaviours of each group;
- try to propagate the observations made for one group of attributes to the representations of the aspectual behaviours of other groups of attributes and, on this basis, try to find common features and note essential differences;
- if appropriate domain knowledge is available, try to explain the common features and the differences.

As a summary of the examples considered, let us recall what types of tasks play the most important role in the various exploratory activities:

- In analysis (division, extraction, and separation):
  - pattern search (matching behaviour fragments to the types of patterns expected);
  - synoptic relation-seeking (looking for major contrasts, changes, and discontinuities; detection of outliers and deviations from the major trend);
  - pattern comparison (differentiation between behaviour fragments).

- In characterisation:
  - pattern definition (ascribing the pattern to a particular type, and summarisation of characteristics);
  - elementary lookup and comparison (establishing the extent of the pattern and characteristic values such as the minimum and maximum, and characterising outliers).
- In synthesis (grouping and integration):
  - pattern comparison (noting similar patterns, grouping, and arranging);
  - pattern search (looking for patterns similar to a given one);
  - connection discovery (looking for correlations, dependencies, and structural links).

For the sake of fairness, it should be noted that exploratory data analysis does not always start with top-down activities, i.e. dividing a whole into parts. Moreover, synoptic tasks do not always play the leading role right from the beginning, as in the examples discussed. As a counter-example, we recall a case of EDA that occurred in our private life when we decided to buy a digital camera and needed to make a substantiated choice of the right model. The basic difference from the previous examples is that we had no full dataset to analyse. In fact, we had no dataset at all; the data had yet to be collected. However, it would not have been a good idea to try to collect all data about all existing digital cameras. This would have taken too much time and effort, and probably would never have ended since new models appear almost every day. Since we wanted to start taking snapshots quite soon, we took a more feasible approach.

From the beginning, we had certain constraints such as the price range and the minimum resolution, which would help us to reduce the set of all models to a subset consisting of potentially suitable models that it would make sense to investigate. In order to define this subset, we applied several query tools (search engines) available on the Web. These tools typically present their results as a list of products with hyperlinks to more detailed descriptions of the products, supplied by their producers or dealers (and hence not complying with any standard characterisation format). Sometimes opinions of customers and even evaluation reports from experts are obtainable, but not for all products.

So, we had obtained the subset of potentially relevant references (i.e. camera models). To establish an order of preference among them, it was necessary to compare the characteristics. However, these characteristics were dispersed among numerous individual descriptions of the cameras and hence could not easily be compared. Since the subset of models was still rather big, and the individual descriptions often quite lengthy, it was

hardly reasonable to try to put all the characteristics of all the models into a common table for a subsequent comprehensive investigation.

Instead, we approached the problem by means of sampling. We chose a few specific models differing in price and appearance, read their descriptions, and compared their characteristics. The purpose was not to assess the goodness of those particular cameras but to extract relevant attributes that could influence our choice, and the possible values of those attributes. Of course, on the basis of our (rather modest) domain knowledge and experience, we had some initial criteria in mind, such as weight, size, the availability of an optical zoom, but we did not know the current ranges of values of those attributes and also did not know how different characteristics were related. For example, were there pocket-sized cameras with the possibility of manual selection or adjustment of the settings for a photograph? What was the relation between the memory capacity and the price?

From an examination and comparison of the characteristics of the selected models, we extracted the set of attributes that needed to be taken into account, and some of their possible values. Then, we sampled other models in order to learn more about the variety that existed in the values of the chosen attributes. In order to obtain more information with less effort, we applied certain heuristics: for example, we looked at models with extreme prices, we compared different-looking cameras with the same price, and we compared cameras from different producers. In this way, we gained a considerable amount of relevant information.

In the course of collecting information, we judged some characteristics as inappropriate for us and, on this basis, pruned the set of options to be considered. For example, we discarded the whole class of pocket-sized cameras, since they did not allow any manual adjustment of the parameters of a shot. Among the remaining models, we already had quite a small subset of favourites, which were chosen with the use of the criteria that we had extracted. However, their characteristics were so close that it was very hard to make a choice between them.

Then, we tried to use the knowledge and experience of experts that was accessible. We read two or three expert evaluation reports about some of the models on our shortlist. This allowed us not only to enrich our knowledge about the models described but also to extract several additional criteria, such as support for focusing in low-light conditions, the lag between pressing the release button and taking a snapshot, and colour balance quality. Then, we had to make an evaluation of the other models with respect to these criteria. Information concerning some of the criteria could be extracted from the official descriptions, while the other criteria required a search through expert evaluations and customer comments. It was not possible to find the data of interest for all models. Therefore, we had to make

a decision as to how to deal with the missing data. Since we considered some criteria (e.g. image sharpness) to be very important, we decided to discard the models for which the corresponding data were missing. Then, we made our final selection from the remaining models on the basis of the available data.

We have written this long story in order to demonstrate a different method of exploratory data analysis as compared with the examples previously discussed. It is easy to note that elementary rather than synoptic tasks prevailed in this case: most of the time, we were examining the characteristics of specific cameras and comparing selected cameras, i.e. we were performing elementary lookup and comparison tasks. It may even seem that the exploration process consisted exclusively of such tasks, but this is not true. In fact, in a situation where we lacked full data, we used elementary tasks to extract the relevant attributes, estimate the ranges or varieties of their possible values, find the most typical values and note outliers, and learn which values of different attributes typically occur together and which combinations never occur. All this can be regarded as building a pattern for the distribution of characteristics throughout the set of cameras and for the links between attributes. Hence, behaviour characterisation and connection discovery tasks were also present in the process of exploring the digital-camera market. Of course, the pattern (mental model) that we eventually built was incomplete and imprecise, but it was sufficient for defining the feasible constraints and arriving at a shortlist of appropriate candidates, and we understood quite well the positions of those candidates with respect to the remaining models.

We hypothesise that such a bottom-up manner of data exploration is not applied only occasionally but takes place every time when complete data are not available at the beginning and cannot be collected because of practical constraints such as the time and effort required. In such a situation, this seems to be the only feasible approach. We should note that, unfortunately, most of the tools that we have considered are not applicable in a situation where one has only fragmentary data. We are not ready at the moment to tell what tools can be used in such a situation and how they might be used. A separate study of exploratory analysis of incomplete and uncertain data would be required to provide an adequate answer. Therefore, we restrict our further consideration to those cases where data have already been given, and there is no task of data collection.

## 5.4 Principles of Selection of the Methods and Tools

In the examples of data exploration that we considered in the previous section, we chose, individually in each case, certain approaches to analysing the data and, accordingly, certain exploratory tools: a time graph and a collection of time graphs, a map with “behaviour” symbols and a series of choropleth maps, smoothing and visual comparison, focusing and selection. We made this choice on the basis of our experience, which is mostly a tacit, subconscious kind of knowledge. The general principles by which we selected the approaches and tools were buried somewhere deep in our minds. Nevertheless, we have tried to externalise and verbalise them.

When we extracted some first, vague ideas, named them and put them in a list, we recalled that we had already encountered similar formulations in the books and papers of Bertin, Shneiderman, Klir, and Arnheim (Bertin 1967/1983, Shneiderman 1996, Klir 1985, Arnheim 1997). At first, we were quite surprised, but we then understood that this was not accidental. If our subconscious knowledge of the principles of data exploration indeed originates from the literature, this means that we have mastered the teachings of the best experts to such a degree that we have started to feel that they are our own. Moreover, this also means that these principles really are usable and useful. It is also possible that we have arrived independently at at least some of the principles (especially taking into account the fact that we read some of the above-mentioned works much later than when we started developing exploratory tools and analysing data with them). In that case, the coincidence means that these general principles really exist objectively, independently of whether we know them or not, and anyone can find them through experimenting and reasoning. Anyway, we feel safe and comfortable standing on the shoulders of giants.

So, let us now enumerate the principles that we extracted from the depths of our minds (or from the literature?). We have tried to give them short but expressive names:

1. See the whole.
2. Simplify and abstract.
3. Divide and group.
4. See in relation.
5. Look for recognisable.
6. Zoom and focus.
7. Attend to particulars.
8. Establish linkages.
9. Establish structure.
10. Involve domain knowledge.

The meaning of each name will be explained in the following sections. We would like to note that the items in this list are not arranged by importance or in the recommended order of application. The items can be grouped according to their relevance to analytic or synthetic activities. The first two principles are equally important for analysis and for synthesis. The principles from 3 to 7 are more pertinent to analytical activities, and principles 8 and 9 to synthetic activities. Domain knowledge (principle 10), when available, can be useful both in analysis and in synthesis.

So, let us now review these principles one by one.

#### 5.4.1 Principle 1: See the Whole

As we discussed in Chap. 3, a task, or question, in exploratory analysis consists of two parts: the target, i.e. unknown information, which needs to be obtained, and the constraints, i.e. known information, which is related to the target in a certain way. Of all the classes of tools that we have considered, only some query tools are designed in such a way that an analyst can formulate questions directly, i.e. specify the target and the constraints, and, in response, receive the information needed. However, as we noted, only elementary questions can be adequately supported in this way, since synoptic questions require the human's capability for abstraction.

All other tools for exploratory analysis suppose that the explorer finds answers to his/her questions by means of perception and reasoning. In other words, the explorer has either to see the answer in some display or display combination or to derive the answer from what he/she sees and what he/she has seen or inferred before (and, possibly, also from his/her pre-existing knowledge). Hence, the tools must create suitable conditions for the analyst to be able to see the answers to possible questions or to note information from which the answers can be derived.

One such condition is that the information that constitutes the answer or allows the answer to be inferred is present on the screen. Since this information is not yet known to the analyst and must be looked for, the attention of the analyst is guided by the constraints of the task. Hence, the information involved in the constraints must also be present on the screen. For example, if we use a graphical display to find the value of the burglary rate in California in 1980, the display must give us an opportunity to find the visual item corresponding to the year 1980, to the state California, and to the attribute "Burglary rate" (these are our task constraints), and then to extract the value of the attribute from this visual item. Hence, not only the item itself must be available, but also the information that allows us to identify it as corresponding to the constraints.

This was an example of an elementary task. Synoptic tasks also require the target and constraints to be present in a display or combination of displays. Since synoptic tasks deal with reference sets and the corresponding behaviours, these sets and behaviours should be seen on the screen. For example, if our task is to characterise the behaviour of the burglary rate in California over the period from 1960 to 2000, we need to see this entire period and the corresponding visual item(s) that represent the behaviour in the display.

In general, synoptic tasks require that the entire reference set is open to view as well as the corresponding characteristics. However, this is not enough. For these tasks to be done effectively, it is desirable that the characteristics are represented in such a way that the corresponding visual items are easily united into a single whole, so that the analyst can perceive them as a behaviour rather than as multiple individual items. Thus, the time graph in Fig. 3.12 was produced from 41 individual values of the burglary rate in California, but the representation in the form of a line allows these 41 values to be perceived all together as a unit. We can grasp the overall behaviour in just one sight.

Unfortunately, such a unified representation of an entire behaviour is rarely possible. In the example concerning multiple crime rates in California, we did not find a way to represent the joint behaviour of seven attributes as a single image. In the example concerning the burglary rates in all the states of the USA, we could not represent the entire reference set, i.e. space plus time, in such a way that the corresponding overall behaviour could be perceived from the display. The representation methods that we used, which involved space embedding and space partitioning (see Sect. 4.3.2), allowed us to see only the aspectual behaviours. A pattern approximating the overall behaviour had to be derived from our perception of the aspectual behaviours.

While the complexity of data rarely gives us an opportunity to create ideal conditions for accomplishing synoptic tasks, one should strive to meet the following requirements:

- *Completeness*. The entire behaviour (and hence the entire set of corresponding references) must be visible.
- *Unification*. The visualisation should support effective linking of the visual items representing parts or aspects of the behaviour into a single perceptual pattern.

Before considering these requirements in more detail, we would like to relate the principle “see the whole” to the teachings of the classics in visualisation. First of all, there is a clear link to Bertin’s views, which were discussed in Sect. 4.3.1 and are briefly summarised below:

- Graphical representation for information processing (i.e., in our terms, for exploratory data analysis) must be comprehensive, i.e. avoid any prior information reduction and allow one to find answers to any potential questions of any type or level.
- The most useful questions involve the overall level of reading.
- Ideally, a visualisation should be perceivable as a single image, in the minimum “instance of vision”. If it is not possible to construct such a visualisation, it is necessary to construct multiple comparable images (as few as possible) to provide answers to all questions.

There is also a link to Shneiderman’s principle “overview first” from his well-known “Information Seeking Mantra” .

We would also like to refer to Arnheim, who does not deal directly with data visualisation and analysis but considers the general properties of human vision and thinking. As we have already mentioned in Sect. 4.2, Arnheim argues that that perception consists in the grasping of relevant generic features of an object, and it is precisely this grasping of the character of a given phenomenon that makes productive thinking possible. Perception does not provide some “raw material” for thinking but immediately forms concepts, which are already quite general and abstract. It is these visual concepts that serve as material and tools for thinking. Hence, human perception and cognition are based on the approach “from above”, that is, from the whole to its constituents. This is highly related to our principle “see the whole”, which may also be formulated in a more precise form as “enable seeing the whole”. Any data analysis should start from an attempt to see the whole, and tools should support this appropriately.

Now, let us have a closer look to the two aspects of seeing the whole, referred to as completeness and unification.

#### 5.4.1.1 Completeness

The requirement of completeness implies that the entire reference set must be represented in a display whenever possible. As we mentioned in the section dealing with visualisation (Sect. 4.3.3), the dimensions of the reference set should preferably be mapped onto display dimensions. The mapping should preserve the essential properties and relations of the reference set such as ordering and distances but inhibit the influence of any emergent properties of the display (e.g. irrelevant ordering or distances) upon the analyst’s perception and reasoning.

Recall that display dimensions provide positions for placing visual items (marks), i.e. create a kind of reference framework for marks. This is close to the role of referrers in data, which provide a reference framework for

attribute values. Therefore, it is natural when data referrers are represented by display dimensions and attribute values are represented by marks and visual properties of these marks, such as colours, shapes, and sizes. Such a representation corresponds better to the observer's expectations and is therefore easier to perceive than if other representation principles are used.

The following display dimensions are available for visualisation:

- the two dimensions of the screen plane, i.e. width and height;
- the third spatial dimension, depth (in perspective views or special media);
- the temporal dimension, or display time;
- various arrangements in space, namely partitioning, embedding, and sharing;
- space transformations, which change the properties of the display space to make them conform better to the properties of the reference set, for example node-link structures, which introduce specific relations, or space segmentation, which eliminates an undesired perception of continuity.

There are two aspects of the requirement for a complete representation of the reference set: the dimensionality of the reference set, i.e. the number of referential components, and the size of the reference set, which depends on the number of different values of each referrer. For a complete representation, all the referrers must be mapped onto appropriate display dimensions, and all the values must be mapped onto appropriate positions in these dimensions. This means that each of the dimensions used must provide a sufficient number of distinguishable positions for representing the values of the referrer mapped onto this dimension.

It is also admissible that a referrer is represented by means of a retinal variable rather than a dimension. Recall that retinal variables correspond to the visual properties of marks: size, colour, shape, orientation, texture, etc. For example, in Fig. 4.4, where chains of arrow-shaped marks represent movements of objects, the orientation of the arrows is used to reflect the succession of the values of the temporal referrer. When a retinal variable is used, it is necessary that the variable can provide a sufficient number of distinguishable values to match the values of the referrer represented. It should be said, however, that representation of a referrer by a retinal variable is seldom effective, and hence is not highly recommended.

Of the display dimensions, the planar dimensions are the least limited with respect to the number of positions that they can provide. Although the display time also provides a potentially infinite number of different positions, this is not a priority choice for visualisation, because these positions

cannot be seen simultaneously, and hence comparison and grouping activities are seriously obstructed. It may be said that involving the display time in representing a reference set (e.g. in an animated display) does not fully meet the requirement of completeness: at each moment, the display shows only a part of the reference set or, more precisely, a slice resulting from fixing a specific value of one of the referrers.

This does not mean, however, that animated displays should never be used. When the spatial display dimensions are already in use, animation may be the only reasonable choice for representing a referrer with a large value set, since the other remaining dimensions (i.e. arrangements) do not provide enough positions. Thus, in Figs 3.16 and 5.6C, we could use a space-partitioning arrangement to represent only selected years of the period from 1960 to 2000. Because of screen size limitations, it is not practicable to display 41 maps simultaneously. Moreover, even if we had a very large screen on which all 41 maps could easily be fitted, the visualisation would be very difficult to perceive, since an explorer perceives each map as a separate image that needs to be studied and related to the other images. This process involves very many comparisons, eye movements, and attention switches. In contrast, an animated map is perceived as a single changing image, and human eyes need only to do their usual job, in which they are very well trained: to observe a dynamic scene and detect movements and other changes. If the changes are coherent rather than chaotic, the animated presentation promotes unification, i.e. perception of the information as a single behaviour rather than a sequence of slices. A space-partitioning arrangement does not have this property; it requires significant mental effort to reconstruct the behaviour from multiple distinct images.

Nevertheless, the transient character of the information representation in an animated display is a serious disadvantage, which needs to be compensated somehow. First of all, the user needs good facilities for controlling the animation: regulating its speed, stopping and resuming, jumping to a specific frame, and moving stepwise back and forth. Besides, the user should be able to retrieve selected frames for detailed examination and for comparison. The frames thus retrieved can be organised on the screen using a space-partitioning arrangement. Hence, rather than choose between utilising the display time or juxtaposing multiple displays, it is better to combine these two approaches to benefit from the strengths of each of them and mitigate their weaknesses.

As we have said before, for a complete representation of a reference set, it is necessary to choose a dimension with a sufficient number of distinguishable positions (or, less preferably, a retinal variable with a sufficient number of distinguishable values) for each referrer. However, it is not al-

ways possible to fulfil the requirement of completeness. Two basic problems may arise:

1. A referrer has so many different values that none of the available dimensions or retinal variables can provide enough positions or values.
2. There are so many referrers that it is impossible to represent all of them simultaneously by appropriate dimensions or variables.

In the case of the first problem, data aggregation is typically used. Strictly speaking, aggregation does not solve the problem. It transforms the reference set so that the number of different references is significantly reduced, and a complete representation through the available display dimensions becomes possible. However, this will be a complete representation of the transformed reference set, not the original one. As we discussed in the corresponding section (Sect. 4.5.4), data aggregation involves significant information loss. Hence, the requirement of completeness is violated at the stage of data transformation.

Still, this does not mean that data aggregation is not recommended for use. It deserves to be recommended not only because there may be no other possibilities for dealing with a very large reference set, but also because data aggregation is a way to simplify the data and display, which is highly desirable for synoptic tasks. Data aggregation helps the analyst to disregard excessive detail and thereby promotes abstraction.

Another possible approach to the visualisation of data with a very large reference set is the representation of selected portions of the data. Of the two approaches, aggregation is much more appropriate for synoptic tasks, while selection is more suitable for elementary tasks, which do not involve grasping of the overall behaviour but require every data element to be accessible. It may be said that aggregation, as compared with selection, is more compliant with the principle “See the whole”, since it provides a condensed representation of the whole dataset. Characteristics of aggregates are not just arbitrarily selected attribute values; they are intended to provide appropriately condensed information about all the members of the aggregates.

As with the use of display animation, certain rules of thumb apply to utilising data aggregation in exploratory data analysis. These rules were discussed in Sect. 4.5.4, and here we shall briefly recall them:

- Be cautious with averaging. Pay attention to the value range, the character of the distribution, and the presence of outliers.
- Prefer positional measures to means.
- Do not rely upon a single aggregation; vary the level and method of aggregation.

The latter recommendation implies that the tool used for aggregation provides sufficient flexibility in defining and characterising aggregates, allows the user to change choices previously made, and reacts promptly to such changes, with corresponding data reaggregation.

Aggregation and selection are also used in the case when there are too many referrers. To be able to visualise highly multidimensional data, the explorer needs to eliminate the variation of the values of one or more referrers and to use the available expressive means for the other referrers. There are two possible ways to eliminate the variation: either the explorer chooses a specific value of areferrer and disregards the others, or the data are aggregated by means of combining all values of areferrer.

Let us explain these two opportunities with an example. In Chap. 2, we described a dataset containing the results of a simulation of forest development under different forest management scenarios (see Sect. 2.3.7 and Fig. 2.11). This dataset includes five referential components:

- two-dimensional geographical space divided into forest compartments;
- time (measured in years; the simulation was done for 200 years);
- management strategy, with four possible values: natural, selective, Russian, and illegal;
- tree species, with six different values: aspen, birch, oak, pine, spruce, and lime (denoted in the dataset by its Latin name *Tilia*);
- age group (represented by an integer number from 1 to 13).

It is hardly possible to visualise these data in such a way that all the referrers are completely represented. We need to eliminate the variation within some of the referrers. The choice of these referrers depends on our goals. Let us suppose that our goal is to compare the development of the forest over time under the different strategies of forest management. This means that we should not eliminate the variation within the temporal referrer and the management strategy referrer. A good candidate for elimination is the spatial referrer: if we choose to represent it fully, this will take two spatial dimensions of the display, while any other referrer may be represented using just a single dimension. For example, we can use the horizontal dimension for the time, the vertical dimension for the age group, and an arrangement in which we partition the display space into four rows and six columns, and each row corresponds to a management strategy and each column to a species. Hence, the visualisation will consist of 24 images, each image corresponding to a particular combination of a management strategy and a tree species. Within each image, the areas occupied by different age groups at different time moments can be represented by proportional sizes of marks (e.g. circles).

However, we need to handle the spatial referrer somehow. There are two opportunities:

- We select a specific value of the spatial referrer, i.e. a specific forest compartment, and consider how the species and the age structure develop in this compartment over time under the different management strategies.
- We aggregate all the compartments together; for example, we compute the total area for each species, age group, management scenario, and year over the whole forest. Then, we analyse how the species and the age structure develop in the forest as a whole over time under different management strategies.

We do not claim that the resulting visualisation will be effective, and we shall not discuss how it may be used and what strengths and weaknesses it has (in fact, the tools that we have at our disposal do not allow us to build such a display). The only purpose of this example is to illustrate how the dimensionality of a reference set may be reduced by means of selection or aggregation.

Nevertheless, to be consistent with our emphasis on the role of visualisation, we would like to give a visual illustration as well. This demonstrates another possible way to reduce the dimensionality of the same data, out of the wide variety of options that exist.

This time, we reduce the dimensionality at the cost of the referrer “Age group”. So, we apply data aggregation; specifically, we sum the areas occupied by different age groups of the same species, and hence disregard the age differences. To visualise the result of the transformation, we construct a collection of four animated maps, each map corresponding to one forest management scenario. Hence, we use a space-partitioning arrangement to represent the scenarios, the two-dimensional space within each image (map) for the spatial referrer, and the display time for the temporal referrer. To represent the species, we apply a space-embedding arrangement: within the space of each map, pie charts portray the areas occupied by different species in each forest compartment. The sizes of the pie charts are proportional to the total area occupied by all the species together, while the angular sizes of the sectors show the proportions of the various species in the total area. Figure 5.8C demonstrates a screenshot from this visualisation corresponding to the 100th simulation year of the 200-year long simulation period, and Fig. 5.9C shows the situation on the 200th year. The map in the upper left corner in each figure corresponds to the natural scenario, the map in the upper right to the selective cutting scenario, the map in the bottom left to the Russian legal system, and the map in the bottom right to the illegal-cutting scenario.

We shall not perform a detailed analysis of the information perceivable from these maps, but instead suggest that readers note the differences between the consequences of the different forest management strategies. However, we would like to make some comments about the illustration in Figs 5.8C and 5.9C.

In order to construct the animated maps with pie charts, we reduced the dimensionality of the data by means of aggregation over the age group referrer. The other referrers did not undergo reduction. However, we could not insert the animated maps in this book as an illustration. Therefore, we actually applied another reduction: we selected two particular values of the temporal referrer and produced static pictures corresponding to each of the values. Hence, both in Fig. 5.8C and in Fig. 5.9C two referrers are in a reduced state: the age group referrer has been reduced through aggregation and the temporal referrer through selection.

Another comment is that, as a consequence of aggregation, we have lost the information concerning the age structure of the forest, which is very important for a comparison of the different scenarios. Therefore, it is necessary also to try other ways to visualise and analyse the same data. Of course, it is possible to aggregate the data over the species dimension and look at the age structure irrespective of the species. Or, as we already mentioned, the dimensionality can be reduced at the cost of the spatial component in order to consider both the species dimension and the age groups dimension. These and other transformations can be done analogously to what we did with the age group component. Of course, not only sums but also other aggregate characteristics can be used, depending on the nature and distribution characteristics of the data and the goals of the analysis.

We would like to demonstrate another approach to reducing data dimensionality. It is also based on the selection of values of the referrer(s) undergoing reduction but applies another selection principle. The approach is demonstrated in Fig. 5.10C. We have taken one screenshot of four animated maps corresponding to four different forest management scenarios. The screenshot corresponds to the 200th year of the simulated development of the forest. In each map, the forest compartments are coloured according to which species and which age group dominates, i.e. occupies the maximal area. Different colour hues represent the species, and the degrees of darkness represent the age groups: the older the trees, the darker the colour. Black signifies the compartments that have no or very few trees because of cutting.

In this visualisation, reduction has been applied to both the species and the age group referrers. The reduction has been done by means of selecting certain values of these referrers. However, there are two differences between this selection and the selection of the 100th year in Fig. 5.8C and of

the 200th year in Figs 5.9C and 5.10C. First, the values are selected individually for each combination of scenario, simulation year, and forest compartment. Second, the selection is not arbitrary but is made according to a certain rule. In this example, the rule prescribes that what is selected is the combination of species and age group for which the corresponding value of the attribute “Area” is the highest among all combinations. In principle, other rules may be applied as well. For example, a potentially useful rule could be to select the combinations for which the corresponding areas reach beyond a specified threshold (which may be specified as an absolute value or as a proportion of the total area of the respective compartment). For the selected combinations, coloured bars could be applied to portray the corresponding areas.

Again, we shall not go into a detailed analysis of what we can learn from the visualisation of the dominant species and ages. Our main goal has been to demonstrate different ways of reducing the dimensionality of data rather than to perform an actual comparison of the different forest management strategies.

Several important notes need to be made concerning dimensionality reduction. First, reducing the dimensionality by means of selecting a specific value of a referrer implies that the explorer must repeat the process of visualisation and analysis for every value of the referrer undergoing the reduction. The partial patterns so derived need to be integrated into an overall pattern. Second, when using aggregation, one should be aware of its pluses and minuses, which have been discussed earlier. The same guidelines as for handling large data volumes are also applicable in this case. Third, dimensionality reduction inevitably results in information loss. It is necessary to ensure that the information thus omitted is not overlooked. For this purpose, the explorer should apply several different ways of reducing the dimensionality, so that components that are reduced in one view be fully represented in alternative views (at the cost of reducing some other components).

It is true that the resulting procedure of data exploration becomes very complicated and requires considerable cognitive effort to join multiple partial views into an integral overall pattern. However, the requirement of completeness can only be fulfilled when appropriate attention is paid to every data component. In a situation where it is impossible to consider all components simultaneously, there is no other way to fulfil this requirement than to use multiple complementary visualisations and try to integrate the partial knowledge derived from them into a coherent picture of the overall behaviour.

It is necessary to say that effective perception of the overall behaviour is not guaranteed even when there are enough display dimensions to repre-

sent the referential components of the data. Let us recall the example of the visualisation of the variation of the burglary rate over the territory of the USA and over the time period from 1960 to 2000. Throughout the book, we have used three alternative mappings of the referrers of this dataset (i.e. two-dimensional space and time) onto the display dimensions:

1. The spatial referrer is mapped onto the two-dimensional display space and the temporal referrer is mapped onto the horizontal dimensions of multiple subspaces embedded in the primary display space. This representation is applied in the maps with superimposed time graphs shown in Figs 3.13 and 5.4C.
2. The spatial referrer is mapped onto the two-dimensional display space and the temporal referrer is mapped onto a space-partitioning arrangement. The resulting display consists of multiple maps corresponding to different years, as in Figs 3.16, 5.5C, and 5.6C.
3. The temporal referrer is mapped onto the horizontal display dimension, and the spatial referrer is mapped onto a space-sharing arrangement (thereby, the inherent properties of the spatial referrer are ignored, and it is treated as a referrer of the population type). This mapping is applied in the time graph shown at the top in Fig. 4.47: the variation of the burglary rate in each state is represented as a line on the graph, and the lines for the different states are overlaid within the same display space.

We have also used an animated map display, where the temporal dimension was represented by means of the display time, but we could not include this visualisation as an example in the book.

Let us compare the first and the second visualisation. Formally, they represent exactly the same information and are therefore equivalent (we cannot say the same about the third visualisation, which omits important properties and relations of the spatial referrer). But are they really equivalent with respect to the information perceived from them?

As we discussed earlier, the first representation allows us to perceive the local behaviours of the burglary rates in different states and the distribution of various behaviours over the territory of the USA. The second representation allows us to perceive the distribution of the burglary rates over the territory of the USA in different years and how this distribution changes over time (the same information can be perceived from the animated map display). These are two different sorts of information; we have called them “aspectual behaviours” and argued that they are not equivalent to each other and that neither of them is equivalent to the overall behaviour of the burglary rate in space and time (see Sect. 3.4.4).

Hence, neither the first nor the second visualisation enables a perception of the overall behaviour, although there is no problem with data volume or

data dimensionality, and all components of the data are appropriately represented. What is the reason for our seeing only the aspectual behaviours instead of the overall behaviour?

The reason lies in the distinction between the perceptual properties of the various display dimensions. The primary spatial dimensions, i.e. width, height, and, to some extent, depth,<sup>29</sup> have uniform properties, first of all, continuity. Moreover, these dimensions are integrative, i.e. they jointly form a unified two- or three-dimensional space. As Bertin claimed and psychological studies have confirmed, this space, possibly filled with visual stimuli differing in a single feature such as colour or size, can be perceived all at once, as a single image. No other dimension has the same capability. Thus, the temporal dimension does not allow us to see all the information simultaneously. The space-partitioning and space-embedding arrangements produce composite displays containing multiple smaller displays. Each of these subdisplays has its internal space, and the individual spaces of the subdisplays are not perceptually integrated. While each of the subdisplays taken separately can prompt holistic perception, the composite display is seen not as a single image but as a collection of images.

When the subdisplays represent slices of the overall behaviour, each of the slices can be perceived holistically, i.e. we see each slice as a behaviour. Thus, each of the time graphs embedded in the map in Fig. 5.4C is seen as a single image of a local behaviour in time. Each map in the multimap display in Fig. 5.5C or 5.6C is seen as a single image of a behaviour in space. Then, the entire composite display is perceived as variation of these images over a larger space. Thus, we see a variation of local temporal behaviours in Fig. 5.4C and a variation of momentary spatial behaviours in Fig. 5.5C or 5.6C.

Hence, when there is no possibility to represent the whole reference set of a dataset using the primary spatial dimensions of a display, the visualisation does not support the perception of the overall behaviour but only the perception of a certain aspectual behaviour. In order to fulfil the requirement of completeness, an explorer needs to consider all possible aspectual behaviours (if there is no special reason to give more priority to particular aspectual behaviours) and, on this basis, try to synthesise a concept of the overall behaviour.

We have not discussed yet the third visualisation of the burglary rate data, that is, the time graph with multiple overlaid lines corresponding to

---

<sup>29</sup> Depth is not fully equivalent to width and height unless a special three-dimensional display medium is used. On a two-dimensional screen, the third spatial dimension has to be simulated by means of depth cues; it cannot be utilised in exactly the same way as the horizontal and vertical dimensions.

different states (Fig. 4.47). We have mentioned that this visualisation is not fully appropriate, because it omits the pertinent properties and relations of the spatial referrer. Therefore, exploratory analysis of these data should not rely upon this visualisation alone. This visualisation can only be used as a complement to other visualisations which preserve the essential properties of the referrers.

However, would this visualisation be sufficient if the referrer were not spatial? Let us suppose that the time graph represents, for example, the variation of the yearly incomes of 50 different people or some other space-irrelevant phenomenon. Can we say that we can perceive the overall behaviour of the phenomenon over the population and over time, since all data are shown within a single space, and there are no separate, non-unifiable spaces of multiple subdisplays?

We dare to say (merely on the basis of our introspection, since we are not aware of any relevant psychological studies) that a time graph has the potential to prompt a holistic view of the entire behaviour. However, this potential is limited with respect to the data volume; in other words, a space-sharing arrangement provides a limited number of distinguishable positions. When the lines on a time graph are too numerous, they typically overlap greatly, and this impedes the perception of the overall behaviour. With fewer lines, it is quite easy to detect some common behavioural features, such as a general increase or decrease. Increases and decreases of value variability are also easily seen. One can make general observations concerning the presence or absence of any periodicity in the data. So, quite a large amount of knowledge can be gained just from the overall appearance of the display, without scanning every individual line.

When the lines are too numerous, data aggregation may be quite helpful to the analyst in building a concept of the overall behaviour. In Fig. 4.75, we have demonstrated an approach to using aggregation in a time graph display. This approach is based on the use of positional measures. Another approach is shown in Fig. 5.11C. We have divided the value range of the attribute “Burglary rate” (from 0 to 2907) into several intervals. The aggregation tool has counted for each year how many values fit into each of the intervals. The resulting counts are represented in the lower part of the display in Fig. 5.11C. Here, each vertical bar corresponds to one year. It is divided into coloured segments with heights proportional to the number of values belonging to each interval. The upper part of the display is the original time graph, with the background coloured according to the division of the attribute value range, which helps in understanding the lower subdisplay. The thick black line on the time graph connects the yearly median values. This is one more, rather crude, variant of aggregation.

The lower subdisplay in Fig. 5.11C offers a highly holistic view of the overall behaviour. We see a rapid decline of the green part (i.e. the number of states with low values) during the 1960s, a rise of the red part (corresponding to high values) over the first two decades and a gradual decrease of this part over the next two decades, a peak of criminality in 1980 and 1981, and a general improvement in the situation at the end of the entire period. This visualisation also allows more detailed observations. However, as we have already said, aggregation inevitably involves information loss. Therefore, the use of only aggregated data displays is not justified when there is a possibility to see the full information, as in the current example. In such cases, aggregation serves as a useful complement to visualisation of the original data but not as a substitute for it.

In our discussion concerning the perception of overall and aspectual behaviours, we referred to the capability of certain display dimensions to be perceptually unified (specifically, the width, height, and depth of the display are perceived together as an integral space). So, we have already touched upon the topic of unification, which is quite closely related to the topic of completeness.

#### **5.4.1.2 Unification**

Data always consist of multiple individual items. Synoptic tasks require these multiple items to be regarded as a single unit, as the behaviour of the underlying phenomenon. Therefore, the tools used for exploratory data analysis, in particular, visualisation, should promote the perception of multiple data items as a single whole.

It is relevant to refer here to psychological theories of visual perception, which attempt to describe how humans organise visual elements into groups or unified wholes. Most contemporary researchers in cognitive psychology recognise the “gestalt” principles, which were originally developed by German psychologists in the 1920s and later extended and refined. The word “gestalt” means a unified or meaningful whole. Gestalt theory arose as a reaction to the prevalent psychological theory of the time, atomism. Atomism examined parts of things with the idea that these parts could then be put back together to make wholes. Gestalt theorists, in contrast, focused on studying how our mind perceives wholes out of incomplete elements. They believed that human perception and cognition proceed “from above”, from the whole to its constituents. Here is a brief summary of the gestalt principles based on material that we have found on the Web (where such material is quite plentiful):

1. *Principle of proximity.* We tend to perceive elements as being associated when they are close together.

2. *Principle of similarity.* Those elements that share qualities (of colour, size, or shape, for example) will be perceived as part of the same form.
3. *Principle of good continuity.* We prefer to perceive smooth, continuous contours rather than abrupt changes in direction. Elements that continue a pattern tend to be grouped together.
4. *Principle of closure.* We tend to enclose spaces by completing contours and ignoring gaps in figures. It follows from the principle of good continuity and allows us to group elements together or to interpret forms as complete even though parts may be missing.
5. *Principle of figure/ground.* We tend to perceive some visual elements as a figure, with a definite shape and border, while other elements appear as a ground, further away and behind the main focus of the figure. There are two other principles related to the organization of the visual material into figure and ground:
  - a) *Surroundedness.* The elements of an image that are seen as surrounded will be perceived as the figure, and the elements that are doing the surrounding will be perceived as the ground.
  - b) *Smallness/area.* When two figures overlap, this principle states that the smaller of the two will be considered as the figure and the larger will be perceived as the ground.
6. *Principle of symmetry.* When elements may be viewed as parts of some symmetrical figure, they are seen as the whole figure. Arnheim considers symmetry as “a special case of fittingness, the mutual completion obtained by the matching of things that add up to a well-organized whole” (Arnheim 1997, pp. 64–65).

The most general principle, which embraces all others, is the principle of Prägnanz: We are innately driven to organise things in as good a gestalt as possible. “Good” can have various meanings, such as regular, orderly, simple, or symmetric, which then refer to specific gestalt principles.

The gestalt principles suggest that human perception has an inherent tendency towards unification. However, the unification occurs only if the visual material allows this. Arnheim writes:

Assimilation is probably the primary condition. Homogeneity prevails unless a sufficiently strong stimulus breaks up the field into separate units, as when a red object is seen on a green ground or when parts of the field are separated by a spatial distance or when an object moves through an immobile environment. Separation by difference imposes itself also when the observer is called upon to make a choice among given items. (Arnheim 1997, p. 65)

While the gestalt principles can be used consciously in art and design, it is not so for data visualisation. An artist or designer is usually quite free to

arrange visual elements and choose their colours, shapes, sizes and other features. In visualisation, the arrangement and visual properties of display items are not arbitrary; their task is to represent data, and hence they are determined by the data. When the positions, colours, sizes, shapes, or other features of display items encode elements of data, it cannot be guaranteed that the resulting picture will allow any of the gestalt principles to work.

Nevertheless, there are certain techniques that promote holistic perception and are applicable to a range of display types. We have recently mentioned one such technique: in a time graph, positions corresponding to attribute values at different time moments are connected by lines. In the resulting graph, we have a single visual element, a polygonal line or curve, instead of multiple separate elements representing individual attribute values. This helps us to grasp the whole behaviour immediately, in one sight. The same technique was applied in our map showing the migration of storks (see Fig. 4.4).

Another technique that works in a similar way is known as adjoining. For example, a bar chart is better perceived as a unified whole when the bars are in contact than when the bars stand separately.

One more technique or, rather, recommendation is to use the integrative display dimensions, i.e. width, height, and depth, as much as possible. Not only are they perceived together as a single space but also facilitate holistic perception of marks positioned in this space. However, as Bertin's image theory claims and psychological experiments mostly support, marks can be seen together as a unified whole only if they differ in one visual feature, for example colour or size. According to Bertin, only a visual construction involving two planar variables and one retinal variable can be perceived as a single image (Bertin did not consider the third spatial dimension, depth). This means, in particular, that only one spatially related attribute can be represented on a map display in a way that prompts unification. In such a map, the two available spatial dimensions of the display are utilised to represent the spatial referrer, and a retinal variable has to be used to portray the values of the attribute.

However, data to be analysed typically contain more than three components. In Sect. 5.4.1.1, we discussed how multiple referrers can be dealt with; here, we shall focus on multiple attributes. The following approaches to visualisation of multiple attributes are possible:

1. Values of several attributes corresponding to a common reference are encoded in different visual features of a mark: position within a display, size, shape, colour, etc. Very often, positions cannot be used to represent attributes, since the spatial display dimensions have been chosen to represent referrers. In this case, only retinal variables may be combined to

encode different attributes. This situation always occurs in map displays of spatially referenced attributes.

2. Multiple attributes are represented by means of charts – compound graphical constructions consisting of multiple, typically uniform graphical elements, where each element stands for one attribute. The elements are arranged in a certain way within a chart; this means that a chart uses at least one dimension of its internal space for the arrangement of the elements. Visual properties of the elements are used to encode the values of the respective attributes. Some examples are bar charts, pie charts, and segmented bars. Multiple charts can be embedded in the space of a map or other display.
3. When there is a spatial dimension that is not used for a referrer, multiple attributes may share this spatial dimension, i.e. a space-sharing arrangement may be applied. For example, this technique can be used to represent the behaviours of several numeric attributes in a single time graph. In this case, the vertical dimension is used for encoding attribute values. As we have mentioned before, this requires the attributes to be comparable, i.e. to have the same or very close value ranges. If this requirement is not fulfilled, the attributes need to be transformed to become comparable.
4. Each attribute is represented in an individual display, and the displays are juxtaposed on the screen, i.e. a space-partitioning arrangement is applied. In particular, several spatially referenced attributes may be represented in multiple maps, which are put side by side for comparison. Bertin and Tufte strongly advocate this technique (Bertin 1967/1983, Tufte 1983, 1990) but not everyone is so enthusiastic about it.

Let us now consider a few examples. We assume that the primary spatial dimensions of the display are not available for representing attributes, since they are used to reflect referrers. With this assumption, map displays will be used for the purposes of illustration.

We shall start with a situation where two spatially referenced attributes need to be jointly explored. Figure 5.12C demonstrates two different approaches to the representation of a pair of attributes on a map display. Specifically, we have taken the attributes “% 0–14 years” and “% 65 or more years” from the Portuguese census dataset.

On the left in Fig. 5.12C, the values of both attributes are “packed” together into the colouring of the districts in the map. From Bertin’s viewpoint, only one retinal variable, colour, is used. However, current researchers distinguish different components of colour and acknowledge the possibility, in principle, of utilising these components for encoding different types of information. While the most widely accepted division of colour is

into hue, saturation, and brightness, an alternative division into red, green, and blue components also exists, especially in computer-based visualisation. We have applied the latter division and chosen the red component to portray the proportion of elderly people and the green component to portray the proportion of children. So, the degree of greenness of the colour encodes the proportion of children in the population (the more children, the greener the colour), and the degree of redness encodes the proportion of elderly people (the more elderly people, the redder the colour). Low values of both attributes appear as yellow shades, and a brown colour, which is a mixture of red and green, would reflect high values of both attributes, but such value combinations do not occur in the dataset.

On the right in Fig. 5.12C, the values of the attributes are represented by visual properties of the triangular marks; specifically, the widths represent the proportion of elderly people, and the heights the proportion of children. So, tall, narrow triangles appear where there are many children but few elderly people, and low, wide triangles mark the places with few children and many elderly people.

Is it possible to say which of these displays is better for the exploration of the joint spatial behaviour of these two attributes? The coloured map prompts unification strongly; it is definitely capable of being perceived all at once, as a single image. We cannot readily say the same about the map with triangles; it requires at least some training to be seen holistically (for us, it took about a minute). On the other hand, the coloured map is more difficult to interpret. It is easy to learn that green is used for children and red for elderly, but understanding the meanings of the various colour mixtures requires the viewer to look repeatedly at the legend. For the interpretation of the triangles, it is sufficient to look at the legend once in order to understand the general principle. As soon as the principle has been grasped, the viewer can concentrate fully on observing the distribution of different shapes throughout the map.

The triangles allow both selective attention to either of the dimensions, i.e. either to the width or to the height, and conjunctive attention to both dimensions, which takes the form of judging the shapes of the triangles: how harmonious they are. This is different from the perception of the coloured map: we cannot selectively attend to the degree of greenness or redness and ignore the other component. However, since current computer-based visualisation tools allow an explorer to use different displays for different purposes, there is no necessity to support selective and conjunctive attention simultaneously.

A disadvantage of the coloured map in comparison with the triangle map is the information loss due to the classification involved: the attribute values are not transformed directly into colours, but instead the value

range of each attribute is divided into five equal-length intervals, and the colours are assigned to combinations of the intervals. However, the classification is quite fine and does not significantly distort the mental picture of the joint behaviour of the two attributes. Although no classification or aggregation is involved in the triangles, they allow in any case only approximate judgement of the values. Generally, there is no strong necessity to combine capabilities to support unification and to enable exact judgement of values in a single tool, since there is the possibility to combine tools.

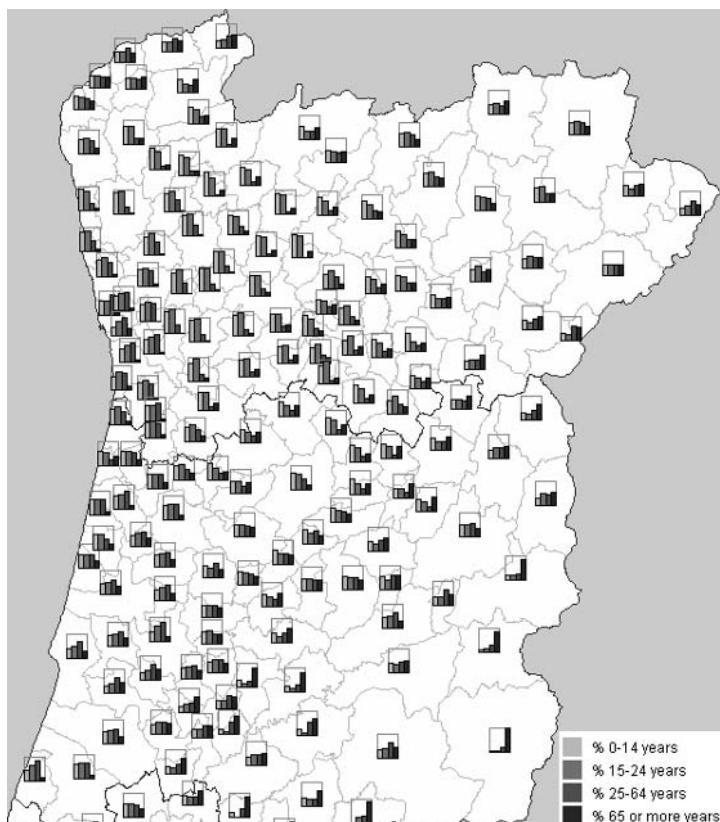
It seems that we cannot reach a final verdict as to which of the visualisations is better. This appears to be largely a matter of personal details of perception, personal preferences, and training. Still, one general comment is relevant here. When data refer to areas in space rather than to points, they may be represented on a map either using area colouring or by means of symbols or charts (when data refer to points, area colouring is not applicable). A disadvantage of the latter approach is that the symbols or charts may overlap significantly, which complicates perception. It is not always possible to regulate the sizes of the symbols or charts so that they do not overlap but still remain clearly visible. In our example, the overlapping of the triangle symbols is tolerable; however, to use diagrams, we need to increase the size of the map, as will be seen from the following examples.

To conclude the current example, we can say that representation of two attributes by a combination of two retinal variables works sufficiently well. However, one cannot expect that any combination of retinal variables will work equally well. Thus, we were not happy with combining symbol size and colour. When the sizes of symbols are small, the colours are poorly distinguishable. The colours of larger symbols attract more attention than the colours of smaller ones. Hence, the attributes represented by the size and by the colour are treated unequally, and the values of the attribute represented by the size distort the perception of the values of the other attribute. Another remark is that trying to combine more than two retinal variables is hardly productive; at least, we cannot give an example where such a combination was effective.

In the next example, we shall try to visualise simultaneously four attributes, specifically, the four age structure attributes of the Portuguese census dataset. This time, we shall apply the approach of building charts and then compare this with the use of four juxtaposed maps (“small multiples”).

In Fig. 5.13, the attributes are represented by means of bar charts embedded in a map. To reduce the overlap of the charts, we had to increase the size of the map. Since the resulting map is very large, only the northern half of it appears in the illustration. The attributes have quite different value ranges and are not directly comparable. Therefore, rather than represent attribute values by proportional bar heights, we have chosen another

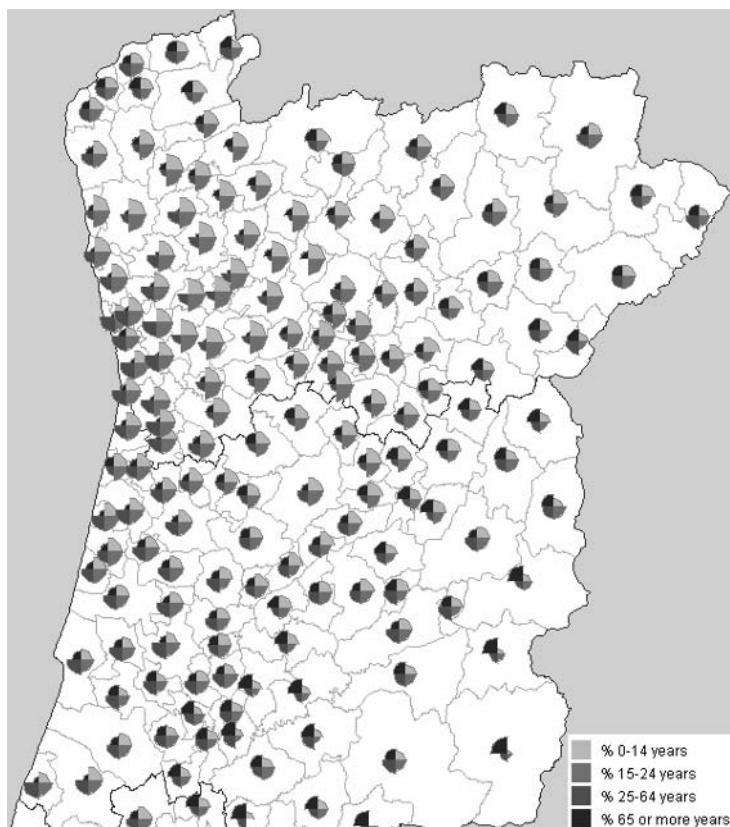
approach. Each bar represents the value of one of the attributes by the position of its upper edge between the bottom and the top of the diagram. For each attribute, the bottom corresponds to the minimum attribute value and the top to the maximum attribute value occurring in the dataset; hence, each bar has its own scale. The bars constructed in this way characterise the position of each district of Portugal in relation to the other districts in terms of the proportions of the four age groups.



**Fig. 5.13.** Four age structure attributes are represented on a map here by means of bar charts. Each bar represents the value of one of the attributes by the position of its upper edge between the bottom and the top of the diagram. For each attribute, the bottom corresponds to the minimum attribute value and the top to the maximum attribute value occurring in the dataset; hence, each bar has its own scale and is not comparable to the other bars in the same chart

However, it is not our current task to investigate the relative position of each individual district with respect to the others. We need to overview the map and get an overall idea concerning the spatial behaviour of the age

structure of the population. For this purpose, we need to focus on the general shapes of the charts and on how these shapes vary over the territory, rather than on the sizes of the bars in any individual chart. The question is whether the map supports this. It is clear that the charts do not form a unified image as readily as the colouring of the map in Fig. 5.12C, and definitely they require more training than do the triangles. Nevertheless, the overall pattern of the spatial distribution of the age structure can be perceived with some effort.



**Fig. 5.14.** The same four attributes as in Fig. 5.13 are represented here by means of “wheels” composed of four segments. The angular sizes of the segments are equal, and the radii portray the attribute values. The principle is the same as in the bar charts in Fig. 5.13: for each attribute, an individual encoding scale is used

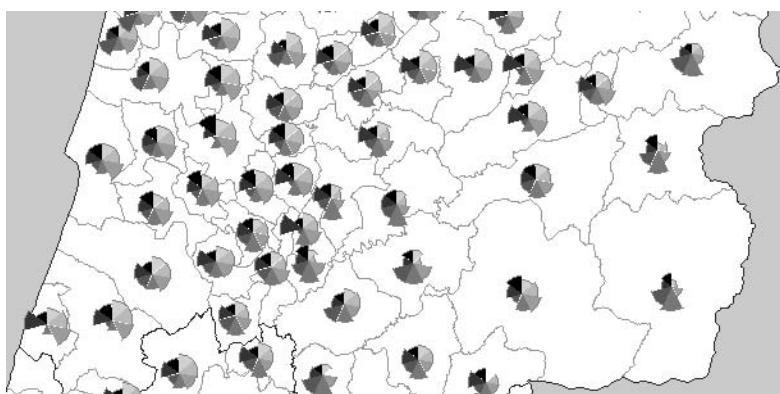
We have also tried another form of chart to represent the age structure. Figure 5.14 demonstrates a map of the same territory with the attributes represented by “wheels” or “wheel charts”. The principle of construction

of these charts is different from that of pie charts. Like a pie chart, each wheel is composed of four sectors, one sector per attribute. Unlike a pie chart, the angular sizes of the segments are equal, and the radii are used to portray the attribute values. As in the bar charts in Fig. 5.13, an individual encoding scale is used for each attribute.

Our personal impression is that the wheels are better suited for grasping general shapes and their spatial variation than are the bar charts. We do not so much attend to the differences between the sizes of the individual sectors as assess the shape of a wheel as “harmonious”, “nearly harmonious”, “distorted”, or “greatly distorted”. A shape close to round is perceived as harmonious, and all other shapes are judged according to their degree of deviation from roundness.

Our sight can cover rather large regions, noting where the shapes are predominantly harmonious and where they are mostly distorted, and what the major character of the distortion is. But it is not only the shapes that are grasped. When the sectors are coloured according to which attribute they represent, we also attend to the distribution of colours over the map and note the regions where the amounts of different colours are balanced and the regions where some colours prevail.

So, we can conclude that charts, at least certain types, can support the overall perception of the joint behaviour of several attributes. However, it can be predicted that increasing the number of attributes will make the task more difficult. Thus, from the map fragment in Fig. 5.15, one may consider whether it is easy to deal with seven-sector wheels.

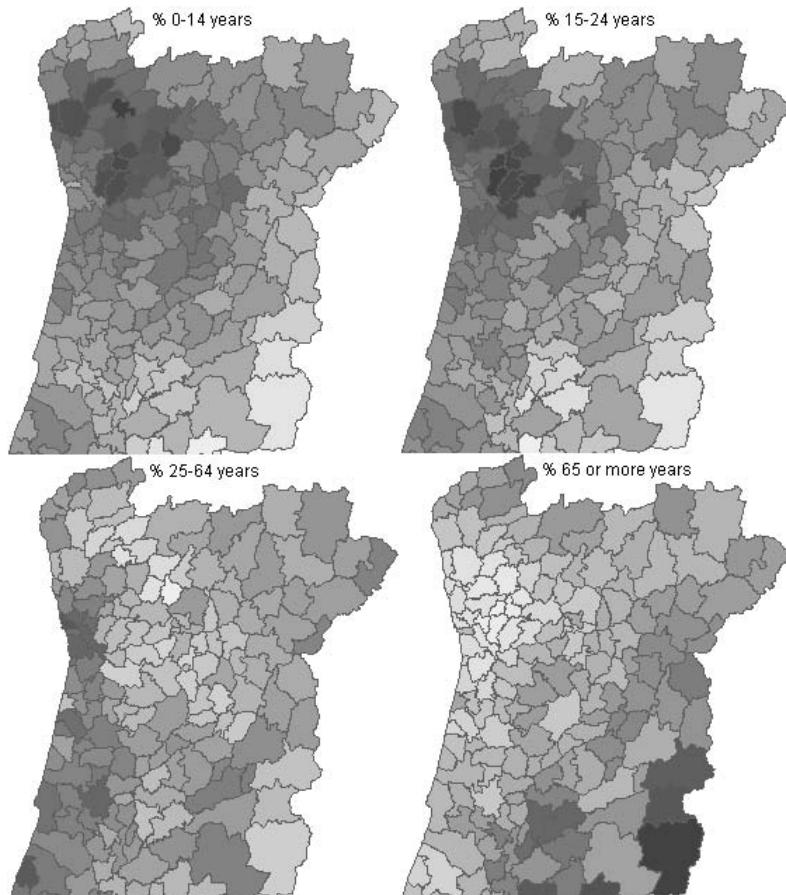


**Fig. 5.15.** Increasing the number of sectors in wheel charts makes them more difficult to perceive

Let us now look at the visualisation of the same four age structure attributes by means of four juxtaposed maps, or “small multiples” (Fig.

5.16). These are unclassified choropleth maps where the attribute values are encoded by degrees of darkness. For consistency with the chart maps considered earlier, we have cut the choropleth maps so that they represent the same part of the territory of Portugal.

Each individual map in Fig. 5.16 can be easily grasped as a single image, but do all the maps together promote the formation of a unified mental image of the joint behaviour of the four attributes, i.e. the variation of the age structure over the territory?



**Fig. 5.16.** The four age structure attributes are represented on separate maps here by means of area shading; darker shades correspond to higher values. Each map uses its individual value-encoding function

The four maps are not perceptually fused into a single image. The task of forming a unified pattern of the joint spatial distribution has to be per-

formed by means of multiple comparisons between the maps. We are not going to say that it is impossible to build an overall pattern in this way. Our point is that the multimap visualisation does not facilitate this task. This is not only our opinion. Thus, MacEachren refers to psychological studies showing that “vision is not particularly well suited to judging spatial correspondence between two variables represented on side-by-side maps” (MacEachren 1995, p. 402), and mentions a group of Earth scientists cooperating with his research team who rejected using “small multiples” for multivariate analysis and preferred composite maps instead.

Even Bertin, a well-known advocate of “small multiples”, admits certain deficiencies of this graphical construction and recommends using it in combination with a display containing all data components together. Thus, when there are three spatially referenced attributes, an array of three maps where each map represents a single attribute does not effectively support answering questions that address all three attributes simultaneously. Therefore, in order to respond efficiently to all types of question, two types of graphic are necessary: the three-map visualisation and a single map showing all three attributes together, for example by marks in which three retinal properties, such as size, colour, and orientation, vary (Bertin 1967/1983, pp. 154–155).

As we have said, deriving an overall pattern from “small multiples” requires numerous comparisons between the images. In such comparisons, an observer notes commonalities and differences between the images, but then complex synthetic work is required to proceed from these observations to a kind of overall pattern. The visualisation itself is not especially supportive of this work. In this connection, we would like to mention Arnheim’s argument concerning how an image is composed of its parts. One of the acknowledged rules of visual perception is the rule of similarity: things that resemble each other are tied together in vision. However, in most examples intended to show that similarity makes perceptual grouping, the effect is not created by similarity alone:

Arrange a number of chips, some white, some black, in a random order, and you will see them loosely related by color without any definite grouping; but let the white chips form a straight line or a circle, and their segregation from the black ones will be immediate and stable. That is, similarity will exert its unifying power only if the structure of the total pattern suggests the necessary relation. (Arnheim 1997, p. 55)

From this general observation, it may be deduced that noting similarities among multiple displays does not by itself stimulate unification. Some order or structure needs to be introduced for the similarity to start working. Therefore, it may be useful to provide the explorer with tools for rearrang-

ing multiple images, analogously to the technique of matrix permutation suggested by Bertin (Bertin 1967/1983, pp. 256–259). This pre-computer method of manipulating a display, which could be applied to data presented on cards, consists in reordering rows and/or columns of a matrix. The elements of such a matrix may be simple marks varying by size or colour, but they may also be more complex graphical constructions such as diagrams or maps. The idea of the technique is to move similar elements closer together by reordering the rows and/or columns until some prominent visual pattern emerges. The technique is also applicable to one-dimensional arrangements of multiple displays, such as an array of curves representing time-series data. MacEachren acknowledged the potential usefulness of this technique in application to multimap displays, noting its particular suitability for analysis of large numbers of attributes, probably a minimum of 16, for a  $4 \times 4$  matrix (MacEachren 1995, p. 403). It should be noted that it is hardly possible to construct an effective composite map that represents so many attributes simultaneously.

What other approaches to handling large numbers of attributes could be used besides reorderable “small multiples”? Computational tools could be applied to reveal groups of correlated attributes. Then, it would be possible to take a single representative attribute from each group and, in this way, reduce the number of attributes considered together. Another approach is clustering, which can group references according to the similarity of their characteristics in terms of multiple attributes. Clustering results can be represented as a single multicoloured image, such as the maps in Figs 4.120C– 4.123C. However, a problem with such an image is that it is very difficult to understand what characteristics each colour corresponds to. Therefore, it is necessary to use additional tools that can provide something like “portraits” of the clusters. In our examples, we have used collections of histograms and parallel-coordinates displays.

The requirement of promoting unification is relevant not only to attributes but also to referrers. We have touched upon the topic of unification several times in our discussion concerning completeness. Referrers are preferably mapped onto display dimensions, of which only the primary spatial dimensions effectively promote unification. Of the remaining dimensions, the display time, while not being ideal from the perspective of completeness, has more unifying power than the various possible arrangements. As Arnheim says, “the views that follow each other in the sequence are fused in such a way as to appear as states of one and the same persisting thing”, and the mind “is not limited to the view it receives at a given moment but is able to see the momentary as an integral part of a larger whole, which unfolds in a sequence” (Arnheim 1997, pp. 49–50). We have

already discussed some advantages and drawbacks of animated displays. Here, we would like only to say that animation is definitely recommended for use in combination with “small multiples”, especially when the display time represents a temporal or, more generally, a linearly ordered referrer (it is relevant to note that reordering of “small multiples” is not encouraged in such a situation, since the order is meaningful and cannot be changed arbitrarily).

The formation of an overall pattern is also greatly promoted by means of data and display simplification. Let us now proceed to the consideration of the second principle, “simplify and abstract”.

#### 5.4.2 Principle 2: Simplify and Abstract

We have already written rather much about simplification and abstraction. We have referred to Arnheim and the gestalt psychologists who characterise the process of human perception and cognition as a process of simplification and abstraction, of organising the stimulus material according to the simplest pattern compatible with it. We have also referred to Bertin, who saw the goal of information processing (i.e. exploratory data analysis) as discovering the synthetic schema which is at once the simplest and the most meaningful. In defining synoptic tasks, we have introduced the notion of a behaviour and the notion of a pattern as a parsimonious internal (mental) or external representation of a behaviour, i.e. a representation that is significantly simpler than an enumeration of all data items.

Besides acknowledging the role of simplification and abstraction, we have also pointed out that these processes can and should be supported by exploratory tools. In discussing various tools, we have described, whenever relevant, how these tools can promote simplification and abstraction. Let us recall what categories of tools have been characterised as suitable for this purpose.

In the section dealing with display manipulation, we discussed techniques leading to the simplification of data displays. Most of these techniques do not change the data but only the visual encoding of that data. The technique of display smoothing involves a data transformation and has therefore been mentioned both in the section on display manipulation and in the section on data manipulation.

Basically, display simplification may be achieved in two different ways:

1. The display is reorganised so that it *appears* simpler, while no information is hidden.
2. The display is simplified at the cost of reducing the information contained in it.

The first way is preferable but not always possible. According to Bertin, the only technique that can lead to display simplification without information loss is reordering of the graphical elements contained in the display. The basic idea is that reordering (or, more generally, arranging) may result in the elements being visually grouped into simple and interpretable shapes in accord with the gestalt principles of grouping by similarity, proximity, and good continuity. We have shown in Sect. 4.4.6 that the operation of visual comparison, which also preserves the entire information, can also lead to the same effect. This operation does not change the positions of the display items but manipulates their colour hues. Similarly coloured neighbouring items are perceptually associated into larger shapes, and this favours simplification and abstraction.

The other tools for display simplification involve information loss. Smoothing and other methods of graphical and cartographical generalisation eliminate excessive detail and random fluctuations. Classification is based on uniform representations of groups of characteristics, that is, different characteristics belonging to the same group become visually indistinguishable. In the resulting display, identical-looking neighbouring marks may be perceptually associated. Removal of outliers by means of focusing helps the explorer to concentrate his/her attention on the bulk of the data and abstract from the deviations that confuse the general picture.

In the application of tools involving information reduction, it is important to take care that the principal features of the behaviour under investigation are not hidden. Thus, we have demonstrated that an operation of smoothing applied to a time graph may eliminate a significant peak in the variation of the values and exhibit a gradual decrease instead. To avoid being misled, the explorer should “play” with the tools, i.e. change their parameters and observe how this affects the display. The tools, in turn, must be designed so as to facilitate such “playing”. It should be very easy to change tool settings, and the reaction of a tool to any change should occur immediately.

A great many data manipulation tools are also directed towards simplification. We have considered two classes of such tools: attribute integration, which reduces the number of attributes under analysis, and data aggregation, which reduces the number of references and, hence, the number of corresponding characteristics. These tools also involve significant information loss and need to be used cautiously. The same general approach, “playing” with tool parameters, may be recommended when one is using data reduction tools. There are also specific recommendations for data aggregation: take into account the characteristics of the value distribution (such as the range, variability, and presence of outliers) and prefer positional measures to statistical means.

When the necessary precautions are taken, data aggregation becomes a very powerful means of simplification and abstraction: it can provide a highly abstract view of an entire dataset and expose its most general characteristics in a compact form. Attribute integration is not so generally applicable. This technique is mostly domain-specific and requires domain knowledge to be applied meaningfully.

Query tools, in particular, filtering techniques, can simplify a display by removing some visual items. However, this is not the sort of simplification that is the focus of this section. This simplification does not promote abstraction and building of an overall pattern; on the contrary, abstraction tends to be hindered. We have discussed the fact that query tools are mostly intended for answering elementary questions. They are not directly suitable for synoptic tasks, since they do not support the abstraction processes which are necessary for such tasks. However, in combination with other tools, query tools may be used for synoptic tasks as well. As we have explained in Sect. 4.6.3, query tools in such a combination serve as means for extracting groups of data elements with similar characteristics, while the other tools, first of all visualisation, aid unification and abstraction. The role of query tools in this symbiosis corresponds to the next principle on our list, “divide and group”.

As to the computational tools, abstraction and simplification are their primary purpose. Both descriptive statistics and data mining aim at deriving a sort of “data essence”, characterising the multitude of particular instances as something unified, something that reveals the distinctive features of the dataset as a whole. Unlike the other tools for exploratory data analysis, computational tools simplify and abstract purely by exploiting the power of mathematics, without involving the abstraction capabilities of human analysts. The role of the latter is to interpret and verify the abstractions provided to them.

Is it always necessary to use aids to simplification and abstraction in exploratory data analysis? In principle, humans have an inherent capability and tendency to simplify and abstract whatever they perceive. In some cases, it may be sufficient to supply an analyst with an appropriate data display and leave the rest to these natural abstraction and simplification processes. However, this is only possible when the data are not too numerous and not too complex in their structure. Look, for example, at the display of earthquake occurrences in Fig. 4.80. Here, the data are too numerous to be managed without aids to simplification such as aggregation (some examples of the results of aggregation of these data are shown in Figs 4.81–4.85C). Another example is the multidimensional dataset containing the results of the modelling of forest development considered in Sect. 5.4.1.1. These data have a very complex structure, and therefore can-

not be analysed or even fully visualised without simplification (some approaches are demonstrated in Figs 5.8C–5.12C).

Many of the tools mentioned in this section produce an effect of simplification by means of grouping data elements and/or the corresponding display items. Hence, these tools are also relevant to the principle “divide and group”, which will be discussed next.

### 5.4.3 Principle 3: Divide and Group

We would like to begin this section by citing an article from the *Electronic Statistics Textbook* (StatSoft 2004). This article, which is titled “Categorizing, Grouping, Slicing, Drilling-down”, says:

One of the most important, general, and also powerful analytic methods involves dividing ("splitting") the data set into categories in order to compare the patterns of data between the resulting subsets. This common technique is known under a variety of terms (such as *breaking down, grouping, categorizing, splitting, slicing, drilling-down, or conditioning*) and it is used both in exploratory data analyses and hypothesis testing. For example: A positive relation between the age and the risk of a heart attack may be different in males and females (it may be stronger in males). A promising relation between taking a drug and a decrease of the cholesterol level may be present only in women with a low blood pressure and only in their thirties and forties. . .

There are many computational techniques that capitalize on grouping and that are designed to quantify the differences that the grouping will reveal (e.g., ANOVA/MANOVA). However, graphical techniques (such as categorized graphs) offer unique advantages that cannot be substituted by any computational method alone: they can reveal patterns that cannot be easily quantified (e.g., complex interactions, exceptions, anomalies) and they provide unique, multidimensional, global analytic perspectives to explore or mine the data. (StatSoft 2004, <http://www.statsoft.com/textbook/glosc.html>)

The importance of dividing and grouping for data analysis is widely recognised. In fact, the primary meaning of the term “analysis” is separation of a whole into its component parts, and it is hardly possible to do any analysis without separation. Therefore, we see no need for an additional argument in favour of dividing and grouping. Concerning the contents of this subsection, we have the following plan. First, we are going to show how the principle “divide and group” is related to our task framework. In parallel, we shall explain the difference and the relations between dividing and grouping and remind readers of how these tools can do this. In the next subsection, we shall speak about comparing the outcomes of dividing/grouping. In this connection, we shall touch upon the topic of combin-

nation of different tools, which can be done, in particular, on the basis of division of the reference set.

Dividing and grouping may be regarded as two sides of the same analytical process aimed at representing the overall behaviour of characteristics over the entire set of references by a compound pattern, i.e. a pattern composed of several subpatterns. Each subpattern approximates some part of the overall behaviour, which is based on a certain subset of the reference set. Each subset is a part of the entire reference set, i.e. the reference set, as well as the overall behaviour, is *divided*. On the other hand, each subset is a collection of individual references, which are considered together as a unified whole. Hence, the individual references are *grouped*, as well as the corresponding individual characteristics, which are united into the subpatterns.

As means of building compound patterns, dividing and grouping are relevant to the synoptic tasks of behaviour characterisation (pattern definition). However, these activities are also strongly related to other types of descriptive synoptic tasks, namely pattern search, pattern comparison, and relation-seeking, which usually appear as subtasks of the general task of characterising the overall behaviour (see the examples in Sect. 5.3). Thus, pattern search implies separation of a subset of references that are the base of a particular pattern from the remaining references. Pattern comparison is typically applied to the outcome of division/grouping but it is also involved in the process of division: an explorer may compare various subpatterns in order to decide how to better divide the overall behaviour. Relation-seeking takes place when the explorer looks for major changes in the behaviour from one part of the reference set to another.

There are various approaches to dividing/grouping:

1. By using domain knowledge (data semantics); for example, divide people into males and females or divide a territory into coastal and inland parts.
2. On the basis of data structure; for example, consider the route of each stork separately.
3. By applying a formal rule; for example, divide a time period into regular intervals.
4. On the basis of characteristics or behaviour variation; for example, group by similarity.

Examples of the first approach to grouping are given in the above article from the *Electronic Statistics Textbook*: “A positive relation between the age and the risk of a heart attack may be different in males and females (it may be stronger in males). A promising relation between taking a drug and a decrease of the cholesterol level may be present only in women with a

low blood pressure and only in their thirties and forties.” In these examples, the analyst expects that some differences will exist between the behaviours based on certain reference subsets: the subset of males versus the subset of females or the subset of females with specific characteristics versus the entire population. These expectations come from the domain knowledge of the analyst. This knowledge suggests to the analyst how it is meaningful to divide the reference set.

Examples concerning spatial and temporal data could be the division of a territory into coastal and inland parts (or into urban and rural areas, into mountains and flatland, etc.) and the division of a time period into working days and weekends (or into high, medium, and low tourist seasons, day and night time, etc.). In these examples, again some differences in the behaviours on these subsets may be expected on the basis of the analyst’s domain knowledge.

Division on the basis of data structure occurs when the data have two or more referential components, for example, spatial and temporal components. In such cases, the reference set and the overall behaviour are often divided into parts corresponding to different values of one of the referrers. Thus, one can consider the spatial behaviours of spatio-temporal data at different time moments, or one can look at the temporal behaviours in different places in the space. We applied such divisions to the dataset concerning the US crime data. In the example concerning the storks, we have a temporal referrer and a population referrer (i.e. the set of storks). We can choose a specific stork and consider its behaviour (i.e. movement) in time. We can also choose a specific time moment and consider the distribution of all the storks in space at this time moment.

Formal rules of division are frequently applied in classification and in data aggregation. Thus, in the classification of references according to values of a numeric attribute, it is customary to apply a division of the attribute value range into a specified number of equal-length intervals. Another widely used rule is to divide the data into equal-size reference groups, that is, the breaks in the value range are chosen so that the corresponding classes of references contain approximately equal numbers of elements. Demographers, who deal with reference sets consisting of districts of a territory, often choose to divide the whole set of districts into subsets with approximately equal total populations.

Space is often divided into regular compartments (cells). We used such a division in the examples where we aggregated the earthquake data (see Figs 4.81–4.83) and the data about the forest structures (Figs 4.86C and 4.87C). Division into regular intervals may be applied to a temporal component of the data. Thus, in the mosaic signs in Fig. 4.12 representing monthly temperature data, we divided the time period into years, i.e. 12-

month intervals. Then, we arranged the representations of the data for these periods so as to reveal the cyclic nature of the temperature changes. While this was an expected periodicity, it is also possible to apply the same approach to test whether a behaviour on some linearly ordered reference set is cyclic and, if so, to find the cycle length. In this case, an explorer iteratively divides the reference set into intervals of different length and looks to see whether the arrangement of the corresponding parts of the behaviour exposes any regular pattern. The interval length for which a regularity appears will be the cycle.

Probably the most exploratory approach, by its nature, is division on the basis of variation of characteristics or of behaviour. In this case, the analyst does not know in advance how the data will finally be divided. In order to decide how the data could best be divided/grouped, the analyst needs to observe the data and detect substantial differences in characteristics or in behaviour tendencies between parts of the data. In Sect. 5.3, we have discussed how the behaviour of a numeric attribute over time can be divided into fragments of increase, decrease, stability, or fluctuation. In this case, the analyst does not divide the time period into any predefined intervals. Instead, the analyst looks for certain expected types of patterns (i.e. performs pattern search tasks) and notes “turning points” where one pattern type changes to another (pattern comparison and relation-seeking). Or, perhaps, the analyst does not look specially for occurrences of particular pattern types such as an increase or a decrease but simply notes the heterogeneity of the behaviour and tries to divide it into relatively homogeneous parts. In any case, the analyst needs to detect the major changes in the character of the behaviour; hence, relation-seeking and pattern comparison tasks are involved. These considerations apply not only to numeric attributes but also to any time-referenced data. Thus, the migration of the storks (a time-referenced spatial attribute) can be divided into movement to the south and movement back to the north plus, possibly, some smaller movements in other directions. Another possibility is to divide the overall migration behaviour according to the speed of movement.

A similar procedure may be applied to spatially referenced data. An analyst may note a heterogeneity in the spatial behaviour and try to divide the entire space into regions with consistent internal behaviours. This procedure is known as regionalisation. It often takes the form of grouping locations or districts according to similarity of their characteristics in terms of one or more attributes. Grouping by similarity can also be applied to a population-type reference set, i.e. a set without ordering or distances. For example, students can be divided into groups according to their performance in various subjects.

Let us now review exploratory tools from the perspective of their relation to dividing and grouping. A tool that is relevant to dividing/grouping may enable dividing or grouping and/or deal with the results of dividing or grouping, which may be obtained by means of this tool or a different tool.

Among the tools for ordering and arranging, the matrix permutation technique suggested by Bertin enables grouping by similarity and, on this basis, division of the overall behaviour into diversified parts. Hence, the division/grouping is done on the basis of characteristics or of behaviour variation (approach 4). The tools for arranging periodic or supposedly periodic data are based on dividing a linearly ordered reference set into regular intervals, which is an application of a formal rule (approach 3).

The classification tools are meant primarily for dividing and grouping. There are tools that allow an explorer to define classes completely arbitrarily, according to any criteria that he/she finds to be appropriate. This supports, in particular, division/grouping on the basis of domain knowledge. For example, we may wish to divide the districts of Portugal into several geographical regions. For this purpose, we may use a sort of “manual classification” tool, which allows us to specify the desired number of classes (regions), give names to the classes and choose colours for them, and iteratively select districts and assign them to one of the classes.

Other classification tools divide references into classes according to the corresponding values of one or more attributes. Different techniques need to be applied depending on the types of the attributes, for example whether they are numeric, qualitative, spatial, or temporal, and on the number of attributes involved. In Sect. 4.4.3, we have considered examples of classification on the basis of a single numeric attribute (Figs 4.16, 4.18, and 4.19), according to the dominant attribute among several comparable attributes (Figs 4.21 and 4.22), cross-classification on the basis of two numeric attributes (Fig. 4.23C), and classification on the basis of a temporal attribute with the value range divided into seasons (Fig. 4.24). Of course, these are not all of the possible variants of classification.

In classification according to characteristics, various approaches can be supported. Most tools allow division/grouping by applying a formal rule. For example, the tool described earlier for classification on the basis of a single numeric attribute can automatically form a specified number of classes by dividing the value range of the attribute into equal-length intervals or by choosing breaks such that the classes have approximately equal sizes. It can also produce a statistically optimal classification, which minimises the variance within the classes and maximises the differences between the classes. Being linked to an enhanced cumulative-curve display (see Fig. 4.71), the tool allows the explorer to use classification rules involving other attributes. For example, the explorer can build classes with

equal populations or equal areas. A rather different example of a formal rule can be seen in classification according to the dominant attribute: the attribute with the highest value determines the class. The user may impose additional rules, for example that the situation where none of the attributes has at least 20% dominance over the attribute with the second highest value must be classified as “mixed”.

A classification tool can also allow the involvement of domain knowledge. Thus, the class breaks in a tool for classification on the basis of a numeric attribute may be chosen according to certain domain-specific criteria. For example, we once dealt with a dataset concerning soil pollution by pesticides. In the relevant domain, there is a standard set of breaks for pesticide concentrations. The resulting value intervals are designated as very low, low, medium, high, and very high concentrations. The tool for classification on the basis of a numeric attribute allows one to specify such standard breaks and use them for definition of classes. Another example of using data semantics can be seen in Figure 4.24, where forest fires are classified according to the season when they occurred into spring, summer, and winter fires.

Many classification tools also support exploratory classification, that is, they allow the analyst to interactively modify class definitions in the search for groupings that promote simplification and abstraction and help in grasping the distinctive features of the behaviour under study. An example is the classification of spatial references (districts or point locations) according to various non-spatial attributes so that the classes form coherent spatial regions. In order to note when this criterion is achieved, the explorer watches a map display where the classes are represented by means of colouring, for example. The map display should react dynamically to changes in the class specifications (e.g. moving class breaks in a classification according to a numeric attribute).

The tool for visual comparison on a map (see Figs 4.34C–4.36C) works in a similar way, i.e. the map reacts dynamically when the user changes the reference value for the comparison. The primary purpose of this tool is the same as in the exploratory classification of spatially referenced data: it supports the finding of groupings that produce simple, clear, easily interpretable spatial patterns (“good gestalt”). In principle, this idea can be generalised from maps to other displays where graphical elements can be coloured according to a classification. Thus, one may also look for a “good gestalt” in a scatterplot (we are not quite sure, however, whether a “good gestalt” can emerge in a parallel-coordinates display).

Since forming a “good gestalt” is an important criterion in exploratory classification, it is appropriate to think how a classification tool can promote this. Typically, classification does not change the positions of display

elements but changes only their colours. Hence, the gestalt principles of proximity, good continuity, closure, and symmetry cannot be consciously exploited. Classification allows identically coloured display elements to be grouped visually according to the principle of similarity, but this principle as such is rather weak and can work only in combination with other principles. However, it may be possible to apply the principle of figure/ground differentiation, in particular, through choosing the “right” colours.

It is beneficial for revealing the distinctive features of the behaviour under investigation when the visual items corresponding to the members of some class can be perceived as a unified figure. Therefore, it would be good if the colours used for the classes could help the differentiation into a figure and a background or at least did not impede such a differentiation. Thus, when all classes have equally bright colours, it may be difficult to perceive any of them as a unity, as a figure on top of a background formed by differently coloured visual items. In the result, the entire image appears too complex; no simplification is achieved. A possible solution would be to “mute” some classes, i.e. represent them by unsaturated colours so that they could produce the effect of a background. The remaining class(es) can then be perceived more easily as an integral image. This idea is demonstrated in Fig. 5.17C. A classification tool can help in perceiving patterns by allowing the user to temporarily “mute” or “switch off” the representation of some classes so that the user can check effectively whether the remaining classes form simple and understandable figures. Some of the classification tools that we know implement this feature.

Another approach is to represent different classes separately in parallel displays rather than all together in a single display. Examples of representation of classes in multiple parallel displays can be seen in Figs 4.20 and 4.24. The benefits of this approach are especially evident when the representation in a common display suffers from severe overlapping of graphical elements (compare, for example, Fig. 4.20 with Fig. 4.19). However, display multiplication can only be recommended when the number of classes is quite small.

The next big category of tools supporting division and grouping is that of data aggregation tools. Generally, the primary purpose of aggregation is not forming groups or partitions but rather data summarisation. However, for the summarisation, the results of grouping/dividing are needed. Therefore, aggregation tools typically include facilities to divide or group. In principle, aggregation tools may use the same methods for defining groups or reference set partitions as do classification tools. Moreover, aggregation tools can, in principle, be implemented in such a way that they use the output of any classification tool as their input and produce summaries of the classes specified. This is quite meaningful to do, since an explorer is usu-

ally interested not only in obtaining certain groups or partitions but also in seeing the characteristics of groups in a compressed form and comparing these characteristics. For example, having defined classes of the districts of Portugal according to their geographical positions, we may be interested to see whether these classes differ with respect to employment in various sectors of the economy. For this purpose, we need the values of the appropriate attributes related to individual districts to be summarised over the classes. We may use an appropriate computational tool to obtain statistics such as the minimum, maximum, median, quartiles, statistical mean (average), and standard deviation of the attribute values for the entire set of districts and for the classes.

An essential difference between classification and aggregation tools is that aggregation, unlike classification, “hides” individual data elements and treats groups as units. The groups and their collective characteristics, such as sizes and various statistics, are usually represented by means of special visualisation techniques. Therefore, aggregation tools are not so limited in the number of aggregates that they can produce and visualise as classification tools are in the number of different classes. Thus, the example maps in Figs 4.81–4.83, 4.85C–4.87C show simultaneously a great number of spatial aggregates. It is interesting that these aggregates may, in turn, be grouped visually on the basis of the gestalt principles of proximity, similarity, etc.

In discussing the possible methods of definition of classes in classification tools, we did not mention dividing or grouping on the basis of data structure (approach 2 in our list of four existing approaches), since this approach is not relevant to classification. However, in using aggregation tools, this method of grouping is often possible and quite useful. Thus, in the example of visualising the multidimensional forest management data (see Figs 5.8C and 5.9C), we used aggregation to reduce the dimensionality of the data. Specifically, we united together all references with different age groups but a common forest compartment, management strategy, time moment, and tree species. In Fig. 5.11C, aggregation has been applied to the entire spatial component of the crime dataset.

The next group of tools that are very relevant to dividing and grouping is that of query tools. Typically, a query divides data into at least two parts: data satisfying the query constraints and data not satisfying the constraints.<sup>30</sup> Some query tools produce finer divisions: when a query includes

---

<sup>30</sup> There are some classes of query tools that do not operate according to this principle. One class is that of direct manipulation tools that provide additional information when the user points on a display item (“What’s this?” queries). Another class includes tools for measuring metric relations such as distances.

several constraints, the data are divided according to how many constraints or which constraint combinations they satisfy. In fact, the outcome of such a tool is nothing other than as a classification of references according to characteristics, and can be visualised in the same way and used for the same purposes as other classifications.

Querying can be a rather powerful and flexible means of defining subsets of references. Most query tools allow the user to make use of domain knowledge in the formulation of query constraints. Thus, one can look for places and time moments where and when soil contamination by pesticides exceeds existing norms. With respect to time, query tools may allow the explorer to consider weekends separately from working days, or a period of growth of vegetation separately from a period of dormancy. With respect to space, one may use query tools to consider urban areas separately from rural areas or mountains separately from lowlands.

Generic query tools (i.e. with a functionality that is not restricted for the sake of dynamic responsiveness), as well as some specialised tools such as Time Wheel, temporal focusing, and temporal brushing, may be used for building queries and thereby defining subsets on the basis of the data structure. Thus, one can select crime data for all states for a particular year or, conversely, retrieve data for a particular state for the whole time period.

There are query tools suited for data selection by using formal rules. For example, Time Wheel allows the user to select data for a particular month in all years, or data referring to particular times of day over a period of several weeks.

Analogously to exploratory classification, where the analyst interactively modifies class definitions and looks for patterns on maps or other displays, exploratory querying is possible. In this case, the goal of the analyst is not to find data with particular characteristics but to find divisions that produce simple and meaningful figures in some visual display. As we noted in Sect. 4.6, query tools as such do not handle the reference subsets that they select as integral entities and do not treat the corresponding characteristics as unified behaviours. It is the job of a human analyst to consider these subsets as wholes, to grasp the characteristics holistically as behaviours, to compare the perceived patterns of these behaviours, and, eventually, to unite them into an appropriate pattern for the overall behaviour. This is the main idea of the exploratory approach to dividing/grouping. Dynamic query tools are especially suited for this purpose.

Among the computational tools, the clustering techniques of data mining are intended for the purposes of dividing/grouping. Clustering tools divide or group references on the basis of characteristics or of behaviour variation, i.e. this is a sort of exploratory grouping. The process is fully automatic but the user can modify the parameters of the clustering algo-

rithm used, which may result in significant changes in the results obtained. However, the user typically does not know in advance how the results will change in response to this or that parameter modification. The user usually “plays” with the parameters in the course of seeking simple and readily interpretable groupings/divisions. Appropriate visualisation of the results of the clustering is needed for the user to understand what unites the references within each cluster and differentiates them from the other clusters.

The characteristics of the various tools and tool categories relevant to dividing and grouping are summarised in Table 5.1.

**Table 5.1.** Tools and tool categories relevant to dividing and grouping

Tool and/or category	Produces division?	Uses division?	Division/grouping principle			
			Domain knowledge	Data structure	Formal rule	Exploratory (behaviour variation)
Interactive permutation (arrangement)	+					+
Periodic arrangement		+	+		+	+
Classification	+		+		+	+
Aggregation	+	+	+	+	+	+
Querying	+		+	+	+	+
Clustering	+					+

Whatever methods and tools are used for dividing or grouping, the outcomes are never regarded as the final result of exploration but rather as material for further analysis. The explorer always strives to understand the characteristic features of each reference subset resulting from the division. For this purpose, the explorer needs to *compare* characteristics of the subsets. As Arnheim states, “to see means to see in relation”; so, it is time to move gradually to the discussion of the next principle.

#### 5.4.4 Principle 4: See in Relation

We would like to start by citing again a statement from Arnheim’s book which was quoted in Sect. 3.7, since it is very relevant to the current topic of our discussion:

Experience indicates that it is easier to describe items in comparison with others than by themselves. This is so because the confrontation underscores the dimen-

sions by which the items can be compared and thereby sharpens the perception of these particular qualities. (Arnheim 1997, p. 63)

In the next few sentences, Arnheim warns that comparison entails certain dangers, which emerge when reference items for comparison are chosen arbitrarily. Thus, a comparison of the United States with China highlights characteristics quite different from the ones that can result from a comparison with France. We believe, however, that arbitrariness in exploratory data analysis can be diminished relatively easily. For example, in a comparison of results of dividing or grouping, an explorer deals with subsets of the same reference set defined in a consistent way by applying a common division/grouping procedure. If the explorer compares each subset with all the others, the factor of arbitrariness is excluded.

Let us now discuss how the results of dividing/grouping may be compared and then consider what other comparisons are useful in EDA and how they can be supported by existing tools.

Whether division/grouping is done interactively or automatically, the explorer uses a visual display to see the results. Thereby, the explorer not only notes how many groups there are and what elements each group comprises, but also compares the general characteristics of the groups visible in the display. Thus, when the analyst uses a map display, he/she compares the spatial positions and extents of the groups. When a parallel-coordinates display is the primary output medium for the results of the grouping/division, the analyst tries to grasp and compare the typical profiles of the groups and the variability of their characteristics, and to estimate the degree of separation or overlap between the characteristics of the elements of different groups.

However, for more comprehensive comparison and characterisation of subsets, the explorer needs to combine several tools. In Sect. 4.8, we considered reference set division as an instrument for the combined use and coordination of several analytical tools working sequentially or in parallel. Various tools can reflect the outcomes of division/grouping by means of highlighting, filtering, focusing, colouring, rearrangement of display items, or display multiplication. Data transformation and computational tools can take the results of the division as their input and perform further data processing on this basis.

Hence, on the one hand, reference set division is used as a means of tool combination; on the other hand, tool combination supports comprehensive examination and comparison of various characteristics of the subsets resulting from the division. By utilising tool combination, an explorer can obtain a many-sided view of any subset. He/she can:

- observe its spatial position, extent, and shape on a map;

- acquire a general idea concerning the distribution of the values of specific attributes over this subset from histograms or dot plots;
- grasp typical characteristic profiles from a parallel-coordinates display;
- look for possible correlations between values of different attributes in scatterplots; and
- obtain a summarised characterisation of the subset from an aggregation tool.

On this basis, characteristics of different subsets can be compared and distinctive features of each subset extracted.

For the comparison, it is often especially advisable to use the display multiplication technique, i.e. to represent each subset in a separate display and to look at several such displays in parallel. We have done this many times throughout this book. We have multiplied parallel-coordinates displays to compare results of classification (Fig. 4.20) and of clustering (Figs 4.126, 4.127, 4.132C, and 4.133C). We have represented the distribution of different classes of forest fires in several maps (Fig. 4.24). We used multiple box-and-whiskers plots to compare summarised characteristics of subsets (Figs 4.73 and 4.124). In fact, a multimap display representing spatio-temporal data referring to different time moments (see Figs 3.16, 4.46C, 5.5C, and 5.6C, referring to crime data, and Figs 4.5 and 4.6, referring to stork movement data) is nothing other than the display multiplication technique, used to represent different subsets (slices) of the reference set. The same applies to the visualisation of the forest management simulation data in Figs 5.8C–5.10C, where the data have been sliced and, correspondingly, the displays multiplied on the basis of the referential component “management scenario”.

The benefit of using multiple displays is that the representations of different subsets do not interfere with each other. As compared with a representation in the same display, no effort is needed for an explorer to separate one subset from another. The advantage of display multiplication is especially evident when representation in a common display results in serious overlapping of marks corresponding to different subsets. Besides the convenience of subset separation, display multiplication is better in helping the analyst to perceive each subset (and the corresponding characteristics) as a unit.

On the other hand, representation in a common display has its advantages as well: as a rule, it is more convenient for perceiving the characteristics of one subset in relation to those of another subset. For example, a common map display is convenient for the estimation of the relative spatial positions of groups of districts, spatial objects, or events. A common parallel-coordinates display shows better the commonalities and distinctions

between the characteristics of two or more subsets in terms of different attributes.

Since the advantages provided by the two approaches are complementary, it is reasonable to use both of them. Thus, a visualisation tool may allow the user to switch between the representation of several subsets in a single display and display multiplication. In the single-display mode, special techniques may be used to combat overlap. These techniques may be based on either aggregation or filtering.

With aggregation, the portraying of individual elements is replaced by the representation of summarised characteristics of subsets. An example can be seen in Fig. 4.79C: in a parallel-coordinates display, the lines corresponding to individual references have been replaced by elliptical shapes representing the deciles (i.e. the 10th, 20th, ... percentiles) of the attribute values for the entire reference set and for two subsets (classes) of references.

With filtering, the analyst may arbitrarily switch the representation of this or that subset on and off. For example, the parallel-coordinates display in Fig. 4.105 represents the elements of only two classes of a classification into five classes. Analogously, the parallel coordinates display in Fig. 4.107C represents three selected classes of four, and the displays in Fig. 4.108 and 4.109 represent a single class, with all other classes omitted.

In principle, aggregation or filtering can also be used in combination with multiple displays. Hence, there are a variety of possibilities for the synoptic comparison of collective characteristics of several reference subsets in terms of one or more attributes (the same process can be characterised, from a slightly different perspective, as the comparison of the partial behaviours of the attributes on these subsets). Table 5.2 summarises these possibilities.

Of course, it is not only characteristics of reference subsets resulting from division or grouping that may be compared. In our task framework, there are quite many subcategories of comparison tasks, both on the synoptic and on the elementary level. An important subcategory on the synoptic level is the comparison of behaviours of different attributes. In principle, the same two general approaches, display multiplication and representation of several behaviours in a common display, are applicable to such tasks.

In Sect. 5.3, we considered an example of comparison of the temporal behaviours of seven different crime rates in the same place. We represented these behaviours on time graphs and looked for similarities and differences (see Figs 5.2 and 5.3). As in the case of comparing characteristics of reference subsets, we used the technique of display multiplication: the behaviour of each attribute was represented in a separate display. We have mentioned that the behaviours could also be overlaid within a single time

**Table 5.2.** General techniques for the comparison of partial attribute behaviours on different reference subsets or of the collective characteristics of these subsets

	<b>Multiple displays, one display per subset/behaviour</b>	<b>Single display with all subsets/behaviours</b>
Technique itself	<p>The analyst grasps the general character of each partial behaviour and its distinctive features or the general characteristic profile of each subset, and compares the general patterns thus derived.</p> <p>Example: compare the spatial distributions (causes, severity, duration, etc.) of spring and summer forest fires.</p>	<p>The analyst estimates differences and detects overlaps between characteristics of different subsets or between major features of the behaviours.</p> <p>Example: are the areas of the highest concentration the same for spring and summer forest fires?</p>
+ Aggregation	<p>The analyst gets a high-level, summarised view of the partial behaviours or collective characteristics, and abstracts from details.</p> <p>Example: compare the spatial variations of the density of spring and summer forest fires (medians of the burnt area, most frequent causes, etc.)</p>	<p>Besides providing a summarised view, aggregation helps in reducing mark overlap in the display.</p> <p>Example: compare the contours of the density isolines (equal-value lines) of spring and summer forest fires.</p>
+ Dynamic filtering	<p>The analyst may focus on particular subsets or subranges of attribute values and locate these with respect to each reference subset and partial behaviour.</p> <p>Example: compare the distributions of the spring and summer forest fires that have the longest durations.</p>	<p>Besides focusing on particular value subsets or subranges, the analyst may switch on/off the representation of an entire subset/behaviour. This helps in reducing mark overlap.</p> <p>Example: where are the most severe spring fires with respect to the distribution of the summer fires?</p>

graph, but for this purpose the attributes need to be transformed in order to make their value ranges comparable.

We have also discussed earlier the fact that multiple juxtaposed displays (“small multiples”) do not promote unification, i.e. perception of the information contained in them as a single whole. Therefore, for the exploration of the joint behaviour of several attributes, it is desirable to represent them in a single display, for example as in the maps using cross-

classification and charts in Figs 5.12C–5.14. However, any display that favours unification of several attributes inevitably impedes or completely prevents the comparison of their individual behaviours. In contrast, multiple displays are very well suited to behaviour comparison tasks.

When a single display is intended to be used for behaviour comparison, a unification effect must be avoided so that the analyst can separate each behaviour perceptually from the others. A suitable approach is the use of a space-sharing arrangement, i.e. overlaying the representations of several behaviours in a common display, such as overlaying several lines in a time graph or overlaying several layers in a map display. Such a representation is not sufficiently powerful for producing a unification effect, and the behaviours of the different attributes can be distinguished and compared.

Mark overlap is a typical problem that arises when multiple behaviours are overlaid in a single display. This may be a serious obstacle to effective perception of the behaviours. To reduce overlap, aggregation and filtering are used, as in the previously considered case of comparing partial behaviours based on different reference subsets.

Analogously to the previous case, multiple displays and a single display with overlaid behaviours have different perceptual properties and provide different possibilities for analysis. Multiple displays are good for grasping the general character of each behaviour and its distinctive features. The analyst compares the overall holistic patterns resulting from this grasping. An overlaid representation is better for the detection of correspondences between the distinctive features of different behaviours. For example, with an overlaid time graph, the explorer can easily check whether similar features of different behaviours (e.g. an increasing or decreasing trend, a peak or a low point) occur in the same time interval or at the same moment.

Let us consider one more example. In Fig. 5.18C, three concurrent map displays represent three attributes from the dataset about forests in Europe, specifically, the percentage of coniferous forest, the percentage of broad-leaved forest, and the percentage of mixed forest. Recall that the data are specified in a raster format, i.e. the values of the attributes refer to cells of a regular grid with rather fine resolution. From the displays in Fig. 5.18C, we can grasp the general character of the spatial distribution of the values of each attribute. When we compare the maps, we see that the behaviours of the attributes are quite different, although there are some similar features, more precisely, clusters of high values with similar, rather characteristic shapes. It is hard to judge from the multiple maps whether these clusters are in exactly the same places and whether they have the same extent. To perform such estimations, it would be more convenient to have the behaviours represented in a common map display.

A map display is, in principle, suitable for an overlaid representation of the spatial behaviours of two or more attributes. These behaviours may form several map layers, drawn one on top of another. Since upper layers may cover lower layers, it is usually necessary to take special measures to ensure that all map layers are visible. One possibility is to represent one of the behaviours by means of area colouring, while another behaviour is represented by symbols or by isolines (i.e. lines connecting points that have equal attribute values). A disadvantage is that it may be difficult to detect similarities when so different representation methods are used. Another option is that a layer drawn on top of other layers is made semi-transparent so that the information beneath it remains visible. A disadvantage is that drawing in a semi-transparent mode distorts the colours in both the upper layer and the background layer(s). This is completely unacceptable when colour variation is meaningful, that is, it is used to encode data.

When a single tool is incapable of satisfying the analyst's needs, tool combination is often helpful. In particular, a combination of a multilayered map display with a dynamic filtering tool may quite adequately support the comparison of several spatial behaviours. The idea is that filtering is applied to each attribute represented in a separate map layer so that only specific selected values are shown in the map. These selected values are portrayed identically, for example using the same colour, while other colours represent selected values of the other attributes. The layers are drawn in a semi-transparent mode, and hence colour mixtures correspond to the places where the selected values of several attributes are present. The filtering tool allows the analyst to change dynamically the selection of values and, in this way, to investigate the behaviours for correspondence.

An example of the use of a multilayered representation of several attributes in a map in combination with a filtering tool is demonstrated in Fig. 5.19C. The same attributes as in Fig. 5.18C are shown in a single map as overlaid layers. The screenshots A–D from the map display correspond to different layer combinations (the map display tool allows one to switch the representation of any layer on and off): A, coniferous and broadleaved; B, coniferous and mixed; C, broadleaved and mixed; D, all three layers. In all the layers, small attribute values have been filtered out by means of a dynamic query tool. The query constraints were selected so as to make the characteristic features of each spatial behaviour well exposed.<sup>31</sup>

The values satisfying the respective query constraints are represented using a single colour for each query constraint. The colour correspondence is the same as in Fig. 5.18C: blue is used for coniferous forests, green for

---

<sup>31</sup> For this purpose, we introduced a lower limit of 9 for the percentage of broad-leaved forest and a lower limit of 3 for each of the other two attributes.

broadleaved, and red for mixed. Layers drawn on top of others are shown in a semi-transparent mode. Hence, colour mixtures occur where the values of two or three attributes satisfy the respective query constraints. Specifically, a turquoise blue colour corresponds to a mixture of coniferous and broadleaved forest, a purple colour to a mixture of coniferous and mixed forest, light brown (or dark orange) to a mixture of broadleaved and mixed forest, and dark brown to all three forest types together.

This overlaid representation of two or three behaviours in a common map display allows us to observe a quite good spatial coincidence of the distinctive features of these behaviours, specifically, the spatial clusters of relatively high values. Thus, from the apparent dark brown shapes in the screenshot D, we see that some clusters are common to all three attributes. We can also see that there is more commonality between the behaviours of coniferous and mixed forests than in the other attribute pairs.

From the discussion of the example visualisations in Fig. 5.18C and 5.19C, an observation can be made. It seems that the representation of behaviours in multiple displays supports direct behaviour comparison tasks better, while the overlaid representation in a single display is more suitable for inverse comparison tasks. Thus, with the multiple maps in Fig. 5.18C, we can answer the question, “What are the similarities and differences between these behaviours?” With any of the multilayered maps in Fig. 5.19C, we can answer another question, “Are the similar subpatterns of the different behaviours based on the same reference subsets?” A similar observation can be made from comparing multiple time graph displays of several time-referenced attributes and a single display with overlaid representations of the behaviours of all the attributes.

Let us also briefly discuss how behaviours of attributes referring to a (statistical) population, i.e. a discrete reference set without ordering or distances, may be compared. Such behaviours are usually characterised as statistical distributions and can be represented, for example, by frequency histograms, cumulative frequency curves, or graphs of the probability density functions. Hence, to compare the behaviours of different attributes, an analyst can look at multiple histograms or multiple graphs. Probability density graphs or cumulative curves of different attributes may also be overlaid in a single display. The two approaches differ in the same way as in the case of maps or time graphs.

When multiple displays are used for the representation of several behaviours, they may be manipulated consistently to allow different types of comparison. An example has been demonstrated in Fig. 4.141, where co-ordinated histogram displays represent the frequency distributions of the

values of the four age structure attributes in the Portuguese dataset.<sup>32</sup> At the beginning, each display has its individual horizontal and vertical scales. The horizontal scales represent the value ranges of the attributes, and the vertical scales represent the frequencies of the values. In this mode, the explorer can compare the general shapes of the histograms. The displays can be manipulated to have a common vertical and/or horizontal scale. A common vertical scale allows the analyst to compare bar sizes between different histograms. A common horizontal scale enables the comparison of the relative positions and sizes of the value ranges, as well as the relative positions of the most typical values of the attributes. As in the other examples considered above, transformation of the values of the attributes may make them easier to compare.

For the simultaneous manipulation of multiple map displays representing different attributes, it is usually required that the displays use a common visual encoding function. This, in turn, is possible when the attributes are comparable, i.e. qualitative attributes have identical or significantly overlapping value sets, or numeric attributes have close value ranges. For example, in Fig. 5.20C, we have represented three age structure attributes, “% 0–14 years”, “% 15–24 years”, and “% 65 or more years”, on three unclassified choropleth maps with a common function for encoding numeric values by colour shades (proportional degrees of darkness). This is different from the four maps in Fig. 5.16 having individual functions for value encoding. It is not occasional that we did not include the attribute “% 25–64 years” in the visualisation in Fig. 5.20C: its value range differs very much from the ranges of the three other attributes characterising the age structure.

In the lower part of Fig. 5.20C, we have demonstrated the effect of applying a common display manipulation tool, specifically “visual comparison”, to all three maps simultaneously. In all the maps, the reference value in the visual comparison operation is the same, and the colour encoding remains consistent. Such a simultaneous manipulation of multiple displays may help in noting similarities and differences between the behaviours.

However, it may be noted that the representation with a common encoding function in Fig. 5.20C does not expose the similarity of the behaviours of the attributes “% 0–14 years” and “% 15–24 years” so explicitly as the

---

<sup>32</sup> Although the referrer of this dataset is space rather than a statistical population, it is nonetheless quite valid to use frequency histograms or other techniques suitable for population-type referrers (e.g. scatterplots or parallel-coordinates displays). Such techniques do not take into account the spatial relations between the references but can be helpful in the exploration of various space-irrelevant features of the data.

visualisation with separate encoding functions in Fig. 5.16. This happens because the value ranges of the two attributes, although quite close, nevertheless differ: the values of the first attribute range from 11.13 to 27.5, and the values of the second attribute range from 8.82 to 21.32. While the general features of the behaviours of these attributes are very similar, the values in any district are different, and hence are encoded by different shades. Therefore, the maps of the two attributes look different, especially when a visual comparison operation is applied. Hence, it is not axiomatically recommendable to use multiple displays with a common encoding function for the visualisation of different attributes even when their value ranges are quite close.

On the other hand, it is very convenient to have the opportunity to manipulate several displays simultaneously. For example, a visual comparison operation can be very supportive for noting spatial patterns, and it is useful to be able to exploit this capability to compare spatial behaviours of several attributes. To make this possible, it is recommended that one transforms the attributes so that they become more comparable. We have mentioned such transformations in the Sect. 4.5.2. For attribute integration, it was important to ensure that all attributes had comparable value scales. The same idea is also applicable to the visualisation of multiple attributes. An example is demonstrated in Fig. 5.21C.

To produce this visualisation, we have transformed the values of the age structure attributes into  $z$ -scores, or standardised deviations from the respective means (see the formula (4.5) in Sect. 4.5.1.1). Note that this time we have included the attribute “% 25–64 years” in the visualisation: after the transformation, its value range has become quite comparable with the other value ranges. In the visualisation, the brown shades correspond to positive values, and the blue shades to negative values. Recall that positive values signify positive deviations from the mean (i.e. the original values are higher than the mean), and original values lower than the mean are transformed into negative  $z$ -scores. The display manipulation tool (“visual comparison”) used here allows the user to change the default midpoint of the diverging colour scale from 0 to any other value.

From the visualisation in Fig. 5.21C, the similarity of the behaviours of the attributes “% 0–14 years” and “% 15–24 years” is clearly visible, while differences are also easily detectable. The behaviour of the attribute “% 65 or more years” appears to be opposite to that of the attribute “% 15–24 years”. The behaviour of the attribute “% 25–64 years” is neither similar nor opposite to that of any other.

Not only the standard normal transformation (i.e. the transformation to  $z$ -scores) may be helpful in behaviour comparison, but also other methods

that transform absolute attribute values to relative values and ensure that multiple attributes have a common value scale. In particular, temporally referenced attributes may be transformed by computing relative changes (ratios or percentages) with respect to their values at a selected time moment, for example the beginning of the period that the data refer to.

When the data are multidimensional, i.e. have two or more referrers, comparison of behaviours becomes quite a difficult job, be it comparison of partial behaviours of the same attribute based on different reference subsets or comparison of behaviours of different attributes based on the same reference (sub)set. Multiple displays may be needed for the visualisation of a single behaviour, for example multiple maps for space- and time-referenced data. It may be a problem to have two or more collections of multiple displays simultaneously on the screen in order to compare two or more behaviours. Limitations are set, on the one hand, by the available screen size and resolution, and on the other hand, by the human perceptual capabilities. Displays that are too small may be not legible, and displays that are too numerous may cause cognitive overload or confusion to the user. Viewing two or more animated displays at the same time is also hardly productive, since it is impossible to pay equal attention simultaneously to all the displays and note similar and different developments over time effectively.

A more feasible strategy may be to consider the behaviours that need to be compared one by one. The analyst is expected to grasp the major features of one behaviour, store them in his/her mind or somehow note them on paper or in an electronic medium, and then try to detect the same features in another behaviour. After that, the explorer tries the same process the other way around. The analysis may require several iterations of the process.

It is sometimes possible to represent multiple behaviours in a common display or collection of displays. For example, if the data about European forests were time-referenced, we could apply the same solution as in Fig. 5.19C but the display would be animated, or we could construct several displays for different time moments and link them to the same filtering tool so that the constraints could be set simultaneously for all the displays. It should be remembered, however, that an overlaid representation of multiple behaviours in a single display provides somewhat different possibilities for exploration than a representation of these behaviours in multiple concurrent displays.

In the section dealing with the principle “see the whole”, we have spoken about the reduction of the dimensionality of the data, which is applied when there are not enough display dimensions and variables for an appropriate visualisation of all components of a multidimensional dataset. The

same approach can be used when it is necessary to compare the behaviours of several attributes on a multidimensional reference set. Recall that dimensionality may be reduced by means of selection or aggregation. Hence, multiple concurrent displays can represent several behaviours in an aggregated form or show selected slices of these behaviours. For example, multiple aggregated views like those in Fig. 5.11C could be used to represent the behaviours of different crime attributes.

Besides synoptic tasks of comparison of behaviours, there are also elementary tasks in which values of attributes or referrers are compared. Since elementary tasks, in general, play a less important role in exploratory data analysis than do synoptic tasks, we prefer to avoid a very detailed discussion of possible support for various kinds of elementary comparisons. So, we shall give only a few brief notes.

Relations between values of attributes or referrers can often be perceived quite well from a representation in a visual data display. Thus, from a representation of numeric values by horizontal or vertical positions or by symbol sizes, one may judge which of two values is greater. From a representation of values of a spatial attribute or referrer in a map, it is easy to see their relative spatial positions (e.g. one value is south of the other) and the distance between them.

Sometimes, when values are close, it may be hard to differentiate them only on the basis of their representation. Another problem is that an exact measurement of differences or distances may be impossible. The design principles developed for paper graphics and maps pay much attention to the accuracy of judgements of values from a visual representation. For example, it is recommended to prefer a representation of numeric values by positions within a display to a representation by symbol sizes since the values can be retrieved more accurately from positions than from sizes. The use of different colours or even different degrees of darkness of the same colour for numeric values is strongly discouraged.

Unlike paper graphics and maps, data displays on computer screens do not force an observer to judge and differentiate values only on the basis of their visual representation. When an accurate estimation is required, the observer may use query tools. There are query tools specifically intended for the measurement of distances (differences) between data items. However, not only these tools are appropriate for comparison tasks. Various dynamic query tools, including direct manipulation tools, which are especially convenient and time-efficient, can also be very helpful. One can use such a tool to retrieve exact values, which can then be compared.

Besides querying, some computational tools also support comparisons quite well, for example a tool for the computation of changes over time. The visual comparison tool discussed in Sect. 4.4.6 is suitable not only for

synoptic tasks (since it favours mark unification and pattern perception) but also for elementary comparison. Thus, one can compare the currently chosen reference value with any other value present in the display. While the tool does not show the exact differences, it is easy to find out which values are greater than the reference value and which are smaller.

It may be seen that elementary comparisons are done in a quite different way from synoptic comparisons and, naturally, require quite different tools. Let us now return to the synoptic level and move on to the next principle in our list.

We have mentioned in Sect. 5.4.3 that an explorer looks at the results of division represented in various displays in search of “good gestalt”, i.e. a simple and easily interpretable pattern. In so doing, the explorer bases his/her actions not only upon innate gestalt principles and aesthetic criteria but also upon his/her expectations of what sort of meaningful figures might be revealed. These expectations depend on the nature of the data and the underlying phenomenon and on the type of display used. In the next subsection, we shall discuss what tools can support looking for the expected patterns or subpatterns and how they can do this.

#### 5.4.5 Principle 5: Look for Recognisable

In the examples of data exploration given in Sect. 5.3, we demonstrated how pattern search tasks are involved in the analysis process. The basic idea is that an explorer does not simply *look at* data but *looks for* certain features of the behaviour. Thus, in an exploration of the dynamics of the burglary rate in California, an analyst may look for increasing and decreasing trends, periods of relative stability, and periods of intense fluctuation. The analyst expects in advance that some of these pattern types will be present in the data. These expectations are based on a knowledge of the nature of the data, specifically, that this is a time-referenced numeric attribute. Besides expecting that certain pattern types may be found in the data, the analyst also has an idea of what these patterns may look like in the type of display used for the exploration, specifically, a time graph. In a time graph, each pattern type appears as a line fragment with certain characteristics of the shape and slope. Hence, the task of the analyst is to divide the entire line into fragments with characteristic shapes identifiable as an increase, a decrease, etc.

A similar activity of detecting expected and interpretable patterns takes place in the exploration of spatial data (i.e. data that have a spatial referrer or a spatial attribute). Thus, when an explorer studies the spatial distribution of events such as crime incidents or disease occurrences, he/she looks

for areas of high concentrations of these events and, perhaps, for something similar to linear arrangements. The explorer may also expect a pattern such as an increasing or decreasing concentration in some direction, for example from north to south, from the coast to inland, or from the centre to the periphery. When the analyst studies the distribution of the values of a spatially referenced attribute, he/she looks for clusters of neighbouring locations or districts with the same or close values. If the attribute is ordered, in particular, a numeric attribute, another type of pattern that can be expected is a consistent change (an increase or decrease) in some direction.

The principle “look for recognisable” means that the tools for analysis need to be selected so as to allow the detection of the expected and meaningful types of patterns, which depend on the nature of the data under analysis. Not only the formal characteristics of the data (such as the number and types of the referrers and attributes) but also domain knowledge concerning the underlying phenomenon are important.

Let us compare, for example, a dataset containing occurrences of a disease that is typically caused by an infection or a contaminant and a dataset containing occurrences of earthquakes. Each dataset consists of a population-type referrer (a set of disease cases and a set of earthquakes, respectively) and two major attributes, space (where a disease case or earthquake occurred) and time (when the event took place). Hence, from a formal viewpoint, the datasets are identical; however, the expected patterns are quite different.

For the disease, an explorer may expect one or a few concentration clusters to appear, spread, shrink, and perhaps move over time. Nothing like this can be expected in the spatio-temporal distribution of earthquakes. At any selected time moment, no clusters typically exist; there will be either a single earthquake occurrence or no occurrences at all. Only when earthquakes that occurred over an extended time period are considered simultaneously will some concentrations perhaps be seen, in particular, around geological faults.

Concerning the behaviour of earthquakes over time, an expected and potentially interesting type of pattern is the occurrence of a series of earthquakes in the same or nearly the same place during a relatively short time period, say, from a few days to a few weeks. Let us call this pattern a “spatio-temporal cluster”. This is different from the major pattern type expected in the behaviour of disease cases, which can be described as “temporal development of a spatial cluster”, including formation, spreading, shrinking, and dissolving.

The definition of the pattern type actually contains a clue as to what kind of tool is needed to detect and observe patterns of this type. “Temporal development of a spatial cluster” suggests the use of an animated map

or a series of juxtaposed maps. In such visualisations, the entire time period that the data refer to is typically divided into short, regular intervals such as days or 10-day periods, depending on the timescale of the development of the underlying phenomenon. Each map state in an animation or each individual map in a series shows the events that occurred during a certain interval and allows an analyst to detect spatial clusters. If the events are numerous, the data on individual events can be transformed into a field (raster) of event density so as to make the clusters better visible. By viewing the sequence of states of the animated map or the sequence of juxtaposed maps, the analyst can observe how the clusters develop over time.

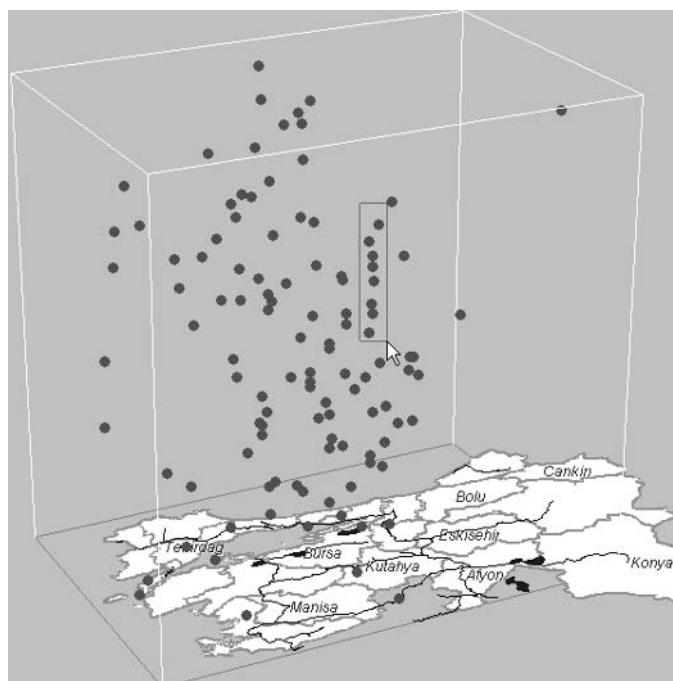
This visualisation, however, would not expose what is meant by “spatio-temporal cluster”. The definition of this pattern type suggests that space and time need to be viewed simultaneously. It may be suitable to use the three spatial display dimensions for this purpose: two dimensions represent space and the third dimension time. The visualisation may look as is shown in Fig. 5.22. In this perspective view, known as a space–time cube, the two horizontal dimensions represent the geographical space, and the vertical dimension the time. The events are represented by circular symbols placed within the cube according to the locations of the events in space and the times of their occurrence. In principle, this type of display may represent not only the times and places of event occurrences but also some characteristics of these events, by varying the size and/or colour of the circles. It is possible to use not only simple symbols such as circles but also charts, which can portray several characteristics of an event at once.

It should be noted that the space–time cube display does not allow us to see all the earthquake data simultaneously since the time period is rather long (24 years, from 1 January 1976 to 30 December 1999) and the earthquakes are too numerous (10 560 events in total). An attempt to represent all of the data at once results in tremendous overlap of the symbols. No patterns can be seen under such conditions. Therefore, it is necessary to apply focusing, and so we did this. The screenshot in Fig. 5.22 corresponds to the 100-day time period from 13 May to 20 August 1990, which was chosen using an appropriate temporal-focusing tool (a possible user interface is shown in Fig. 5.23). The display has been automatically adjusted to use the whole display height available to represent the selected time interval. The bottom of the display corresponds to the beginning of the interval, and the top to the end.

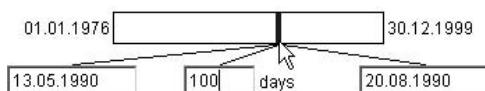
In a three-dimensional visualisation of the locations and dates of events such as earthquakes, a spatio-temporal cluster appears as a vertically aligned sequence of circles (or other symbols that are used for representation of events). One such sequence is enclosed in a frame in Fig. 5.22. It should be borne in mind, however, that visual grouping of symbols can

also be a mere projection effect resulting from the representation of a three-dimensional space on the two-dimensional computer screen. In order to verify the genuineness of this or that apparent grouping, the cube display must allow the user to change his/her viewing perspective.

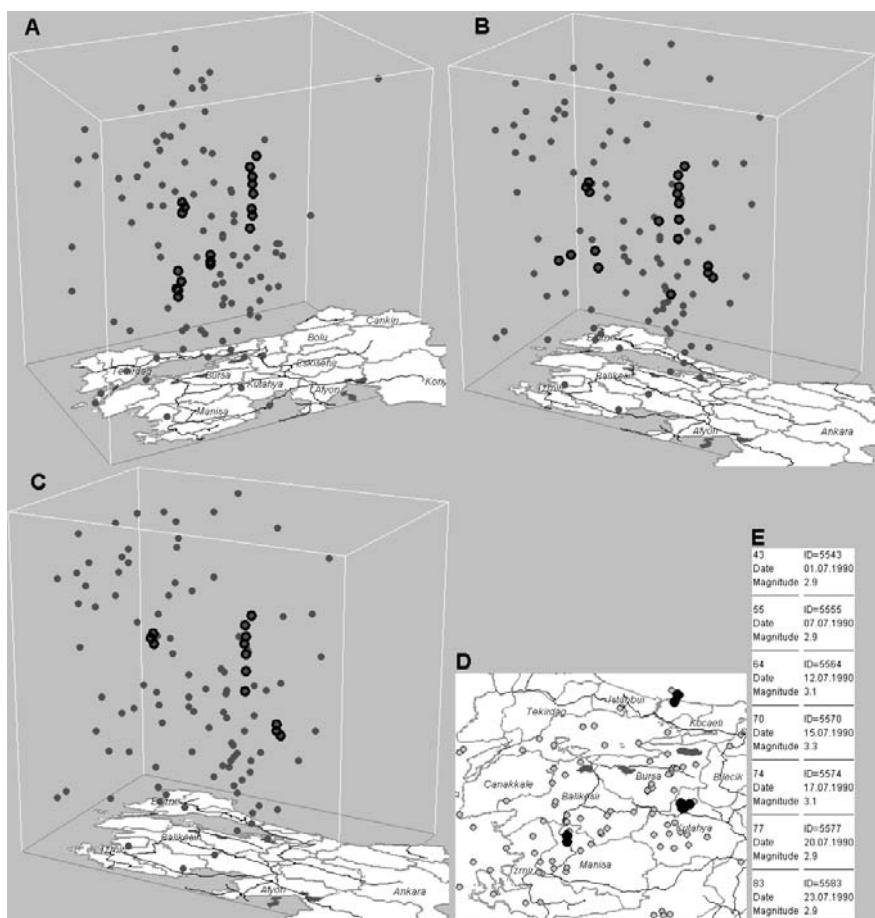
In Fig. 5.24A (upper left), we have used a direct-manipulation query tool to select several groups of vertically aligned symbols, which potentially indicate spatio-temporal clusters. The selected symbols are marked by thick black outlines. Then, we have changed the viewing perspective, i.e. rotated the cube. Figure 5.24B (upper right) shows the result: one group has dissipated, and two symbols have separated from two other groups.



**Fig. 5.22.** Earthquake occurrences are represented here in a perspective view (space–time cube). The horizontal dimensions represent the geographical space, and the vertical dimension the time



**Fig. 5.23.** The time interval represented in Fig. 5.22 was selected using an appropriate temporal focusing tool



**Fig. 5.24.** By changing the viewing perspective (rotating the cube), an observer can verify whether apparent alignments correspond to real clusters or result from projection effects

Using the same query tool, we have deselected the events that separated from the groups (Fig. 5.24C, lower left). Three groups remain. Two of these groups consist of three events each, and one includes seven events. From the relative positions of the symbols representing these seven events in the cube, it may be seen that the time intervals between the first and the second event and between the second and the third event are longer than the intervals between the subsequent events (recall that earlier events have lower positions in the cube).

Using a link between the cube and a map display, we have looked at the positions of the selected events on the map (Fig. 5.24D) and found that each of the three groups is compactly located in the geographical space

and, hence, can really be treated as a spatio-temporal cluster. The most interesting is, of course, the largest of the clusters, which consists of seven events; it is situated in the district of Kutahya on the east of the map fragment shown in Fig. 5.24D. The dates and magnitudes of these seven earthquakes can be seen in Fig. 5.24E. The first of the earthquakes in the sequence occurred on 1 July 1990 and the last on 23 July 1990. The time intervals between the earthquakes, in days, are 6, 5, 3, 2, 3, and 3. The magnitudes are 2.9, 2.9, 3.1, 3.3, 3.1, 2.9, and 2.9.

On the map, it may also be noted that there is one more earthquake located very close to the selected cluster of seven earthquakes (to the right of it). This earthquake, with a magnitude of 3.0, occurred on 24 June, a week before the sequence of seven earthquakes.

To our regret, we have no appropriate domain knowledge to judge whether the observations that we have made are meaningful; in particular, whether the earthquakes forming the clusters are related in some way. Nevertheless, this example allowed us to demonstrate how to look for spatio-temporal clusters of events using a space–time cube display, which seems to be a quite appropriate tool for this purpose. In general, any three-dimensional shapes that emerge in such a display owing to visual proximity of event symbols, for example, inclined chains or conical structures deserve attention as an indication of possible interactions between events. Imagine, for example, the exploration of a set of events in which durable clusters may be expected, such as the above-mentioned hypothetical dataset about disease occurrences (unfortunately, we have no appropriate data in reality). The clusters would appear as “clouds” in a space–time cube. However, taking into account the nature of the expected patterns, i.e. formation and development of spatial clusters, visualisation by means of an animated map or map sequence may be more convenient and productive.

Returning to the space–time cube display, we would like to mention that this tool is also suitable for the detection of expected patterns in the movement of objects in space when the objects are not too numerous. For example, we could use this tool for the data about the seasonal migration of storks. According to this technique, points in the three-dimensional space represent the positions of an object at different time moments. Lines connect the points corresponding to consecutive moments. In this representation, gently sloping path segments indicate fast movement, i.e. a long distance in space travelled in a short time, while steep segments correspond to slow movement. Vertical lines occur when an object stays for some time period in the same place. Readers interested in learning more about the space–time cube technique may be referred to Hägerstrand (1970), Hedley et al. (1999), Kraak (2003), and Gatalsky et al. (2004).

We have discussed how various types of patterns can be detected visually. Very often however, data do not readily allow this method of investigation; they may be too voluminous or too complex. Thus, in the example concerning earthquakes, we had to zoom into a 100-day interval within the overall 24-year period in order to be able to see the patterns. When all 10 560 earthquakes that occurred during the 24 years were represented in the cube, we could not see anything but myriads of overlapping circles. Although the selected interval (“time window”) can be shifted back and forth in time so that the entire period can eventually be surveyed, the pattern search job becomes very laborious and time-consuming. Similar problems arise when it is necessary to look for certain types of patterns among multiple lines drawn on a time graph. It would be very beneficial for analysts if pattern search could somehow be automated.

Unfortunately, there are not many computational tools capable of performing automatic pattern search. A probable reason is that each pattern type requires a specific method of search. We can imagine, for example, what type of tool could help in searching for patterns in earthquake occurrences. Such a tool would scan the sequence of earthquakes ordered according to the time of their occurrence. For each earthquake event, starting from the second one, the tool would check whether there was an event before it such that the distances between the two events in space and in time did not exceed certain user-specified thresholds, for example, 100 km in space and 5 days in time. If this condition was fulfilled, the tool would mark these events in some way as possible members of a spatio-temporal cluster. If the earlier event had already been marked, the later event would receive an identical mark and would thereby be attached to the previously constructed chain; otherwise, both events would receive a new, unique mark. At the end, the tool would retrieve the groups of events with identical marks including not less than some user-specified number of events. The results could be visualized in a space–time cube display without including the events that do not belong to any group. Such a tool would be helpful in detecting not only vertical alignments of events in the space–time continuum but also other structures such as inclined chains and cloud-like shapes. However, to our knowledge, no such tool exists yet.

Among the existing computational tools that we are aware of, there are tools capable of searching for a specified pattern of temporal behaviour among a collection of numeric time series data, such as the collection of local behaviours of the various crime rates in the states of the USA. Such tools can be categorised as query tools according to their function, but also as computational (data-mining) tools since they are based on rather intensive computation. We have mentioned these tools in Sect. 4.6.3, as well as some problems involved in judging the similarity between a user-specified

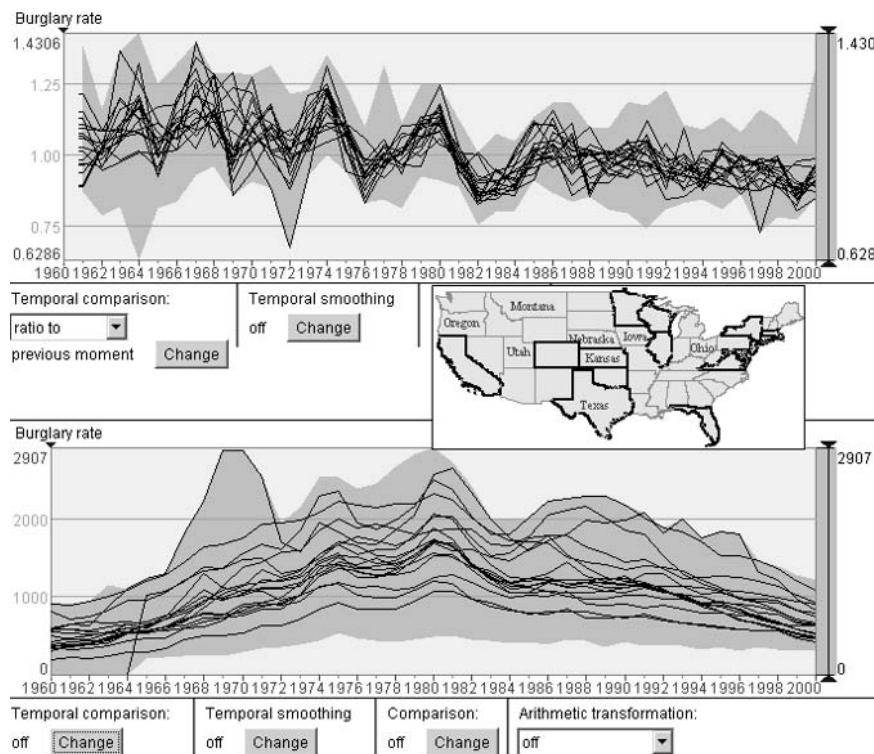
general shape of the pattern the user is looking for and a specific time series from the collection being scanned. Depending on what is understood by “similarity”, one can use a tool that simply computes the distance between the two lines and compares it with a specified threshold or a sophisticated, computationally intensive algorithm of “dynamic time warping”, which smoothes, stretches, or shrinks each line in the collection to bring it into the maximum possible correspondence with the model shape (see the review in Keogh and Kasetty (2003)).

In Sect. 4.6.3, we have also described some visual query tools that support a search for particular patterns among multiple lines on a time graph. One of these tools selects lines according to a user-specified search mask. The other tool shows line fragments that have a specified inclination.

To some extent, various data transformations, in combination with query tools, may be helpful in searching for patterns. Thus, values of a time-referenced numeric attribute may be transformed into differences or ratios with respect to the previous time moment. Then, if we need to detect increasing trends over sequences of time moments, we should look for sequences of positive values in the case of differences or values greater than 1 in the case of ratios. To detect decreasing trends, conversely, we look for negative values or values below 1, respectively.

An example can be seen in Fig. 5.25, where a transformation of the original values into ratios with respect to the previous time moments has been applied to the attribute “Burglary rate” in the dataset concerning the crime statistics over USA. The transformed data are represented in a time graph at the top of Fig. 5.25. Using a direct-manipulation query tool, we have selected the lines with transformed values in the last five years of the period covered by the dataset (i.e. from 1996 to 2000) below 1. This corresponds to a decreasing trend over this interval, as can be seen from the time graph at the bottom, where the transformation of the values has been cancelled, and the original shapes of the lines can be seen. In both screenshots of the time graph display, only the selected lines are visible, and the general outlines (“envelopes”) of all the lines, while the lines not satisfying the query are hidden. In the middle, we have shown a map in which thick black boundaries mark the states that the selected lines correspond to.

It may be noticed that the decreasing trends in many of the selected lines started earlier than 1996. Using the transformed time graph, we can modify the query constraint and look for longer decreasing trends over the last years of the dataset. Thus, Fig. 5.26 shows us where decreasing trends took place over the period 1992–2000 (top) and where they were observed as early as 1990 and continued until 2000. The same transformation allows us to look not simply for decreasing trends but for decreasing trends with particular rates of decrease. For example, to find where the decrease is

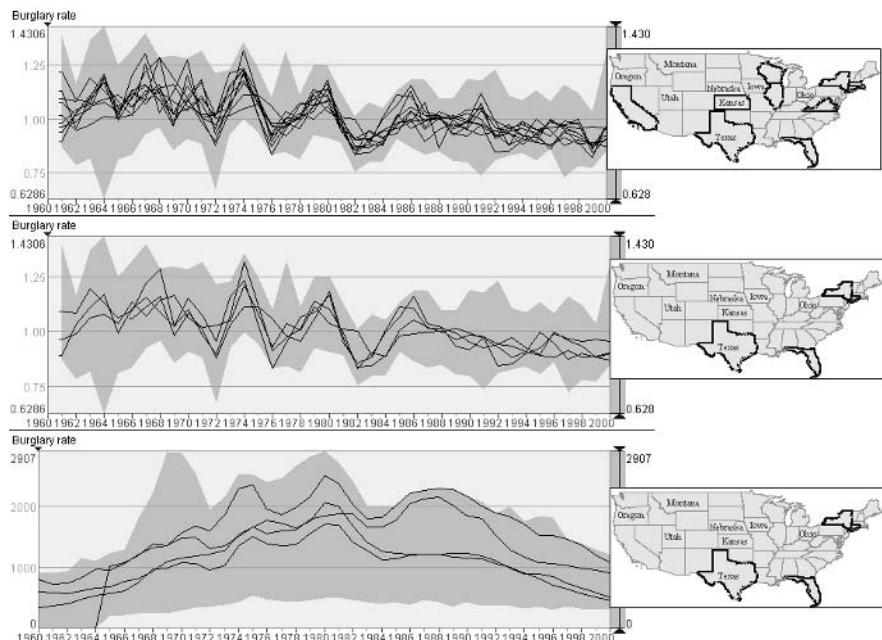


**Fig. 5.25.** A transformed time graph showing differences or ratios between attribute values at consecutive time moments is suitable for searching for increasing or decreasing trends in specific time intervals. Here, we have selected the lines that have a decreasing trend during the last five years of the period, i.e. from 1996 to 2000. At the top, the time graph is shown in the transformed mode; at the bottom, the transformation has been switched off, and the original shapes of the selected lines are visible. In the map in the middle, thick black boundaries mark the states that correspond to

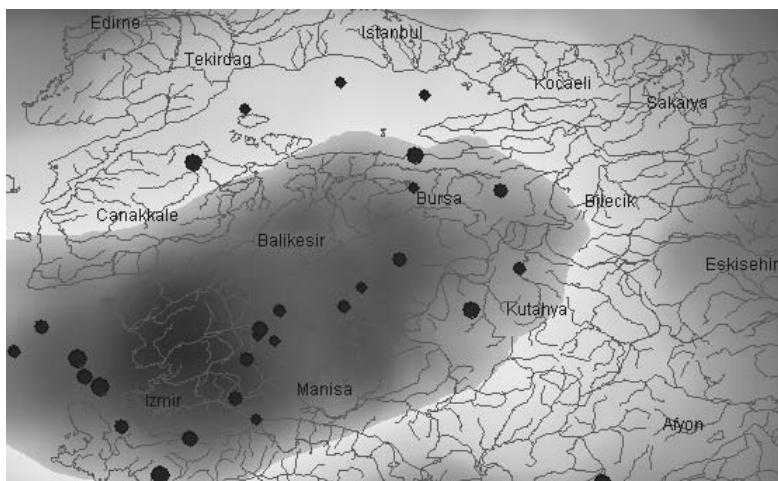
10% or more, we need to look for transformed values that are not higher than 0.9.

Similarly to the use of computed changes in time for the detection of particular temporal trends by means of a query tool, it is possible to compute changes between locations in space (see, for example, Fig. 4.51) and apply a query tool to the output in search of various spatial trends. Data about spatially dispersed objects, events, or movements can be transformed into spatial densities, and an analyst can detect clusters of high concentrations of the objects, events, or movements by applying a query tool to the density or just by viewing a visualisation of the density. For example, the map in Fig. 5.27 represents the earthquake density computed from the data

about individual earthquakes. Densities are portrayed by means of background colouring. The dark spot covering the districts of Izmir, Balikesir, Bursa, Kutahya, and Manisa is the area of high earthquake density, and the area with the darkest shading, near Izmir, is where the density is the highest. In the original map, we used a diverging colour scale with varying degrees of darkness of red and green colours so that the dark spot was originally red and the territory around it green. Hence, data transformation plus visualisation plus display manipulation (specifically, visual comparison) allowed us to detect a concentration cluster (a kind of association pattern), as well as a spatial trend of decreasing earthquake density in the outward direction from the area near Izmir. We can also add some detail to the characterisation of this spatial trend, in particular, that the decrease rate is higher in the latitudinal than in the longitudinal direction.



**Fig. 5.26.** Using the transformed time graph, we can probe how long the detected decreasing trends lasted in different states. The time graph and map at the top show the lines and the corresponding states for which a decreasing trend was observed over the last nine years starting from 1992, and the next two pairs of screenshots show the decreasing trends over 11 years, starting from 1990. The time graph at the bottom is the version of the time graph in the middle obtained when the transformation is switched off



**Fig. 5.27.** This map represents the density of earthquakes over the territory of Turkey by background colouring. The area near Izmir, with the darkest shading, corresponds to the highest earthquake density. The circles on top of the shading show the locations of the strongest earthquakes (with magnitudes of 5 or higher). It may be seen that the strongest earthquakes occur outside the area of the highest earthquake density and sometimes even quite far from it

Generally, after some feature (subpattern) has been detected in a behaviour, an explorer often wishes to examine it in more detail. Similarly, when a behaviour has been divided into parts, according to the principle “divide and group”, it may be necessary to take a closer look at the parts obtained. For this purpose, the explorer usually needs to zoom and focus, and hence the tools that he/she uses must support these operations. The importance of these operations is stressed in Ben Shneiderman’s Information Seeking Mantra: “Overview first, zoom and filter, and then details-on-demand”.

#### 5.4.6 Principle 6: Zoom and Focus

The purpose of zooming and focusing is twofold:

1. To visualise a selected part of the data with the maximum possible expressiveness so that more detail is visible and more differences are detectable.
2. To disregard the remaining data so that they do not distract the explorer from the portion of data of interest.

We have discussed the tools for zooming and focusing in a dedicated section (Sect. 4.4.4) and shall not repeat ourselves. We would like to note,

however, that not only specific zooming and focusing tools are suitable for this purpose. Other tool types may sometimes be applicable and even more appropriate than specialised devices, especially with regard to the second purpose above. Thus, it may be impossible to focus, by means of only zooming and focusing operations, on a particular group of lines on a time graph or parallel-coordinates display and disregard the other lines. A query tool of the filtering type is in this case more suitable for this purpose. It can be noted that the Information Seeking Mantra cited above mentions filtering rather than focusing.

Filtering not only helps an analyst to get rid of distracting display items but can also increase the expressiveness and legibility of the display, which is achieved at the cost of reducing mark overlap rather than by transforming the display scale or the visual encoding function used in it.

Sometimes, one and the same query tool may be used either in a marking or in a filtering style. This is possible in a display that enables a special, optionally used visualisation mode showing only selected data items. The time graph tool demonstrated in Figs 5.25 and 5.26 offers such a possibility, which we utilised for producing the illustrations. This tool can be easily switched back to the “normal” mode, where all the lines are visible.

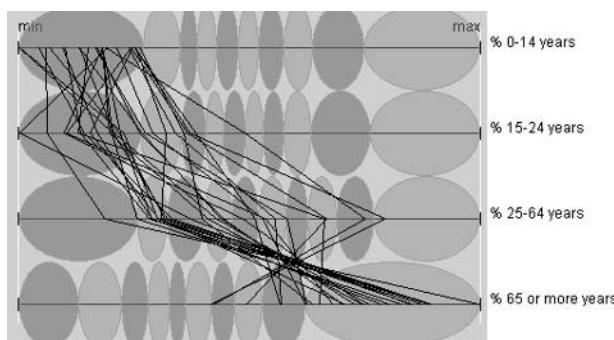
Figures 5.25 and 5.26 demonstrate also the concept known as “focus plus context”, which has recently received much attention in the area of information visualisation (see, for example, Spence (2001)). The idea is that a visualisation tool allows the user to focus on part of information, while the remaining information is not entirely removed from the display. Instead, it is shown in a reduced or generalised form and provides the “context” for the portion of information of interest, that is, it shows the position of this portion with respect to the entire data collection. In Figs 5.25 and 5.26, only a selected part of the data can be viewed in detail as lines on a time graph. However, the time graph also shows the general outline, or “envelope”, of the whole collection of lines. The outline provides a kind of context for the selected data, which allows the analyst to judge how the values for the selected states at different time moments are positioned with respect to the minimum and maximum values attained at these moments over the whole country. A similar example can be seen in Fig. 4.76, where a line for a selected state is drawn upon a background formed by an aggregated representation of the remaining data where not only the minima and maxima are present but also the medians and quartiles.

These examples, however, cannot be qualified as classical “focus plus context” display techniques. It is more typical to relate this concept to zooming, i.e. enlargement of some display items at the cost of other items. “Focus plus context” zooming tools increase the sizes of selected items and reduce the sizes of the remaining items. The best-known examples of

such techniques are Fisheye View (Furnas 1986) and Perspective Wall (Mackinlay et al. 1991).

In our opinion, the “focus plus context” concept in its classical sense has more to do with navigation than with exploratory data analysis. However, if it is treated in a broader sense, this concept also subsumes such approaches as the combination of an aggregated view of the whole dataset with a detailed representation of a selected portion of the data, which can be very useful for data exploration.

We have demonstrated the combination of aggregated and detailed views with an example of a time graph display. Other display types can also be modified to allow this. Thus, in Figs 4.77–4.79C, we have shown a modification of the parallel-coordinates technique that represents aggregated characteristics of a dataset rather than individual items of data. This representation can easily be combined with portraying individual characteristics corresponding to a selected subset of references. For example, the display in Fig. 5.28 represents the data on the age structure of the population in the districts of Portugal in an aggregated form: the elliptical shapes reflect the relative positions of the deciles (i.e. the 10%, 20%, ..., 90% percentiles) of the attribute values in the entire dataset. In addition to this, there are lines portraying the individual age structure characteristics of a selected subset of the districts, specifically, the districts with the lowest 10% of the values of the attribute “% 0–14 years”. This combination allows the selected districts to be easily positioned among the remaining districts with respect to their age structure.



**Fig. 5.28.** This parallel-coordinates display represents aggregated characteristics of a whole dataset together with individual attribute values for a selected subset of references

Similarly to the case of the parallel-coordinates visualisation technique, there is a modification of the scatterplot display, the binned scatterplot, which represents data in an aggregated form (see Fig. 4.64). Such a display

can be combined with the representation of a selected subset of individual data items. For this purpose, it is better to use the variant of the binned scatterplot shown on the right in Fig. 4.64: this can serve as a background for individual dots just as the ellipses in Fig. 5.28 serve as a background for individual lines.

It may well be that any visualisation technique suited to the representation of individual data items is, in principle, extendable to the combined display of aggregated characteristics and selected individual data items. In contrast, the techniques specifically intended for the representation of aggregated information, such as histograms, cumulative curves, treemaps, and mosaic plots, cannot be so easily modified to include the representation of individual data.

While the combined display of aggregated data and selected individual items can be considered as a specific realisation of the concept “focus plus context” (treated in a broad sense), there are a few other concepts in information technology that are related to focusing and at the same time to aggregation. Thus, “drill down” means moving from summary information to detailed data by focusing in on some part of the data. “Slice and dice” implies a systematic reduction of a body of data into smaller parts or views that will yield more information. This term is also used to mean the presentation of information in a variety of different and useful ways.

For an analyst, the most convenient way of drilling down or slicing and dicing is through direct manipulation of a display that provides aggregated information. For example, when the analyst clicks on a display item representing a data aggregate, the item may “expand” to show more detailed information, which may be the individual data items included in this aggregate or the data on a lower aggregation level. Many of the existing software systems for exploratory data analysis provide this possibility. For example, Fredrikson et al. (1999) describe a system that applies various methods of aggregation to data about traffic incidents (spatial aggregation by road fragments; temporal aggregation by dates; times of the day, or days of the week; and categorical aggregation by event types) and displays the data in an aggregated form. However, the user can drill down into any aggregate by clicking on a display item representing this aggregate. The detailed data about the events included in the aggregate are then shown in an additional window.

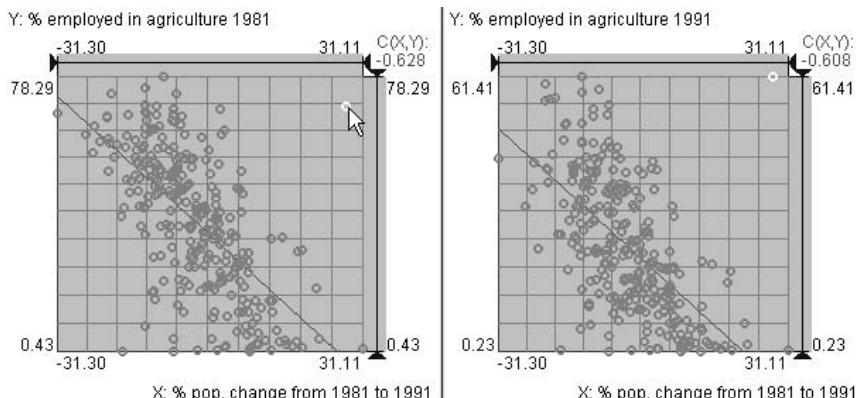
Zooming and focusing may be viewed as a bridge from the overall level of analysis to a detailed consideration of specific places, times, and individuals. The next subsection deals with access to specific data items. We say briefly why an analyst may need this and what tools are suitable for this purpose.

### 5.4.7 Principle 7: Attend to Particulars

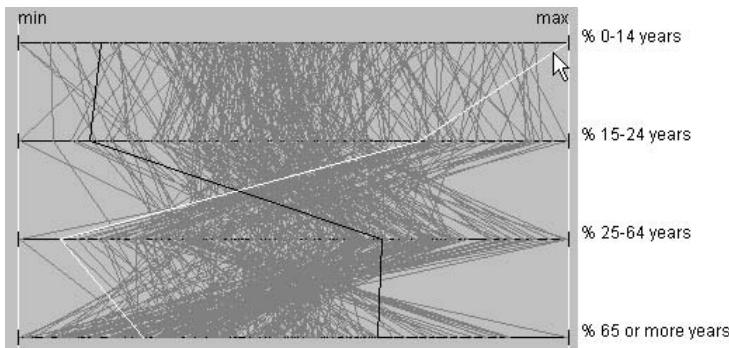
In a data display, some visual elements may immediately attract the viewer's attention because they look "strange", being substantially dissimilar to all others or to their neighbours. Such outstanding display elements correspond to outliers in the data, i.e. extraordinary attribute values or value combinations. For example, among the values of a numeric attribute, there may be one or a few extremely high or extremely low values standing far apart from the bulk of the data, such as the extremely high population densities in three districts of Portugal (see Fig. 4.27). An example of "local" outliers, i.e. unusual attribute values in comparison with the values in their neighbourhood, is the burglary rates in the District of Columbia in the years 1969 and 1970, which amount to 2869.9 and 2873.7, respectively, while the remaining values in these two years range from 249.3 and 286.4 in North Dakota to 1676.1 and 1753 in California. On a time graph (for example, in Fig. 4.3), the two outstanding values appear as a high peak against the positions on the horizontal axis corresponding to the years 1969 and 1970. Nevertheless, these are not the highest two values in the entire dataset. The highest burglary rate over the country, 2906.7, was attained in Nevada in 1980. However, the peak for the year 1980 does not look so prominent as the one for 1969–1970, since quite many states also had very high burglary rates in 1980. The maximum value attained in Nevada is not so far from the values of 2559.7 in the District of Columbia, 2506.8 in Florida, 2316.5 in California, and so on.

Examples of unusual value combinations (more precisely, examples of how such combinations appear on scatterplots or parallel-coordinates displays) can be seen in Figs 5.29 and 5.30. In Fig. 5.29, we see that the proportion of people in the districts of Portugal employed in agriculture in both 1981 and 1991 is negatively correlated with the relative change in the population from 1981 to 1991. However, each scatterplot has a dot in the upper right corner indicating that some district that had a high proportion of agricultural employees also had a significant population increase from 1981 to 1991. This is rather untypical, and the relative positions of these dots (which are marked in white) with respect to the remaining dots clearly manifest the uniqueness of the respective value combinations. By the way, the marked dots in both scatterplots correspond to the same district, Sao Joao da Pesqueira, with a population change from 1981 to 1991 equal to 27.76% and the proportion of people employed in agriculture equal to 69.89% in 1981 and 61.41% in 1991.

Figure 5.30 demonstrates that atypical value combinations manifest themselves in a parallel-coordinates display as line segments that differ in their inclination from the surrounding lines.



**Fig. 5.29.** An unusual combination of values of two numeric attributes appears on a scatterplot as a dot standing apart from the main mass of dots. The mouse cursor points at such a dot in the display on the left. This dot corresponds to a district of Portugal with a high population growth from 1981 to 1991 and a high proportion of people employed in agriculture in 1981. In the scatterplot on the right, the dot corresponding to the same district is highlighted (shown in white). It is situated in the upper right corner and indicates that the proportion of agricultural employees in this district was very high in 1991 as well



**Fig. 5.30.** In a parallel-coordinates display, line segments differing in their inclinations from the surrounding lines indicate atypical combinations of attribute values. Thus, the black line has an “unusually vertical” segment between the axes for the attributes “% 25–64 years” and “% 65 or more years”. The white line is “unusually oblique” between the axes for “% 0–14 years” and “% 15–24 years”

Of the two highlighted lines in Fig. 5.30, the black one has a segment between the axes for the attributes “% 25–64 years” and “% 65 or more years” that is unusually close to vertical. The corresponding district, Aljezur, is characterised by quite high proportions of people in the age groups 25–64 years and 65 or more years (50.68% and 25.33%, respectively),

whereas the districts with values of the attribute “% 25–64 years” similar to that for Aljezur usually have quite a low proportion of elderly people and, vice versa, the districts with a proportion of elderly people close to that for Aljezur have a lower percentage of people aged from 25 to 64 years. The white line has an atypically slanting segment between the axes “% 0–14 years” and “% 15–24 years”. The corresponding district, Povoa de Lanhoso, has a very high proportion of children from 0 to 14 years old (27.5%). While the other districts that have a high proportion of children also have a high proportion of young people (i.e. aged from 15 to 24 years), the proportion of young people in Povoa de Lanhoso is, unusually, not very high: 17.94%, which is not very close to the maximum of 21.32%.

In the examples given above, we have provided quite detailed and precise information about the outstanding attribute values and value combinations, as well as the corresponding references, that is, states and years in the case of the crime data and districts in the case of the Portuguese census data. This information resulted from a number of elementary tasks that we performed:

- Relation-seeking: Find references with characteristics differing greatly from those of the other references.
- Direct lookup: What are the values of this or that attribute corresponding to this or that reference?
- Inverse lookup: What reference corresponds to this (unusual) attribute value or combination?
- Comparison: Compare the characteristics corresponding to this reference with the characteristics of other references

We have said earlier that elementary tasks play a marginal role in exploratory data analysis. This does not mean, however, that elementary tasks do not emerge at all or that they can be skipped without any harm to the analysis process. The existence of global and local outliers and unusual value combinations is a classic case where elementary tasks necessarily arise: the analyst does need to pay attention to any “strange” thing present in the data that is exposed by visualisation. The analyst needs to understand whether this strangeness signals an error in the data. If not, the analyst will seek an explanation of the odd thing detected. This may be not easy: “Strange things sometimes require years of thinking before it becomes clear how they fit into our picture of the world” (O.Chertov, personal communication). Nevertheless, the strangeness cannot be ignored; it demands proper attention.

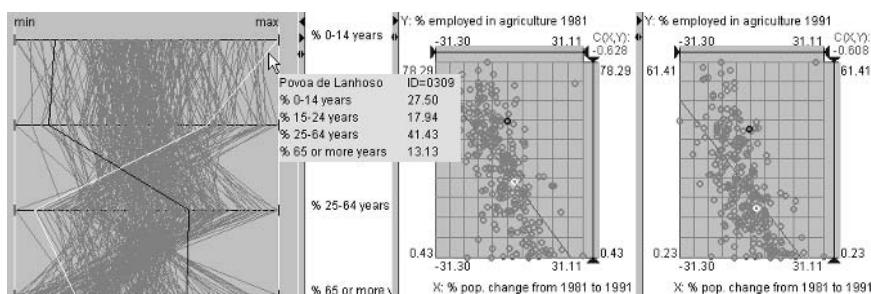
Hence, the explorer needs to ascertain the attribute values that lie behind an odd display item and to determine the corresponding reference. The

explorer also needs to compare these values with other values and, if the reference set has some kind of organisation (e.g. ordering or other relations between the elements), position the reference in relation to the other references. In particular, when the data are spatially and/or temporally referenced, the analyst must determine the relative spatial and/or temporal positions of the unusual attribute values within the entire space segment and/or time period that the data refer to. Furthermore, fitting the strange thing into our picture of the world (or, less generally, into the overall pattern being constructed in the course of the analysis) requires the explorer to look for values of other attributes corresponding to the reference in focus, and to its neighbours. Are these values or their combinations also bizarre? If yes, how are the various unusual features related to each other?

Finally, the explorer needs to find out the reason for the untypical characteristics of the reference in focus. Sometimes, the domain knowledge possessed by the analyst allows him/her to explain the oddness without any additional analysis. However, it often happens that the explorer needs to extend the scope of the analysis by using additional data which characterise certain phenomena potentially related to the phenomenon under study. For example, to understand the reasons for rises or falls in criminality, the explorer may need to look for changes in legislation and/or to consider the dynamics of economic indices, unemployment, migration, etc.

So, what tools does an analyst need in order to attend to particulars in a proper way? First of all, a convenient querying tool, which provides various “details-on-demand” (Ben Shneiderman). Thus, in describing the examples included in this section, we have used a direct-manipulation query tool of the “What’s this?” type: when the mouse cursor is positioned on a display element, a pop-up window appears, in which the data items represented by this display element are listed. This includes the reference and the corresponding values of the attributes portrayed by the display. An example is shown in Fig. 5.31: the mouse cursor points to a line in a parallel-coordinates display, and a pop-up window below the cursor provides the relevant information. Specifically, as the parallel-coordinates display represents the characteristics of the districts of Portugal in terms of the four age structure attributes, the pop-up window shows the name and identifier of the district that the line corresponds to and the values of the four attributes for this district.

Of course, this way of accessing detailed information is not the only possible solution. Moreover, a query tool of this kind has not only advantages (ease of use and quick response) but also drawbacks (transience of the pop-up window containing the information, and covering of the original content of the display). Therefore, the user may need a combination of solutions. Thus, the query tool that we used can also display information in



**Fig. 5.31.** A direct-manipulation query tool shows the reference and the exact attribute values represented by a selected display item here

a special, permanently existing window. Information appears in this window when the user selects one or more display items by clicking on them or dragging a frame around them. The information is removed from the window when the user explicitly deselects the previously selected items. The user may choose the values of which attributes will be displayed in the information window. Again, this is only an example of a possible solution.

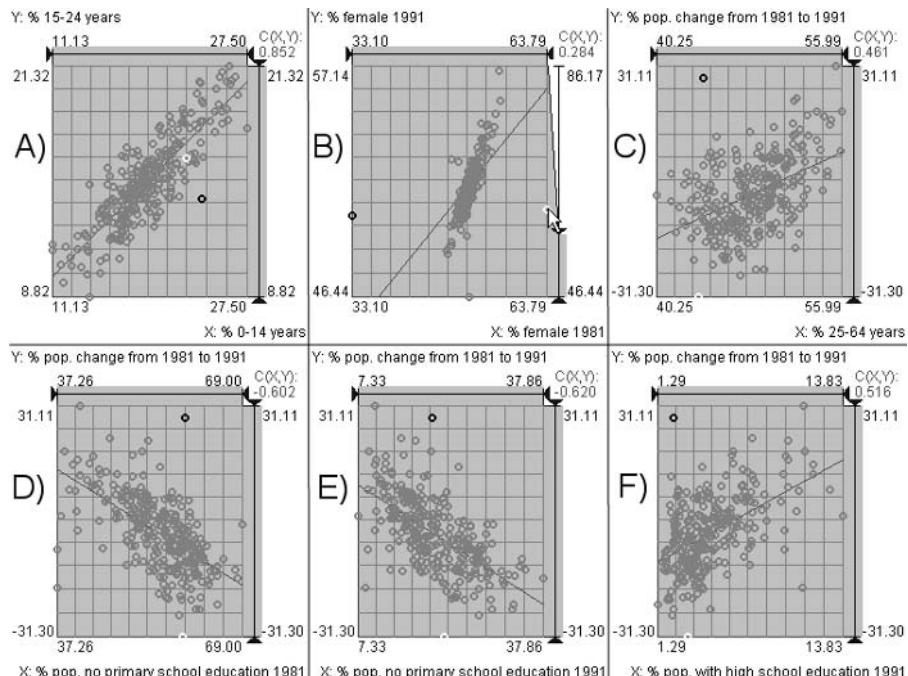
Besides querying tools, the analyst needs the possibility to locate the references in focus on various data displays present on the screen. This is necessary for (1) positioning these references in relation to the entire reference set, in particular, space and/or time; (2) detecting other “strange” values or value combinations among the characteristics of these references or their neighbours; and (3) possibly looking for corresponding characteristics in data concerning other phenomena.

Finding display items corresponding to a particular reference or a few references on all displays is adequately supported by display-linking tools, specifically, the simultaneous reaction of multiple displays to selection of references by the user in the form of highlighting (special marking) of the corresponding display items.<sup>33</sup> Thus, in Fig. 5.31, the line in the parallel-coordinates display pointed at with the mouse cursor is highlighted in white while the other lines are grey. Simultaneously, in the two scatterplots beside the parallel-coordinates display, the dots corresponding to the same district of Portugal as the highlighted line are also highlighted. For consistency, the same highlighting colour (white) is used in all the displays.

It can be noted that the displays in Fig. 5.31 also contain some items coloured in black. These items represent another district, which was selected earlier by clicking on the corresponding line in the parallel-coordinates display (this is the line with the unusually steep segment be-

<sup>33</sup> Not all types of displays can behave in this way. In particular, displays representing aggregated data characteristics, such as histograms, cumulative curves, and box-and-whiskers plots, may be unsuitable for marking selected references.

tween the lower two axes). Let us explain the presence of two highlighting colours, white and black, in the same displays. The display-linking tool that we used to produce the illustration supports two selection modes, transient and durable. Transient selection of a reference occurs when the mouse cursor points to a corresponding visual element in a display. Highlighting by white colouring appears in response to such a selection. The selection is cancelled and the highlighting disappears as soon as the mouse cursor is moved away from the visual element. Durable selection may be done through clicking on display items, enclosing them in a frame, or in some other ways. A reference, once selected, remains in this state until the user explicitly deselects it. Marking in black is applied to display items corresponding to such durably selected references.



**Fig. 5.32.** Six scatterplots, linked through simultaneous highlighting of the dots corresponding to selected references. The black dots in all of the displays correspond to the district of São João da Pesqueira, which has an unusual combination of values of the attributes “% employed in agriculture” and “% population change from 1981 to 1991”, exposed by the scatterplots in Fig. 5.29

To demonstrate the use of this display-linking tool, let us look to see whether the district of São João da Pesqueira, with an unusual combination of values of the attributes “% employed in agriculture” and “% population

change from 1981 to 1991" (see Fig. 5.29) has any other peculiarities. We select this district durably by clicking on the corresponding dot in one of the scatterplots shown in Fig. 5.29. Then, we generate other displays and look in them for items highlighted in black. Displays where the highlighted items stand apart from the others or look different from the others are especially interesting for us.

Figure 5.32 contains six scatterplots containing black dots separated from the rest. We can see that the district of Sao Joao da Pesqueira has some quite peculiar traits in addition to the untypical population increase for a district with so many people employed in agriculture.

Thus, the display A shows us that this district has unusual proportions of children and young people: the proportion of children is rather high, while the proportion of young people is relatively low, in comparison with the other districts, in which the proportions of these two age groups are more balanced. The display C demonstrates that the district under examination has an unusually low percentage of people aged from 25 to 64 years in comparison with the other districts with a high population increase. The displays D, E, and F show that the proportions of people without basic education in 1981 and 1991 and the proportion of people having high school education in 1991 are also uncommon for a district with such a population growth.

The display B is, perhaps, the most interesting. In this scatterplot, the proportions of females in 1991 are plotted against the proportions of females in 1981. For better legibility, we have applied a zooming tool and removed one outlier, specifically, an exceptionally high value of 86.17 in 1991, which may be an error in the data. The black dot corresponding to the district of Sao Joao da Pesqueira lies on the left edge of the scatterplot. Its position in relation to the other dots shows us that this district had an exceptionally low percentage of females in 1981, but that this proportion became quite regular in 1991. This could be regarded as another error in the data, but instead of jumping to a conclusion, let us look carefully at the display. We can notice another quite peculiar dot on the right edge, opposite to the black dot. The mouse cursor points to this dot in the display B, and the dot is highlighted in white. The position of this dot means that the corresponding district, in contrast to Sao Joao da Pesqueira, had a very high proportion of females in 1981, which reduced in 1991 to a regular value. Furthermore, from the positions of the white dots in the displays C, D, E, and F, we can see that the respective district, named Sabrosa, had an extreme population decrease between 1981 and 1991. This suggests the idea that a large number of women moved between 1981 and 1991 from Sabrosa to Sao Joao da Pesqueira so that the resulting gender structure in both districts became more balanced. This idea is supported by the obser-

vation that the two districts are geographical neighbours, as may be seen from the map fragment in Fig. 5.33.



**Fig. 5.33.** The districts that had peculiar gender structures of the population in 1981 but quite regular structures in 1991 are geographical neighbours. In this map fragment, these two districts are highlighted in black and white

Moreover, if the hypothesis about a migration of the female population from Sabrosa to Sao Joao da Pesqueira is true (despite seeming weird), it may explain the unusual ratio in Sao Joao da Pesqueira between the proportions of children aged from 0 to 14 years and of young people aged from 15 to 24 years, which is exposed in the scatterplot A in Fig. 5.32. Taking into account that the values of the age structure attributes, in particular, “% 0–14 years” and “% 15–24 years”, refer to the year 1991, we may guess that the increase in the female population in Sao Joao da Pesqueira and the consequent improvement of the balance between the genders could have resulted in a higher birth rate, which, in turn, would lead to an increase in the relative number of children.

Of course, all these speculations need to be verified either on the basis of domain knowledge or by analysing additional data. Unfortunately, we are not experts in the demography of Portugal and have no other potentially relevant data. Therefore, we propose to regard our investigation as just a demonstration of the use of querying and display-linking tools for performing various tasks on the elementary level of analysis.

Besides detecting surprisingly unusual values and value combinations in various data displays, an explorer may pay special attention to particular references with uncommon characteristics that are expected. Thus, one may expect that the age and employment structure of the population in big cities will differ from that in the surrounding districts. Similarly, for a time-referenced attribute, values corresponding to the dates of public holidays may differ from values corresponding to other days of the year. The

origin of such expectations is the explorer's domain knowledge or common-sense knowledge. As with unexpected deviations from the majority of the data, querying and marking can also help the explorer in investigating expectable "peculiar" cases.

In multidimensional data, there may be not only unusual individual values and value combinations but also unusual aspectual behaviours. For example, the behaviour of the burglary rate in the District of Columbia over the time period from 1960 to 2000 looks rather odd in comparison with the behaviours in the other states. Such "behavioural outliers" are usually clearly visible in appropriate displays. Like atypical individual values, they require close investigation. Querying and display-linking tools are also helpful in this case.

In this subsection, we have concentrated on elementary tasks and showed how display linking supports them. Display linking is also very important for synoptic tasks, including such intricate tasks as connection discovery. In the next two subsections, we shall switch our focus again to synoptic tasks, and display linking will be given proper attention.

#### **5.4.8 Principle 8: Establish Linkages**

As we have said before, the primary goal of exploring a dataset is to characterise the overall behaviour of the data function and, thereby, the behaviour of the underlying phenomenon. "Characterise" means to derive a sufficiently precise and, at the same time, simple (parsimonious) generic pattern reflecting essential features of the behaviour. The essential features are the features that pertain to the entire reference set or a substantial part of it rather than to individual references. They can be called "reference-invariant". Hence, it can be said that a pattern is a reference-invariant depiction of a behaviour.

The goal of deriving a reference-invariant depiction can be best achieved if the entire behaviour is exposed to the explorer by means of an appropriate visualisation, in accord with the principle "see the whole". However, this is possible only in very simple cases, where the number of referential and characteristic components in the data is rather small. Such cases are very rare. A more typical case is where there are not enough display dimensions and retinal variables for the simultaneous representation of all data components. It should also not be forgotten that not every display dimension supports unification. It often happens that, despite every data item being represented on the screen, the overall behaviour cannot be grasped.

Another difficulty is that the overall behaviour is often too complex to be approximated by a sufficiently simple and, at the same time, expressive generic pattern, i.e. the explorer may be unable to derive a meaningful reference-invariant description that would be valid for the entire reference set.

The only possible approach to coping with these problems is to consider manageable parts and slices of the overall behaviour. As is schematically shown in Fig. 5.1, the explorer characterises these parts and slices (called “partial behaviours”), i.e. approximates them by suitable patterns. Then, these partial patterns need to be integrated into a unified pattern approximating the overall behaviour.<sup>34</sup> In this subsection, we shall focus on the possible approaches to the synthesis of a unified overall pattern from partial patterns.

We need to consider separately the approaches to dealing with multiple attributes and with multiple referrers. For multiple attributes, there is a possibility in principle to analyse any attribute independently from the others since the value acquired by an attribute is fully determined by a combination of values of referrers and does not depend on the values of the other attributes. The possibility of independent consideration is excluded for multiple referrers because only all referrers taken together provide complete and unambiguous references to the values of the attributes.

Let us first discuss how to build an overall pattern approximating the joint behaviour of multiple attributes. Two extreme approaches are possible:

1. The values of all attributes associated with each reference are visually or computationally integrated, and the explorer tries to grasp the behaviour of the resulting integrated characteristics over the reference set. The pattern thus derived needs then to be interpreted in terms of the original attributes.
2. The behaviour of each attribute is explored independently of the others and approximated by an individual pattern. In order to bring the patterns together, the explorer tries to establish links between the attributes and their behaviours. The explorer not only notes similarities and differences but also looks for correlations and influences between the attributes.

There is also an intermediate approach between these extremes. The whole set of attributes may be divided into several attribute groups. The first approach is then applied to each group. In order to integrate the resulting patterns, linkages between the attribute groups are established, according to the second approach. The criteria for defining the attribute groups

<sup>34</sup> Such an overall pattern is often called a “model”, particularly, in the data-mining literature.

may come from domain knowledge or result from previous analysis. For example, the census attributes characterising the districts of Portugal may be divided, on the basis of domain knowledge, into age structure attributes, occupation attributes, education-level attributes, etc. Alternatively, the attributes may be grouped according to the similarity of their spatial behaviours, i.e. on the basis of some previous analysis.

It is clear that the second approach to the characterisation of the behaviour of multiple attributes involves a decomposition of the overall behaviour, as does the intermediate approach. However, it is not only this sort of decomposition that is possible. The first approach often results in a compound pattern, derived by dividing the reference set into subsets and characterising the corresponding partial behaviours. This is done when the overall behaviour is too complex to be approximated by a single atomic pattern.

The decompositions occurring in the first and second approaches are different; let us call them Decomposition A and Decomposition B, respectively (Decomposition B includes the consideration of attribute groups as well as individual attributes). Decomposition A produces a set of patterns such that each pattern includes all the attributes under exploration but refers to a subset of the overall reference set and reflects the reference-invariant features of the behaviour based on this subset. This can be viewed as partial, or local, invariance. The overall, globally invariant pattern is built from the partial patterns by specifying the domain of applicability of each pattern, i.e. the reference subset for which it is valid.

The partial patterns produced by means of Decomposition B, in contrast, apply to the entire reference set, i.e. they are globally invariant. The patterns are partial because each of them reflects only a part of the available characteristics, specifically, only characteristics in terms of one attribute or subset of attributes. The overall, exhaustive pattern is built from the partial patterns by establishing linkages between the characteristics that the patterns reflect.

Both Decomposition A and Decomposition B may result in an overall pattern that has a hierarchical structure: the overall pattern is integrated from partial patterns which, in turn, are also composed of smaller patterns, and so on. Moreover, Decomposition A and Decomposition B may be applied to each other. Thus, one can divide the whole set of attributes of a dataset into individual attributes or attribute groups and then characterise the behaviour of each attribute or group by dividing the reference set into appropriate subsets. Or one can first divide the reference set into subsets and then characterise the behaviour over each subset by considering individual attributes and attribute groups.

When we described the principle “divide and group”, we were actually referring to Decomposition A, which is based on dividing the reference set into subsets. Here, we shall not discuss any further how the division is done; instead, we shall briefly touch upon the problem of joining the partial patterns thus derived into an integrated overall pattern.

As we have mentioned before, with Decomposition A, the overall pattern is built from the partial patterns by specifying the applicability domain for each partial pattern. It is usually inappropriate to specify the applicability domain by just enumerating all the references included in this domain (this may work only in the case of very few references). Hence, the subsets that the reference set is divided into should be easily describable without enumerating the individual elements. This is, in fact, an important criterion to be taken into account during the division process: the explorer needs to divide the reference set so that, on the one hand, the behaviour over each subset can be characterised effectively in a reference-invariant manner, and, on the other hand, the subsets themselves can be described meaningfully, parsimoniously, and consistently.

In Sect. 3.4.2, we tried to formulate some rules for dividing the reference set into subsets in building compound patterns. The general idea is to take account of the properties and relations pertaining to the reference set. Thus, a spatial referrer should usually be divided into spatially contiguous subsets rather than collections of scattered locations, and a temporal referrer should usually be divided into temporally contiguous intervals rather than groups of chaotically dispersed moments. The reference set may also be divided on the basis of qualitative differences between the references. The choice of the qualities that the division should be based upon is driven by domain-knowledge-based expectations of substantial differences in the behaviour. For example, in analysing medical data, it is appropriate to divide a set of persons into men and women and/or into age groups. In analysing time-referenced data in which cyclic changes may be expected, the time period can be partitioned according to phases of a cycle rather than into contiguous intervals. For example, one can divide a time period with a length of several years not into years but into January, February, and so on, or into spring, summer, autumn, and winter. A geographical space can be partitioned according to the characteristics of the relief and/or land cover. Partitions of reference sets on the basis of qualitative differences are usually easy to describe, but they are only justified when the corresponding partial behaviours differ substantially from each other.

It may happen that a particular division of a reference set does not yield good results. Thus, the explorer may try to divide the reference set according to the variation of the behaviour (i.e. so as to minimise the diversity within subsets and at the same time maximise the differences between sub-

sets), but the resulting subsets can be hard to identify and interpret. Or the explorer may divide the reference set according to some qualities of the references but detect no significant differences between the behaviours over the resulting subsets. Hence, the explorer needs tools that allow him/her to try various divisions and compare them until it becomes possible to build an appropriate integrated pattern with a clear and parsimonious definition of each subpattern and its application domain. This is not always achievable, and the explorer often needs to find a suitable trade-off between the two criteria.

Let us refer to some examples of Decomposition A that have occurred earlier in this book. In Sect. 4.4.3, we discussed various tools for classification of references according to the values of one or more attributes. Classification divides a reference set into subsets so that the corresponding characteristics lie within certain ranges. These ranges are, in fact, reference-invariant characterisations of the parts of the overall behaviour based on these subsets, or, in other words, partial patterns (subpatterns). A good classification is achieved when the ranges are compact, with little internal variance, while the subsets are easily describable. For example, when we classified the districts of Portugal, we attempted to define the classes so that the territory was divided into coherent parts that could be given meaningful names or descriptions such as “north-western coast”, “areas around big cities”, or “deep inland”. For this purpose, we need highly interactive and dynamic classification tools, which allow us to modify the definitions of the classes quickly and easily, and to immediately observe the resulting division of the reference set.

In Sects. 4.7.4 and 4.7.6, we considered the use of clustering tools, which are also intended for dividing a reference set into subsets according to values of multiple attributes so as to minimise the variance of the characteristics within the subsets and maximise the differences between the subsets. As in the case of classification, we wish the subsets to be easily interpretable, for example, to form coherent regions in space. From this perspective, some of the clustering results presented in this book are not bad, for example the results shown in Fig. 4.120C. However, clustering entails another problem: clustering tools do not provide reference-invariant descriptions of the characteristics pertaining to the subsets. Such descriptions need to be constructed by the explorer. For this purpose, display linking is helpful. Thus, in our examples, we transmitted a division of the reference set obtained from a clustering tool to histograms (Figs 4.120C–4.123C) and parallel-coordinates displays (Figs 4.132C and 4.133C) that visualised the attributes used for the division. This helped us to define the general profiles of the subsets, i.e. the partial patterns from which the overall pattern could be constructed.

So, Decomposition A requires mainly a good division of the reference set, and a description of the logic of this division serves as a means of linking the partial patterns. There is, in principle, no need for any additional links because the reference subsets are parts of the whole reference set, and it is quite clear how these parts are related to each other. Decomposition B is quite different: it is based on considering individual attributes and attribute groups, which are not viewed as elements and subsets of some unified whole. Therefore, the integration of the partial patterns requires establishing explicit linkages between the attributes or attribute groups.

Linkages between attributes are reference-invariant associations of characteristics, where certain values of two or more attributes tend to occur together throughout the reference set. “Together” means that the values are associated with the same references or with groups of references related to each other in an identifiable way, in particular, neighbouring references. A co-occurrence of values with common references is usually called a correlation. For example, we have detected in the Portuguese data that a high proportion of elderly people in the population of a district is correlated with a high percentage of people who have no primary school education, and that districts with a high proportion of people working in services tend to have a low percentage of uneducated people but quite many people who have high school education. To our regret, in the datasets that we have, we have not found any associations between values of different attributes characterising different although related references. As an example, we could refer to the known historical case where an association between a higher than usual content of fluoride in a lake providing drinking water to the surrounding area and a lower than usual frequency of dental caries in that area was detected.

The latter example demonstrates that the role of establishing linkages between attributes or attribute groups is not limited merely to the integration of partial patterns. This is a way not only to a description of the overall behaviour but also to an explanation of it. So, we have in fact moved from descriptive synoptic tasks to explanatory, or connectional synoptic tasks (see Fig. 3.23), the goal of which is to derive a special kind of pattern (we have called it a “connection pattern” or “linkage pattern”) that characterises the behaviour of different attributes or phenomena with respect to each other (a “mutual behaviour”). These tasks are usually very complex and require much more imagination and creative thinking than do descriptive tasks. There are no recipes for how to discover essential connections between attributes or between phenomena. We can refer to potentially helpful tools and approaches, but nobody can guarantee that they will always be effective and lead to satisfactory results.

First of all, we would like to mention the realm of computational methods, in particular, statistics, with its highly developed apparatus of correlation analysis. It should be noted, however, that the techniques of correlation analysis deal mostly with numeric attributes. Moreover, they treat any reference set as a statistical population, i.e. they do not take account of distances, ordering, and other relations between references. Many algorithms for data mining are also intended for the discovery of correlations between attributes or typical associations between attribute values. Some of these algorithms are based on statistical techniques, while others involve information theory or other mathematical apparatus. Recently, specific data-mining methods that search for correlations and associations in spatially referenced data have appeared (Openshaw and Openshaw 1997, Miller and Han 2001).

Of the visual techniques, the scatterplot is recognised as the best tool for detecting correlations between numeric attributes. A scatterplot can expose a relatedness of two attributes irrespective of whether it is linear or non-linear, while computed correlation coefficients are only suited to linear dependencies. The use of the scatterplot technique may be problematic when the data volume (i.e. the number of references) is very large. In such a case, the classical scatterplot may be replaced by a modification of this technique involving data aggregation, such as the binned scatterplot (Fig. 4.64) or the bagplot (Fig. 4.74).

Since the scatterplot can be used only for a pair of attributes, other techniques are necessary for dealing with three or more attributes. Scatterplot matrices are displays consisting of multiple scatterplots, each representing one of the possible pairs of attributes. In such a composite display, one can detect pairwise correlations between attributes. Another applicable technique is the parallel-coordinates display. For two attributes represented on adjacent axes of a parallel-coordinates display, the indicator of a positive correlation is that the lines between these axes are close to parallel, whereas all the lines crossing indicates a negative correlation. However, nothing can be said, typically, concerning attributes represented on non-adjacent axes. Therefore, the parallel-coordinates tool must allow the user to change the order of the axes. Some implementations of this technique involve computation-based optimisation of the arrangement of the axes so that neighbouring axes correspond to the most related attributes.

Another useful visual tool is the variant of the table display known as the “table lens” (see Fig. 4.10), in which the values of numeric attributes are represented by bars drawn inside the table cells so that the sizes of the bars show the relative positions of the values between the minimum and maximum of the respective attribute. In such a display, the rows may be sorted according to the values of one of the attributes. Then, the bars in

columns corresponding to attributes correlated with this attribute will unite visually into shapes close to triangular. Thus, the screenshot in Fig. 4.10 suggests that the attribute “% employed in agriculture 1991” (the one according to which the rows are ordered) is positively correlated with the attribute “% pop. no primary school education 1991” and negatively correlated with the attribute “% pop. with high school education 1991”.

Any display of numeric attributes may be insufficiently expressive with regard to exposing correlations when some of the attributes (or all of them) have outliers or skewed distributions. In such cases, display manipulation tools for removing outliers (focusing) and transformation of the visual encoding function (e.g. from linear to logarithmic) are appropriate.

Besides pairwise correlations, there may be more complex relationships which involve more than two attributes, for example when combinations of values of two or more attributes influence values of other attributes. We do not know of any visualisation technique that would effectively support detecting such dependencies. Dynamic querying tools combined with appropriate visual displays have been suggested by some researchers as instruments for the exploration of multiattribute links, for example, Attribute Explorer and Influence Explorer (Spence and Tweedy 1998, Spence 2001). The idea is demonstrated in Fig. 4.102: the explorer sets limits on the values of some attributes, these limits are used for filtering the data, and the explorer can see which values of the other attributes have been completely removed by the filter and which still occur in the active data subset. The visualisation can also show how many value occurrences have been removed and how many remain. In principle, such a tool can work not only with numeric attributes but also with qualitative ones.

In a tool such as Attribute Explorer, linked displays play a significant role. Linked displays can also be used in combination with direct-manipulation query tools to allow the user to select visual items in any of the displays and immediately see what items correspond to the selected items in the other displays. For example, in Fig. 4.101, the explorer has selected a group of dots on a scatterplot corresponding to low values of the attributes “% pop. no primary school education 1991” and “% employed in services 1991”. From the two histograms linked to the scatterplot, the explorer can see that the selection corresponds to mostly low values of the attribute “% employed in agriculture 1991” and medium to high values of “% employed in industry 1991”.

A disadvantage of using any sort of query tools for the discovery of links between attributes is that the explorer, in principle, needs to try all possible variants of setting limits or making selections, and this is an unfeasible task. Therefore, such tools can be used when the explorer has made some guesses concerning possible dependencies, in particular, which

of the attributes can influence the others. Thus, in the Influence Explorer tool, the attributes are divided into input and output attributes, with the assumption that the values of the output attributes depend on the values of the input attributes. The goal is to examine how the choice of various value ranges of the input attributes influences the values of the output attributes.

In searching for links between groups of attributes, a combination of Decomposition B with Decomposition A can be utilised effectively. The idea is that the explorer divides the attributes of a dataset into groups (Decomposition B). Then, the reference set is partitioned into subsets according to the values of the attributes of one of the groups. For this purpose, the explorer applies appropriate classification or clustering tools. After that, the explorer considers the statistics of the values of the other attributes for the reference subsets thus obtained and/or transmits the reference set division to various displays of these attributes by applying display coordination tools. An example is demonstrated in Figs 5.34C and 5.35C.

To produce these illustrations, we have applied a clustering tool in order to divide the districts of Portugal into subsets (classes) according to the values of three attributes characterising the structure of the employment of the population in different sectors of the economy: agriculture, industry, and services. To interpret the meaning of the computationally derived classes, we have visualised their characteristics in terms of the employment attributes in a parallel-coordinates display. On the left in Fig. 5.34C, the characteristics are shown in an aggregated form, as “envelopes” containing the lines for each class and divided into stripes according to the deciles of the attribute values in the classes. It may be seen that the “blue” class is formed by the districts with high employment in agriculture and low employment in the other two sectors, the “green” class consists of the districts with high employment in industry, the “yellow” class is characterised by high employment in services, and the “red” class consists of the districts where the employment in all three sectors is from small to medium, and none of the sectors prevails significantly.

Now, we would like to find out how the employment structure is related to the educational level of the population in the districts. On the right in Fig. 5.34C, we can see the statistics of the values of four attributes reflecting the education level of the population, specifically, the proportions of people without primary school education, with primary school education, with preparatory school education, and with high school education, for the entire country and for the four classes of districts. The statistics show us that high employment in agriculture (the “blue” class) correlates with a high proportion of people without education, a low proportion of people with preparatory school education, and a still lower proportion of people with high school education. The districts with high employment in industry

(the “green” class) tend to have low proportions of people without education and of people with high school education, while the proportions of people with preparatory and, especially, primary school education tend to be high. The districts with high employment in services are characterised by mostly low proportions of people without education and (surprisingly) of people with preparatory school education, a low to medium proportion of people with primary school education, and, notably, a high proportion of people with high school education.

In a less aggregated form, the “educational profiles” of the classes can be seen in the four parallel-coordinates displays in Fig. 5.35C. Each display contains only lines for a single class. The background colouring indicates the positions of the first and ninth deciles (i.e. tenth and 90th percentiles) of the values of the education attributes in the respective classes. We shall not describe the displays in detail since their meaning is quite clear. It is important that we have more or less succeeded in establishing linkages between two groups of attributes by utilising Decomposition A, i.e. partitioning of the reference set according to the values of several attributes, and tool combination, specifically, application of a statistical tool to the division obtained and reflection of that division in visual displays by means of multicolour marking and display multiplication.

The tools discussed so far are primarily intended for attributes referring to statistical populations, i.e. reference sets without ordering, distances, or other relations between the elements. More precisely, such relations are not taken into account; hence, one can, in principle, apply scatterplots or Attribute Explorer to spatially and/or temporally referenced attributes, but one should keep in mind that a great deal of potentially relevant information is thereby simply ignored. Therefore, the explorer should not rely only on such tools when dealing with spatial or temporal data.

The appropriate visual tools for data with spatial and/or temporal components are tools that reflect the essential properties and relations pertaining to these components. In particular, an appropriate representation of geographical space is a map display, while time can be represented by one of the planar display dimensions, as, for example, in a time graph.

One possible approach to detecting links between spatially or temporally referenced attributes is to compare visual displays of them, for example several maps, as in Figs 5.16 and 5.21C, or several time graphs, as in Fig. 5.2. A similarity between the displays indicates relatedness of the attributes. However, this approach may be ineffective when the attributes are heterogeneous, for example when one attribute is numeric while the other is qualitative.

An overlaid representation of several attributes or even of several heterogeneous phenomena within a single display is a more universal and

quite effective approach, providing that visibility of all the information layers is ensured (some solutions have been discussed in Sect. 5.4.4). Thus, links between several temporally referenced attributes or several phenomena that have temporal components may be investigated using a combined representation of these attributes or phenomena in a time graph or some other display in which one of its dimensions represents time. A map display can be used for a combined visualisation of several spatially referenced attributes or phenomena with spatial components. For example, to look for links between the movement of storks and the relief, land cover, and climate, one can use a map display with an overlaid representation of all these phenomena.

It is interesting to note that in this case the phenomena not only differ in their nature but also have different reference sets: relief and land cover refer to geographical space, climate refers to space and time, and the movement of the storks refers to time, while space is the value domain of the attribute reflecting the location of the storks. The presence of spatial components in all of the phenomena makes it possible to represent them in a common map display. Since some of the phenomena refer also to time, it is necessary to incorporate time into the visualisation. For this purpose, one can use map animation or multiple juxtaposed maps representing different time moments.

To our regret, we have no climate data related to the movement of the storks, but a representation of the movement trajectories on top of a satellite image showing relief, water, and land cover allowed us to note that the storks avoid flying over the sea but prefer to go around it. On the way back from Africa to Europe, they avoid flying over deserts, and on the way south, they move over a desert but with an increased speed as compared with the other segments of their trajectories. These observations provide a simple example of a link between several phenomena that may be detected using an overlaid representation of them. In Fig. 5.36C, two screenshots represent the movement of four storks during the period from 20 August 1998 to 31 January 1999 (with mostly southward movement) and from 1 February 1999 to 1 May 1999 (when the storks returned to Europe).

In looking for spatial or temporal correspondences between phenomena or attributes, it should be borne in mind that influences in space and time may be “lagged”, that is, the effects of events or characteristics may not be observed exactly in the same place and at the same time as where and when the event occurred or the characteristics were attained, but at a certain distance in space and/or time. When all data can be viewed simultaneously, for example, on a map or a time graph, it is usually possible to detect such lagged influences. Animated displays or “small multiples” are less supportive for such kinds of observations. It may be useful to try a

“shifted” representation of the temporal development of several attributes or phenomena. For example, each individual frame in an animated representation or a “small multiples” display may represent the values of one attribute referring to the time moment  $t$  and the values of another attribute referring to the moment  $t + \Delta$ , where  $\Delta$  is a user-specified value for the temporal shift. For two time-referenced numeric attributes, it may also be useful to look at a scatterplot in which the values of one of the attributes referring to the moment  $t$  are plotted against the values of the other attribute for the moment  $t + \Delta$ . Of course, if the user has no expectation concerning the possible “latency period” after which effects may appear (such expectations may come from the user’s domain knowledge), various values of  $\Delta$  need to be tested.

As we have mentioned in Sect. 3.5.2, essential links may exist not only between different attributes or phenomena but between parts of a single phenomenon; for example, hot, dry summers may be correlated with subsequent cold winters. This means that, when applying Decomposition A, i.e. decomposition on the basis of dividing the reference set, an explorer may go beyond the generation of a compound descriptive pattern. He/she may try to derive a connectional pattern by establishing linkages between the subpatterns of the compound pattern.

On the basis of our experience, we believe that such tasks can be appropriately supported by tools that allow the explorer to visualise the characteristics pertaining to different reference subsets in separate displays and flexibly arrange these displays for the most convenient comparison. For example, if the explorer is analysing monthly data over many years, it could be useful to divide the data into yearly portions and represent these portions in a collection of displays arranged in a stack, one below another. This would facilitate comparisons between the corresponding months of different years. It might also be beneficial to construct an overlaid representation of all of the data portions within a single view.

The tools we have at our disposal do not provide such possibilities. Since we would still like to give an example of looking for links between subpatterns, we shall describe an attempt to use the available visualisation facilities to test whether any relations between the weather in summer and in the subsequent or previous winter can be detected in the monthly climate data we have for the period from January 1991 to May 2003. Although the data were collected at 43 weather stations in Germany and hence have a spatial referrer, we are not interested in the spatial component; our goal is to find space-invariant correspondences. Therefore, it is quite appropriate to reduce the dimensionality of the data by means of aggregation over the spatial referrer.

For the investigation, we have used the visualisation shown in Fig. 5.37C, which consists of four vertically aligned displays representing the space-aggregated monthly values of various climate attributes:

- the monthly mean of the daily mean temperature (degrees Celsius);
- the monthly mean of the daily minimum temperature (degrees Celsius);
- the total monthly sunshine duration (hours);
- the total monthly precipitation (millimetres).

The order of the attributes in the above list corresponds to the order of the displays, from top to bottom.

The horizontal dimension of each display represents the temporal referent. Each display consists of segmented bars, one bar per month. To produce the display, the monthly data have been aggregated over Germany by dividing the value ranges of the attributes into intervals and counting the attribute values fitting within these intervals. The sizes of the segments of the bars are proportional to the counts. We have specified the following interval breaks:

- $-10, -5, -1, +1, +5, +10, +15$ , and  $+20$  for the monthly mean of the daily mean temperature (the topmost display);
- $-10, -5, -1, +1, +5, +10$ , and  $+15$  for the monthly mean of the daily minimum temperature (the second display from top);
- $50, 100, 150, 200, 250$ , and  $300$  for the total monthly sunshine duration (the third display from top);
- $20, 40, 60, 80, 100$ , and  $200$  for the total monthly precipitation (the display at the bottom).

In the upper two displays, shades of blue correspond to temperature values below zero (more precisely, up to  $-1$ ), yellow to values around zero (from  $-1$  to  $+1$ ), and red to values above  $+1$ . Hence, cold periods are indicated by high amounts of blue, and hot periods by high amounts of dark red. In the display of sunshine duration (the third from top), yellow corresponds to a duration between 150 and 200 hours, shades of blue to shorter durations, and shades of red to longer durations. Hence, an abundance of blue indicates periods with low sunshine, in particular, winter, when the days are short. In the display of the precipitation, shades of brown correspond to low precipitation, and shades of green to high precipitation. Hence, a high proportion of brown in a bar indicates a dry month, and a high proportion of green corresponds to wet weather.

For the purposes of our investigation, we have divided the entire time period into “summer” and “winter” seasons, assuming the summer season to include the months from May to October and the winter season to in-

clude the months from November to April of the following year. For orientation, we have put at the bottom of Fig. 5.37C a horizontal bar divided into red and blue segments indicating the summer and winter seasons.

From the visualisation in Fig. 5.37C, we can see that the hottest summers were in the years 1994, 1995, and 1997: the corresponding bars in the upper two displays contain the highest amounts of dark red colour. From the relative amounts of green and blue in the corresponding bars of the lowest display, we can see that these summers were not very dry, except for the second half of the summer of 1997. While the winter after the summer of 1995 was very cold, the winter after the summer of 1994 was not so cold, and the winter after the summer of 1997 was exceptionally mild. The winter before the summer of 1994 was quite mild and, apparently, cloudy: the sunshine duration in this winter is one of the lowest. The winter before the summer of 1995 was also rather mild; however, the winter before the summer of 1997 was one of the coldest. Hence, there is no stable association between hot weather in summer and the character of the weather in the previous or subsequent winter.

Let us look at whether there is any association between cold winters and the weather in the previous or following summer. The coldest winters were in the years 1995–1996, 1996–1997, and 2002–2003: the corresponding bars in the upper two displays contain the highest amounts of blue colour. The winter of 1995–1996 was also exceptionally dry, as may be seen from the display at the bottom; the other two winters were also quite dry. The summer preceding the winter of 1995–1996 was quite warm and not dry. The summer before the cold winter of 1996–1997 was not very warm and not dry either. Before the winter of 2002–2003, the summer was quite warm and extremely wet; this was a summer of tremendous floods in many European countries, including Germany. The summer following the cold winter of 1995–1996 was not very warm, while the summer after the winter of 1996–1997 was one of the hottest. The dataset does not contain data for the summer of 2003 following the cold winter of 2002–2003 but everybody in Europe still remembers this extremely hot, dry summer, with extensive forest fires and many deaths from heatstroke.

So, the available data do not allow us to detect any stable association either between hot summers and the weather in the preceding or following winter or between cold winters and the weather in the preceding or following summer. It would be better, of course, to have data for a longer time period, since the number of occurrences of hot summers and cold winters in the given period is too small for drawing any conclusions.

In our exploration, it was important that the visualisation tool that we used allowed us to find, grasp, and compare parts of the overall behaviour

corresponding to various reference subsets, in this particular case, the winters and summers of different years. In general, this is what is basically required for any exploration of possible links between partial behaviours over reference subsets. Perhaps it would have been a little more convenient for us if the available tools had allowed us to split the visualisation into views of the behaviours over different subsets and to arrange such views for easy comparison. However, it is not always possible to obtain an ideal tool for a task, and it is often necessary to find a way of making appropriate use of what is available. Usually, a visual display complying with the principle “see the whole”, i.e. one that shows the entire overall behaviour (or the aspectual behaviour that is the focus of the investigation, as in the present case), is also suitable for making comparisons between parts of the behaviour corresponding to various reference subsets.

Let us now switch to the problem of building an overall pattern in the case of multidimensional data, i.e. data with multiple referrers. When it is impossible to have a complete and unified view of the entire reference set, the explorer needs to consider various slices of the overall behaviour, i.e. fix the value(s) of some referrer(s) and explore the behaviour with respect to the other referrer(s). For example, when a phenomenon varies in space and time, the explorer, on the one hand, fixes various time moments and looks at the spatial behaviours at these time moments, and on the other hand, fixes various places and looks at the temporal behaviours in these places. For a sound, comprehensive investigation, the explorer must consider all possible slices. However, it will not be a useful result of the analysis if the explorer just describes every individual slice. The explorer needs to use these slices to derive an integrated pattern.

It is usually quite clear how to unite the slices resulting from choosing different fixed values of the same referrer(s). Thus, for a spatially and temporally referenced phenomenon, the series of spatial behaviours referring to different time moments can be jointly characterised as the evolution of the spatial behaviour over time. Similarly, the collection of local temporal behaviours in different places can be characterised as the variation of the temporal behaviour over space. However, this will result in two unrelated patterns characterising different aspects of the overall behaviour.

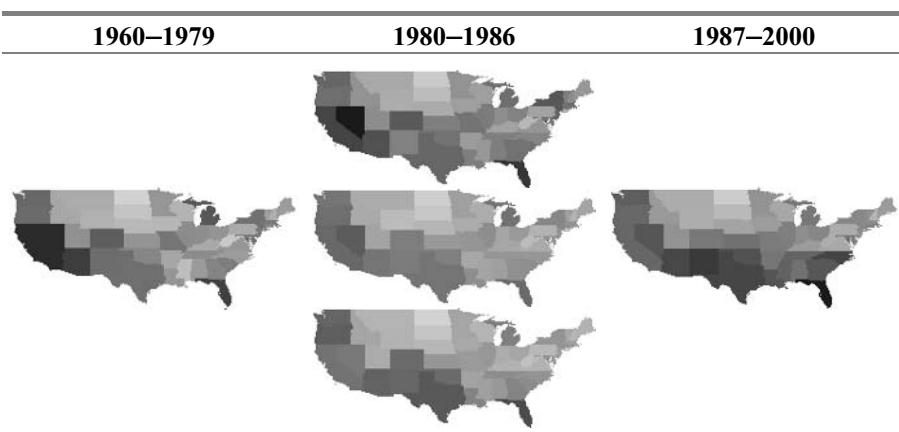
In Sect. 5.3, we introduced an approach to relating such aspectual patterns using the example of the burglary rate data for the states of the USA, which refer to space and time. Let us now try to present this approach in a general form. The main idea is to use the partitioning or grouping performed in the course of characterising each aspectual behaviour for cross-sectioning of the reference set. Then, the parts of the overall behaviour corresponding to the reference subsets obtained in this way are characterised in such a way that all the aspects are accounted for.

Thus, the study of the evolution of the spatial behaviour over time can result in dividing the time period into several intervals such that either the spatial behaviour does not change significantly within an interval or the character of the changes is consistent throughout each interval and differs from that in the other intervals. The study of the spatial variation of the local temporal behaviours may result in a division of the space into parts according to similarity of the local behaviours. On this basis, the explorer can divide the entire two-dimensional reference set into subsets corresponding to the possible combinations of a time interval and a part of the space. If the time period has been divided into  $N$  intervals and the space into  $M$  parts, there will be in total  $N \times M$  subsets of the reference set. Then, for each subset, the explorer needs to characterise what was going on in the respective part of the space during the respective time interval.

Let us take once again the example of the burglary rates in the USA and try to consider it in a slightly different way than in Sect. 5.3 by consistently applying the idea of cross-sectioning.

On the basis of observing the evolution of the spatial behaviour of the burglary rate (for this purpose, we used an animated choropleth map display and a “small multiples” map display), we have divided the period from 1960 to 2000 into three intervals: 1960–1979, 1980–1986, and 1987–2000. During the first interval, the general character of the spatial behaviour did not change significantly; its “averaged portrait” is shown in the first column of Table 5.3. During the third interval, the character of the spatial behaviour was also quite stable, but different from the behaviour over the first interval. The corresponding “averaged portrait” is shown in the third column of Table 5.3. The change in the character of the behaviour is clearly visible: over the interval 1960–1979, the highest burglary rates

**Table 5.3.** The evolution of the spatial behaviour of the burglary rate over time



were in the western states while the interval 1987–2000 is characterised by a cluster of high values in the south of the country. The second interval was a time when the spatial behaviour varied significantly; therefore, it would not be appropriate to produce a corresponding “averaged portrait”. Instead, in the second column of Table 5.3, we have put three screenshots corresponding to the years 1980, 1983, and 1986.

When high precision of behaviour characterisation is required, it may be appropriate to divide the temporal referrer into smaller intervals so that finer changes in the spatial behaviour can be reflected. However, for the purposes of illustration, the three intervals defined above are sufficient.

To divide the territory of the USA into parts according to the similarity of the local temporal behaviours of the burglary rate, we applied a clustering tool, which produced, in accordance with our request, four clusters. These clusters and the corresponding local behaviours are shown in Fig. 5.38C. The local behaviours in each cluster are represented in a separate time graph by lines of the particular colour assigned to the cluster. The thick black line in each time graph shows the “running average”, i.e. the segments of the line connect the average values for the entire country in consecutive years. It may be seen that the green cluster is formed by states with mostly low burglary rates and the red cluster consists of states with high and very high burglary rates. The blue cluster is characterised by medium burglary rates, which decrease significantly in the second part of the time period and become mostly lower than the average for the country. The behaviours in the magenta cluster, in contrast, start with values below the average but end with values above the average for the country.

Spatially, the clusters are not equally well formed. While the magenta cluster is spatially continuous, the blue cluster is rather scattered. The red and green clusters have main bodies formed by adjoining states, and a few additional non-adjacent pieces. We could consider each part of a spatially disjoint cluster as a separate partition. However, to avoid extending the length of this example description, we prefer to stick to four clusters.

So, we have divided the temporal referrer of the dataset into three subsets (time intervals) and the spatial referrer into four subsets (clusters of states). These divisions result from our study of the aspectual behaviours: the first division from an investigation of the development of the spatial distribution of the burglary rate values over time, and the second division from an investigation of the variation of the local temporal behaviour of the burglary rate over space, i.e. the territory of the USA.

Now, according to our general approach, we cross-section the entire reference set, consisting of space and time, into  $3 \times 4 = 12$  subsets, and characterise the behaviour in each spatial partition during each time interval. For convenience, we have produced 12 pictures of the sub-behaviours

corresponding to these reference subsets and organised them in Table 5.4, which is placed together with the colour figures at the end of the book. Let us now briefly describe what we can see from Tables 5.3 and 5.4. It is convenient to organise our observations as is done in Table 5.5.

**Table 5.5.** Description of the behaviour of the burglary rate by parts of the territory (groups of states) and by time intervals

Area	1960–1979	1980–1986	1987–2000
Green (central north and north-east)	Low values; gradual increase; low fluctuations	Low values; decrease followed by flatness; low fluctuations	Low values; very gentle decrease; low fluctuations
Blue (around Great Lakes, plus Utah and Alaska)	Medium values; increase rate higher than in the green cluster; fluctuations	Medium values; quite sharp decrease followed by flatness; low to medium fluctuations	Medium values; gradual decrease; low fluctuations
Magenta (south-east except Florida)	Values mostly between those in the green and blue clusters; increase; fluctuations	Medium values; decrease followed by increase; low fluctuations	Values higher than in the blue cluster; gradual decrease; low to medium fluctuations
Red (west and south-west plus Florida and Michigan)	High values; high fluctuations; general increasing trend	High values; quite sharp decrease followed by slight increase; high fluctuations	High values; decrease rate higher than in the other clusters; medium fluctuations
Entire country	Low values in the “green” area, medium values in the “magenta” and “blue” areas, and high and very high values in the “red” area	Low values in the “green” area, medium values in the “blue” area, and unstable appearances of the “red” and “magenta” areas due to incoherent internal changes	Low values in the “green” area and medium values in the “blue” area. The “red” and “magenta” areas, with high values, have merged together visually.

This table can be viewed as a verbal compound pattern characterising the behaviour of the burglary rate in space and time. It links together the patterns characterising the aspectual behaviours. This sort of linking can be described as structural linking, as it is based on introducing a certain structure over the reference set and, accordingly, in the characterisation of the behaviour. It should be noted, however, that cross-partitioning can also

prompt the detection of cause–effect links. Thus, we can relate the change in the general character of the spatial behaviour (i.e. the difference between the behaviour in 1987–2000 and that in 1960–1979) to the character of the local behaviours in the magenta cluster. It is the increase in the values in this cluster in the second half of the interval 1980–1986 (in contrast to the flatness in the green and blue clusters) that lead to the high values in this cluster attained by the beginning of the interval 1987–2000, and to the perceptual merging of this cluster with the red cluster. The character of the behaviour in the red cluster over the intervals 1980–1986 and 1987–2000, specifically, a sharper decrease in the values than in the other areas, has also contributed to the process of cluster merging: the range of values in the red cluster became very close to that in the magenta cluster, whereas the respective ranges during the interval 1960–1979 were quite different.

In order to finish properly the topic of linking aspectual patterns, we need to answer several questions:

1. Is it always necessary to partition the value set of each referrer?
  2. How do we analyse the behaviour of multiple attributes?
  3. How does the procedure for analysis change when there are more than two referrers?
- *Question 1.* Partitioning of the value set of a referrer is only necessary when the respective aspectual behaviour varies substantially over this set. Thus, in the example above, we had to divide the time period into intervals because the character of the spatial distribution of the burglary rate was not the same during the whole period. Analogously, we had to divide the territory of the USA into groups of states (clusters) because of the substantial differences in the local temporal behaviours. If this were not so, i.e. the character of the spatial distribution were constant and the values in all of the states behaved coherently over time, the description of the overall behaviour would be much simpler, for example “an increasing spatial trend from north-east to south-west; a general increase in the values during the first half of the time period and a decrease during the second half”.
  - *Question 2.* To analyse the behaviour of multiple attributes, for example several different crime rates, we can apply either Decomposition A or Decomposition B described near the beginning of this subsection. Thus, we could use a cluster analysis tool to divide the territory according to the local behaviours of all crime attributes (we indeed did this, and the clusters were similar but not identical to those produced using the burglary rate). We could also represent the joint spatial behaviour of the crime rate over the country in each year by maps with appropriate dia-

grams, and then try to grasp the general character of the behaviour and its change over time. If we manage to do this, we can then proceed as we did with the burglary rate. This is Decomposition A. It is also possible to characterise the behaviour of each crime rate attribute separately, as we did with the burglary rate and then try to establish linkages between the attributes. This is Decomposition B. If the attributes do not behave in the same way (and this is in fact the case), it is appropriate to consider the behaviours on different reference subsets. If divisions and cross-divisions of the reference set are performed in the course of the analysis of the individual behaviours of the attributes, it is highly desirable that these divisions are consistent between the attributes.

- *Question 3.* Every additional referrer increases dramatically the complexity of the analysis process. In the case of two referrers, we dealt with two aspectual behaviours. As we have shown in Sect. 3.4.4, in the case of three referrers, the number of aspectual behaviours is six rather than three, and in the case of four referrers, there are 24 different aspectual behaviours. With such complexity, the full extent of the task of characterising the overall behaviour reaches far beyond the cognitive capabilities of a human explorer. However, some possibilities for simplification often exist.

First, depending on the goals of the exploration, not all aspectual behaviours may be of equal interest. Thus, the dataset concerning the simulation of forest dynamics is used primarily for comparison of different forest management scenarios. Therefore, the explorer is interested in characterising the behaviour of the characteristics of the forest corresponding to each scenario and in detecting the similarities and differences between these behaviours. However, the analyst is not very interested in characterising the overall behaviour across the scenarios or any aspectual behaviour referring to the set of scenarios as a whole. Moreover, it is even inappropriate to consider the set of scenarios as a unified whole, because the scenarios are mutually exclusive.

Second, referrers of the population type (i.e. those which have discrete value sets without ordering or distances) are usually easier to deal with than spatial and temporal referrers. Thus, if the values of a referrer are not very numerous, it is possible to treat the slices of the overall behaviour corresponding to different values in the same way as different attributes in a dataset with lower dimensionality. For example, in the case of the forest simulation data, we can treat characteristics related to different tree species as different attributes: the area covered by aspen, the area covered by birch, and so on. If the values of such a referrer are too numerous, it may be possible to categorise them and, on that basis, aggregate the data. For example, we could categorise the species into

coniferous and broadleaved or into hardwood and softwood, compute the total areas for each category, and consider these areas as different attributes. The approaches to dealing with multiple attributes have been discussed earlier.

Third, depending on the explorer's goals, the variation of the characteristics with respect to some of the referrers may be of minor interest. Thus, for a comparison of the forest management scenarios, the spatial distribution of the characteristics of the forest and the spatial variation of the local behaviours are not so important. In such a case, it is appropriate to aggregate the data over the irrelevant referrer(s) and thereby reduce the dimensionality of the data, as was described in Sect. 5.4.1.

When we described the example of the analysis of the burglary rate data, we mentioned that we had achieved a sort of structural linkage between the subpatterns, which emerged from introducing a certain structure (specifically, cross-partitions) into the multidimensional reference set. While introducing a structure may organise and simplify the process of exploration, some phenomena may have an inherent structure. When this structure is known, it needs to be taken into account; if it is unknown but its presence is suspected, it needs to be discovered. This is what the next principle is about.

#### 5.4.9 Principle 9: Establish Structure

It is widely known that many temporally varying phenomena vary in cycles. Thus, yearly cycles are relevant to climate and to various weather- and season-related phenomena and activities such as vegetation, the migration of animals, forest fires, agriculture, and the tourist industry. People's activities are subject to weekly and daily cycles, as are related phenomena such as transport, traffic incidents, and energy consumption. Any exploration of such data cannot be valid if it does not take proper account of the cyclical structure of the behaviour.

The general approach is to split the behaviour into its cyclical and long-term components and to characterise each component. There may be several nested cyclical components, as in the case of the variation of air temperature (daily and yearly) or of people's activities (daily and weekly). Appropriate techniques for splitting numeric time-series data into components and analysing those components exist in statistics, but other sorts of tools may also be appropriate.

One such tool is the appropriate ordering and arranging of display items. Thus, in Fig. 4.12, we represented the local behaviours of the monthly average temperature at various weather stations in Germany by special "mo-

saic” signs containing “tiles” arranged in rows corresponding to the years from 1991 till 2003 and columns corresponding to the months of the year. Each sign represents simultaneously the yearly variation and the long-term tendency, but only for a particular location. When we need to consider the behaviour of the weather over the entire territory of Germany, we cannot apply the same approach, and need to look for other solutions.

Rather than proposing any solutions for this particular case, let us try to find some general approach or principle. If we think of a time-related behaviour with a cyclical structure, we may note that the cyclical structure of the behaviour responds to the cyclical structure of its temporal referrer. Thus, in the case of the climate data, the temporal referrer consists of years, which, in turn, have a common internal structure. Hence, the temporal referrer is not simply a linear sequence of time moments but a chain of repetitive occurrences of one and the same structure. The temporal referrer consists of two or more components, a linear one (the sequence of the years) and a cyclic one (the internal structure of a year, e.g. its division into days, weeks, months, quarters, or seasons). The components of the behaviour reflect the components of the temporal referrer, and splitting the behaviour into components is based on splitting the temporal referrer into its structural components, i.e. *replacing it by two or more referrers*.

Since one temporal referrer is replaced by two or more, *the dimensionality of the data increases*. Thus, in analysing the climate data for Germany, we need to consider its reference set as three-dimensional rather than two-dimensional. The three dimensions are:

- space, specified as a set of discrete locations over the territory of Germany;
- linear time, i.e. the sequence of years;
- the yearly cycle, i.e. the sequence of months in a year.

When we want to visualise the temporal behaviour of a climate attribute in a particular place, we need to use two display dimensions for the representation of the two temporal referrers, the linear and the cyclic one. This is done in the “mosaic” signs in Fig. 4.12: in the subspace of each sign, the vertical dimension represents the linear time and the horizontal dimension the cyclic time.

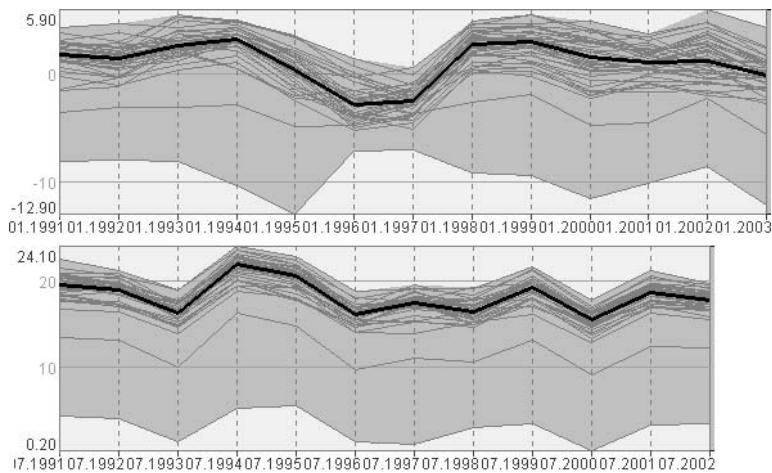
Increasing the dimensionality of the data makes them more and more difficult to analyse. In particular, it may be impossible to visualise such data in full agreement with the principle “see the whole”, i.e. so that all the data can be seen at once and that the display prompts unification. Thus, the visualisation in Fig. 4.12 with multiple mosaic signs scattered over the map of Germany does not support unification: we can see the local behaviours but not the global one.

The approaches to the exploration of multidimensional data that result from splitting a temporal referrer into linear and cyclic components are, basically, the same as for originally multidimensional data. An explorer analyses the data by slices, i.e. selects specific values of one referrer and looks at the behaviour with respect to the other referrers. The explorer can also aggregate the data over the entire value set of one referrer and look at the behaviour of the aggregated characteristics with respect to the other referrers. The explorer needs to consider all aspects of the overall behaviour, including the sub-behaviours with respect to the linear and cyclic times.

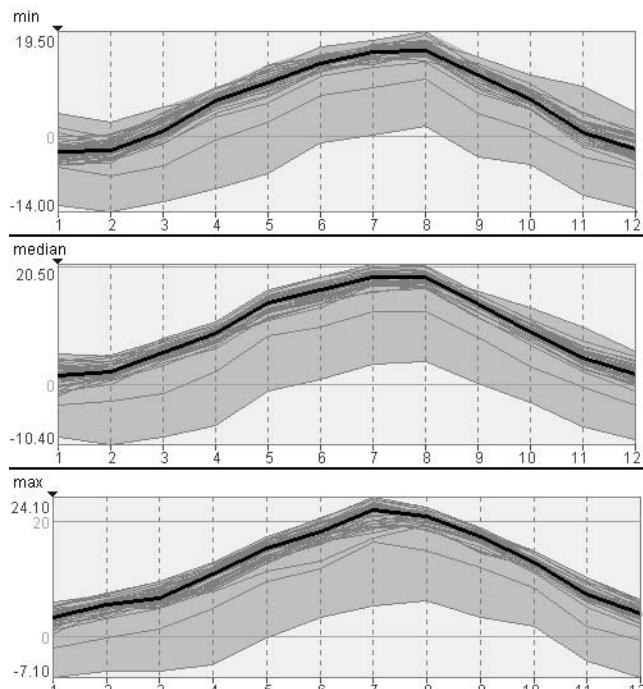
Figures 5.39 and 5.40 present two different aspects of the temporal variation of the mean monthly temperature in Germany. The time graphs in Fig. 5.39 show the behaviours of the January and July temperatures over multiple years. The thin grey lines correspond to the individual weather stations, and the thick black line to the running median among all the weather stations. It may be seen that the long-term behaviour of the January temperatures differs from that of the July temperatures. Analogous displays can also be produced for the other months.

Figure 5.40 demonstrates the cyclic aspect of the behaviour of the mean monthly temperature. To produce the time graphs, the data corresponding to each month of a year were aggregated over all the years; specifically, the minimum, maximum, and various percentiles of the data for the same month and weather station and for different years were found. Of these aggregated characteristics, we have visualised the minima, maxima, and medians. These values are shown in the time graphs as varying over a year. As before, the thin grey lines correspond to the different weather stations, and the thick black lines to the running medians over all the weather stations in Germany.

We have demonstrated two major techniques commonly applied in the exploration of time-referenced data with a cyclic character of variation, namely filtering (i.e. selection of specific elements of a cycle from a sequence of cycles) and aggregation (i.e. grouping of the data referring to specific elements of a cycle over all the cycles and computing summary characteristics of the groups). In Sect. 4.6.1 and 4.6.4, we have mentioned the existence of special tools for temporal querying, such as Time Wheel and temporal brushing, which take account of the possible temporal cycles and allow the user to select elements of a cycle or even several nested cycles. There are also specific aggregation tools that aggregate data by elements of a temporal cycle across many cycles, for example the tool described in Fredrikson et al. (1999). The same operations can also be done using sufficiently flexible general-purpose querying and aggregation tools.



**Fig. 5.39.** The variation of the mean January (top) and mean July (bottom) temperature over several years at various weather stations in Germany



**Fig. 5.40.** The mean temperatures for each month have been aggregated over years here. These time graphs present the yearly variation of these temperatures: top, minimum temperatures for each month from all the years; centre, median temperatures; and bottom, maximum temperatures

We have now discussed the approaches to dealing with cyclically varying behaviours when the cycles are known in advance. There may also be cases where the presence of a cyclical variation in the data may be suspected but the cycle(s) are unknown. In such a case, the explorer needs to discover the cycle(s). For one-dimensional data, a straightforward trial-and-error approach can be used. The idea is to cut the sequence of values into sub-sequences of equal length and visually represent these sub-sequences in horizontally or vertically aligned displays. The user tries different lengths until the maximum similarity between the displays is achieved. The corresponding length will be the length of the cycle. For such an investigation, the user needs an interactive tool that allows him/her to change the interval length and immediately observe the effect of this change.

When the data are multidimensional, and cyclical variation over one of the dimensions (i.e. referrers) is suspected, the explorer can apply the one-dimensional cycle discovery procedure to data corresponding to selected values of the other referrers. If a cycle is detected in such a selected data series, the explorer can check whether the same cycle exists in other series. For this purpose, the analyst can apply the tools and methods suitable for data with a known cycle.

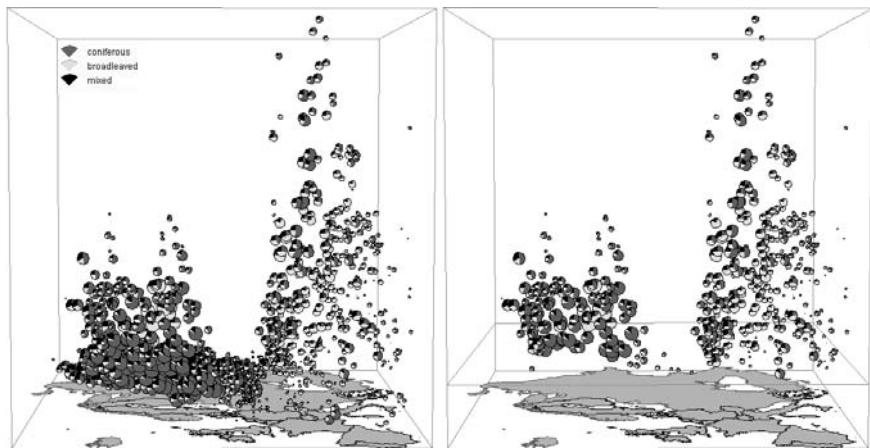
Cyclic variation with regard to time is not the only possible case of a behaviour that has an internal structure. Let us give an example of quite a different kind, using the data for the forest structure of Europe. Although the example is quite trivial, that is, it does not reveal anything really new concerning the distribution of forests, it still demonstrates the idea.

When we look at various representations of the forest structure data (see, for example, Figs 4.86C, 4.87C, 4.132C, 4.133C, and 5.19C), we get the impression that the spatial behaviour of the forest structure consists of two different components. On the one hand, there is a spatial trend: the amount of forest clearly increases from the south to the north, as does the proportion of coniferous forest. On the other hand, there are areas in the centre and in the south of Europe where the amounts of forest are also quite high, and some of them also have a high proportion of coniferous forest. If we compare the behaviour of the forest structure with relief (this can be done using Figs 4.86C and 4.87C), we can note that the high amounts of forest in the centre and south correspond to mountain areas. Hence, we can suspect that the overall spatial behaviour of the forest structure contains, besides the latitudinal trend, an elevation-dependent component also.

In order to investigate how the two components of the overall behaviour are related to each other, it is convenient to treat the dataset as having, besides the spatial referrer, an additional referential component, the altitude.

This is analogous to dealing with a cyclic component of a time-dependent behaviour, where the cycle is regarded as an additional referrer. The altitude can be regarded as one more dimension of the space that the forest structure data refer to: originally, we dealt with the two-dimensional geographical space, and now the space becomes three-dimensional when the altitude component is included.

As a method suitable for the visualisation of data referring to three-dimensional space, we can use, for example, the perspective-view display shown in Fig. 5.41. The horizontal dimensions of this display represent the two-dimensional geographical space, and the vertical dimension represents the altitudinal component. We have chosen the viewing direction so that the territory of Europe is seen from the west; hence, the left side of the display corresponds to the north, and the right side to the south.



**Fig. 5.41.** In order to see how the forest structure is related to the latitude and altitude, the data may be visualised in a perspective view where the horizontal dimensions represent the geographical space and the vertical dimension represents the altitude. In these displays, the viewing direction is from the west of Europe, and hence the left side of a display corresponds to the north, and the right side to the south. The pie charts represent the proportions of coniferous, broadleaved, and mixed forest in the cells of a regular rectangular grid. On the right, the pies are shown only for the cells where the maximum elevation is 500 m or more

The visualisation has been constructed from data originally specified in a raster format but transformed as is described in Sect. 4.5.4.6, i.e. aggregated by cells of a regular rectangular grid. The forest structures in the cells are represented in this perspective view by pie charts located according to the geographic positions of the cells and their maximum altitude. The sectors of the pies represent the proportions of coniferous, broad-

leaved, and mixed forest, and the sizes of the pies show the total proportion of the area of the cell covered by the three types of forest. We have applied a focusing tool to the display so that charts have been drawn only for cells with at least 10% of their area covered by forest (i.e. the sum of the proportions of the coniferous, broadleaved, and mixed forest is not less than 10%). Additionally, the tool allows us to focus on subsets of the data corresponding to specified ranges of altitude. Thus, the screenshot on the right in Fig. 5.41 results from focusing on altitudes of 500 m or higher. The transparent horizontal plane corresponds to an altitude of 500 m; it can be dynamically moved up and down. The representation of the data in the currently selected altitude range can also be “stretched” so as to utilise the whole display height available.

Using the facilities provided by this visualisation tool and a link between it and a map display, we can conveniently and comprehensively explore how the forest structure behaves with respect to the two-dimensional space and to the ground surface elevation. Thus, we can see that for small altitudes (below 500 m), the highest amounts of forest are in the north of Europe, except for the far north. Towards the south, the amount of forest at these altitudes rapidly decreases. The structure also changes from a prevalence of coniferous forests to domination by broadleaved forests.

At altitudes between 500 and 1000 m, the contrast between the north and the south is not so dramatic. In the southern part, the amounts of forest are notably higher than at altitudes below 500 m. Coniferous forest tends to prevail in the north, and broadleaved forest in the south.

For the altitude range from 1000 m to the maximum (3592 m), there are quite a few places in the north where such altitudes are reached. In these places, the amounts of forest are quite small compared with the lower altitudes. Coniferous forest still prevails, while the proportions of broadleaved and mixed forest tend to increase in the northward direction. In the southern part, the amounts of forest at these altitudes vary from rather small to quite large. There is a tendency towards decreasing amounts in the southward direction. At lower altitudes, broadleaved forest clearly prevails in the centre, and coniferous forest in the south. At higher altitudes, which occur mostly in the centre of Europe, the relative proportions of the different forest types vary, but some dominance of coniferous forest can be seen.

In these observations, we have characterised the overall behaviour (i.e. spatial distribution) of the forest structure of Europe as an interplay of two components, latitudinal and altitudinal. Of course, we did not discover anything new with these observations: it is generally known that vegetation is influenced not only by the geographical position (which includes, besides the latitude, many other aspects, such as the closeness of water masses and warm or cold oceanic currents) but also by the elevation. The

reason for giving this example was to demonstrate, using available data, how a behaviour with a suspected internal structure can be explored. The idea is to treat the attribute or phenomenon suspected to be the basis of some component of the behaviour as an additional referrer of the dataset.

In the above example, we were able to visualise the data in such a way that all the referrers, including the additional one, were represented by display dimensions. This was rather convenient, but this is not the only possibility. We could also apply, for example, the technique of cross-partitioning of the reference set, as was described in the previous subsection. We could divide the geographical space into the north, centre, and south and the altitudes into low, medium, and high and then consider the behaviour in each cross-partition: low altitudes in the north, medium altitudes in the north, and so on.

Our practical experience in the exploration of behaviours comprising several components is, unfortunately, insufficient for us to be completely sure that the technique of introducing additional referrers is always helpful. It would be good to have more examples of appropriate data in order to test this. Another problem is how to guess that a behaviour results from an interaction of several components. Generally, we believe that such interactions necessarily manifest themselves in an appropriate data display, i.e. one that complies with the major visualisation principles and the principle “see the whole”. Thus, the presence of a cyclic variation will be seen as a repeated pattern, and the interplay of a spatial trend with something else will appear as a kind of intrusion disrupting the general picture. However, we would like to have more examples to test this and, we hope, to find some general rules for how to detect internal structures in behaviours of phenomena.

Very often, an explorer does not really need to detect the presence of a structure; he/she simply knows that it exists and uses this knowledge in decomposing and characterising the behaviour. Thus, an explorer does not need to detect the presence of a seasonal component in the behaviour of climate, vegetation, or prices of holiday apartments. He/she knows in advance that it exists, and considers separately the variation over a year and the long-term development. Generally, any relevant bits and pieces of domain knowledge can simplify and direct the process of data exploration, and therefore should be used whenever possible.

#### **5.4.10 Principle 10: Involve Domain Knowledge**

We do not feel it really necessary to convince readers of the usefulness of involving domain knowledge in data analysis. It should be quite clear that

domain knowledge can save the explorer's time and effort: he/she will focus rather than wander, distinguish easily between what is relevant and irrelevant and between what is typical and peculiar, and understand quickly what he/she sees through recognising what is known and looking for what is expectable. Besides, domain knowledge can prevent the explorer from making errors and coming to wrong conclusions.

In fact, we have mentioned the involvement of domain knowledge already many times throughout this chapter, as we discussed the various general principles of exploratory data analysis. Let us now try to bring all these mentions together for easier reference.

1. In discussing the principle "simplify and abstract", we have said that simplification may be achieved by means of attribute integration, which must be based on appropriate domain knowledge. In fact, not only integration but also virtually any attribute transformation requires the nature of the attribute(s) to be properly understood. Thus, before transforming absolute values into relative values, the explorer must know definitely that the original attribute values are absolute and must understand what they can be related to (population, area, an established standard, etc.). Accumulation over time is possible only for attributes that express quantities of new items that have appeared in each moment of the measurement as compared with the previous moment.
2. In the subsection dealing with "divide and group", we discussed the fact that dividing/grouping can be based on the explorer's domain knowledge. This knowledge induces certain expectations concerning the diversity of the behaviour with respect to specific groups/subsets of references. For example, in medical studies, differences may be expected between males and females, in geography-related studies, differences may be expected between coastal and inland regions, etc. Later, in discussing the principle "establish structure", we also mentioned that the reference set of a time-referenced phenomenon can be divided into linear and cyclic components when the explorer's domain knowledge suggests that the phenomenon may vary in cycles.
3. With regard to the principle "look for recognisable", we have mentioned that the explorer's expectations concerning the sort of patterns that may be present in the data direct the choice of the visualisation methods, as well as the appropriate data transformations and querying tools. Furthermore, the explorer knows how the expected patterns may show up in the display obtained, and looks purposefully for particular visual elements or particular arrangements of visual elements.
4. In the discussion of the principle "attend to particulars", domain knowledge has been mentioned in two contexts. First, when the analyst en-

counters “strange” things differing significantly from the bulk of the data, domain knowledge may help him/her to understand the reason for this strangeness. Second, the analyst may intentionally pay attention to expectable deviations from the bulk of the data and give them a different treatment, for example, the dates of public holidays in time-referenced business data or big cities in demographic studies.

5. The use of domain knowledge may be extremely helpful in attempts to “establish linkages”, i.e. build a unifying overall pattern from a number of partial patterns characterising different parts, slices, and aspects of the overall behaviour. In fact, establishing linkages starts not when the partial patterns have been produced but at the stage of decomposing the overall behaviour into parts, slices, and/or aspects. Domain knowledge, when available, can help significantly in finding an appropriate decomposition that reflects pertinent links and essential differences. Domain knowledge may suggest how to divide the reference set into subsets and what attribute groups to consider, what attributes or phenomena may influence other attributes or phenomena, and what may be the scope of this influence in space and the latent period in time after which the effect of a change may appear.
6. In “establishing structure”, the explorer may know in advance what structural components exist or might exist in the behaviour of some phenomenon. Then, the task is to “distil” and characterise each component, as well as how the components are related. Thus, the explorer often either knows definitely or at least suspects what cycles exist in a time-referenced phenomenon.

We have thus produced a list of cases in which the use of domain knowledge is either necessary or helpful. We do not think that this list is complete; however, there is no need to try to make a complete enumeration: an analyst who has relevant domain knowledge will intuitively recognise the situations where this or that piece of knowledge may be helpful. What is really needed is that the tools that the analyst uses allow him/her to take this piece of knowledge into account. So, let us try to review the arsenal of tools for data analysis from the perspective of how domain knowledge may be involved in the use of these tools to increase the effectiveness and efficiency of the operations performed.

Many visualisation tools permit one to involve domain knowledge through the use of the display manipulation tools attached to them. Thus, when an explorer uses tools for ordering or arranging visual items, he/she may use his/her domain knowledge in defining the ordering or arrangement. Tools for the elimination of excessive detail usually allow the user to specify the appropriate degree of simplification and level of detail, and this

may be based on the user's domain knowledge. Tools for interactive classification also provide suitable opportunities for involving domain knowledge, which may influence the definition of classes. Zooming and focusing tools allow the user to focus on subsets of the data of primary interest, according to the user's notion of "interestingness" and his/her expectations as to the part of the data where the most interesting things may be found. In using tools for visual comparison, an analyst may specify the reference values on the basis of his/her knowledge of existing standards, typical characteristics, or danger thresholds.

Many tools for data transformation not only allow the user to involve domain knowledge but also assume that domain knowledge is involved. This concerns, first of all, tools for attribute transformation. We have already mentioned attribute integration, and computation of relative values from absolute ones. In this context, we would also like to say a few words concerning one of the examples of data considered in this book, specifically, the climate data for Germany.

Climate is a spatially continuous phenomenon; therefore, it is appropriate to view the climate data as distributed continuously over the whole territory. However, in the dataset that we have at our disposal, the values of the climate attributes are specified only at sample locations – weather stations. In order to view the data as continuous, the data need to be interpolated between the sample locations so that values at any location can be accessed. The operation of computational tools for spatial interpolation is based on a certain definition of the notion of neighbourhood, i.e. what sample locations should be considered as neighbours of a given arbitrary location. In some cases, a formal definition of the neighbourhood may suffice; for example, all sample locations within a specified distance from the given location may be treated as its neighbours. However, this approach is not valid in climate studies. Two locations may have quite different climates despite there being a short distance between them if they are separated by a mountain range or if one of them is at a high altitude and another in a valley. There are also other factors that influence the climate. Therefore, interpolation of climate data from sample locations to the whole territory must be controlled by an expert, who defines the neighbourhoods on the basis of domain knowledge concerning climate variation.

Of the tools for data aggregation, some tools allow users to define arbitrary aggregates, while others aggregate data on the basis of regular intervals, grids, or other formal methods. In the first case, users can involve their domain knowledge in the definition of the aggregates; in the second case, users may choose an appropriate degree of granularity for the division, according to their domain knowledge.

Querying tools may be divided into two major categories:

- dynamic query tools, which are typically quite restrictive concerning the sort of questions they are intended to answer but are easy to use and quick in response;
- comprehensive query tools, which provide opportunities for the formulation of a wide variety of queries by using special query languages but are more difficult to use and less dynamic.

The latter category of tools provides better opportunities for involving domain knowledge than does the former. With comprehensive query tools, the user may select arbitrary subsets of data for further examination, arrange these subsets in aggregates, and obtain various summary statistics. The selection of the data may be based on expected differences in the behaviour or on a knowledge of the structural components present in the behaviour, such as temporal cycles. For example, a query tool allowed us to choose the January temperatures for all years in the German climate data in order to look at the long-term behaviour of these temperatures on a time graph in Fig. 5.39, and then we did the same for the July temperatures. Aggregation of the temperatures for each individual month over multiple years, which was used for constructing the displays in Fig. 5.40, is also possible with the use of comprehensive querying tools that provide sufficiently powerful query languages.

Although computational tools for data analysis are supposed to find various tendencies, dependencies, and regularities in data automatically as well as anomalies, the involvement of a human analyst's domain knowledge is still possible and useful. Of course, this knowledge cannot be directly entered into a computational tool in order to be used in the course of the operation of the tool; however, indirect ways exist. Some computational tools allow the user to specify a sort of template for the patterns that the tool must look for, or some criteria for the evaluation of the partial results that the tool achieves in order to prune useless search directions. Such templates or criteria are formulated on the basis of domain knowledge.

Even when a tool does not suppose the user to direct its operation in any way, the user can still take the available domain knowledge into account by dividing the data into subsets according to expected substantial differences in their behaviour and running the tool separately for each subset. For example, if we decided to analyse the stork movement data with the use of a statistical or data-mining tool, we would apply the tool separately to three subsets of the movements: (1) the movements that took place in August and the beginning of September, when the storks flew from Europe to Africa; (2) the movements inside Africa that occurred in the period from September to February of the next year; (3) the movements during the period when the storks returned to Europe.

Many data-mining tools can work only with attributes that have discrete value domains. In order to apply such tools to attributes with continuous value domains, such as numeric, temporal, or spatial attributes, one needs to discretise their value domains, i.e. introduce some equivalence classes of values. In defining the equivalence classes, the explorer can and should apply the available domain knowledge. Thus, if we were to classify the July temperatures for Germany into low, medium, and high temperatures, we would do this separately for the two stations located in mountain regions and for the remaining stations, since the notion of low and high temperatures differs between high and low altitudes.

It appears that the most universal way to involve domain knowledge in data analysis is through appropriate dividing/grouping, which is applied to the set of references or to the values of attributes.

At this point, let us finish the individual discussion of each principle and try to bring them together and relate them to data analysis tasks.

## 5.5 General Scheme of Data Exploration: Tasks, Principles, and Tools

The primary goal of our study has been to establish the principles for choosing appropriate tools for exploratory data analysis. The main idea is that the tools must support finding answers to the various questions that can potentially arise in the course of data analysis. We call these questions data analysis tasks.

In Chap. 3, we have shown that the potential tasks are determined by the structure of the data under analysis, in particular, the division of the data components into referential components and characteristics, and by the properties of the data components. Hence, knowing the structure of a dataset and the properties of its components, one can, in principle, enumerate all the tasks that can arise in the course of an analysis of this dataset.

The tasks differ in their generality level, and less general tasks typically appear as subtasks in the course of performing more general tasks. The most general task is the task of characterising the overall behaviour of the phenomenon reflected in a dataset. This task is often accompanied by the task of explaining the overall behaviour, which is classified as connection discovery in our framework.

In order to understand what kind of tool could be appropriate for any given type of task, the structure of the task needs to be considered. According to the general model of a task introduced in Chap. 3, a task consists of two parts, the target, i.e. an indication of the unknown information that

needs to be obtained, and the constraints, i.e. a specification of the known information, which is related to the target in a certain way and limits the set of items of information that are suitable as an answer. A tool can support the performance of a task in one of two different ways:

- The tool allows the user to ask the question explicitly, i.e. to specify the target and set the constraints, and, in response, provides the required information. This operation mode is realised in some querying tools.
- The tool represents the data visually, in particular, the components and items relevant to the task. Either the organisation of the information in the display or some additional tools, for example an index, attached to the display allow the user to identify the display items corresponding to the constraints of the task. The required information that the target refers to is represented in these display items or in their surroundings, and the user needs to extract this information by viewing the appropriate parts of the display.

As we have discussed in Sect. 4.6, querying tools are suitable mostly for elementary tasks. The second mode of obtaining answers to the explorer's questions is more universal. It is in this way that synoptic tasks are typically performed.

Hence, the *general strategy* for choosing a tool or tool combination to accomplish some task is to seek such a tool or combination that can

- appropriately represent the information referred to in the task target, and
- allow the explorer to locate effectively the display items satisfying the constraints.

The appropriateness of the representation means that the required information must be perceivable from the display items, i.e. the items must be legible, the information encoded clearly, etc. For synoptic tasks, it is also highly desirable that the display items providing the required information can be perceived as a unified whole.

Thus, for the most general task of characterising the overall behaviour of a dataset, an ideal supporting tool is a visualisation tool which

- represents the entire reference set of the data by appropriate display dimensions so that the essential relations between the elements of the reference set, such as ordering and distances, are reflected; and
- represents all the characteristics from which the overall behaviour is formed, in a way that promotes perceptual unification.

In fact, this is what the principle “see the whole” basically says. Hence, this principle may be viewed as an outcome of applying the general strat-

egy for choosing a tool for a task to the high-level task “characterise the overall behaviour of the attributes of the dataset over the reference set”.

The other principles (perhaps with the exception of “involve domain knowledge”, which is fairly self-evident) also follow logically from the consideration of certain task categories and trying to define the appropriate supporting tools or the requirements of such tools. Besides taking into account the structure of the task, i.e. what is in the target and what is in the constraints, we also thought about the possible complications that may be caused in the analysis by certain characteristics of the data, such as

- multidimensionality of the data, i.e. the presence of several referrers;
- multiple attributes that need to be jointly analysed;
- a very large data volume, i.e. a great number of elements in the reference set.

The principles that we have thus formulated, on the one hand, suggest approaches to performing various types of tasks, and define the tools that can support this. On the other hand, these principles suggest approaches to dealing with various complexities that may pertain to the data under analysis, and define the tools that can help in this. In the description of each principle, we have referred to the type(s) of tasks and/or to the complexities that it applies to, and the tools that can help in implementing this principle. Now, we would like to make a sort of summary in which the links between the principles, the task categories, and the tools are reiterated in a maximally explicit way.

However, we would not like to summarise our study by going once again through the list of principles and saying what tasks and tools they correspond to. We would also not like to go through the list of task categories and say what principles and tools are relevant to each category. Analogously, we would not like to go through the list of tool categories relating them to tasks and principles. In all of these approaches, each task type is considered separately from the others, whereas we would like to bring all task categories together by defining their places in the common context of exploratory data analysis. So, let us try to do this.

As we have already said, exploratory data analysis may be viewed as accomplishing the general task “characterise the overall behaviour of the phenomenon represented by a given dataset”, which is equivalent to “characterise the overall behaviour of the characteristics contained in the dataset over the entire set of references”. When the phenomenon consists of several parts, the task of characterising its behaviour implies that the links between the parts are also revealed and characterised; hence, the highest-level behaviour characterisation task may include subtasks of the connection discovery type, but certainly not only subtasks of this type. In the

course of performing the top-level task, this task undergoes gradual decomposition: subtasks of various types arise, and their results contribute to the final result of the original task.

Hence, we can try to bring all task types together by envisioning how the process of performing the overall behaviour characterisation task may develop, and by determining the places of the other task types in this process. In parallel, we shall relate the stages of the process to the general principles of data exploration introduced in this chapter and to the major categories of tools that may be appropriate for the tasks that the stages comprise.

We have previously mentioned that an ideal tool for accomplishing an overall behaviour characterisation task is a visualisation that allows the explorer to grasp the entire behaviour as a unified whole, i.e. it complies with the principle “see the whole”. Although various complexities of the data make this optimum rarely achievable, let us start with the situation where such a holistic view of the entire behaviour is possible. Moreover, we shall assume, to begin with, that the dataset contains only one referential component, for example time, or space, or a population, but not a combination of two or more referrers. Later, we shall consider how the exploration process will change in response to the possible complications from the data side, i.e. a large data volume, multidimensionality, and multiple attributes that cannot be visualised together in a holistic manner.

### 5.5.1 Case 1: Single Referrer, Holistic View Possible

As we have said, the visualisation must properly reflect the essential properties of the reference set, in particular, the presence of ordering, distances, and/or other relevant relations between elements. The same requirement also applies to the representation of the characteristics, i.e. the attribute values associated with the elements of the reference set. To fulfil this requirement, the data display must be built according to the principles of visualisation overviewed in Sect. 4.3.

Some peculiarities of the data may prevent one from getting a clear picture of the behaviour. Thus, minor fluctuations of characteristics can obscure the view of the general character of the behaviour. Or excessive detail may attract too much attention and obstruct “seeing the wood for the trees”. It may also happen that the presence of outliers in the data makes the representation of the bulk of the data insufficiently expressive (see, for example, the map on the left in Fig. 4.31, which represents the population densities in the districts of Portugal). In such cases, the visualisation or the underlying data should be transformed in order to make the view simpler

and clearer and to allow the analyst to abstract from particulars and concentrate on generalities. This is what the principle “simplify and abstract” is about. The tools that can lead to display simplification include ordering and other ways of arranging the display items, smoothing, cartographic generalisation, outlier removal (focusing), transformation of the visual encoding function, classification, and aggregation.

When the display is sufficiently simple and expressive for the explorer to perceive the entire behaviour represented in it, the explorer can note whether the behaviour is homogeneous throughout the reference set or heterogeneous.

### **5.5.1.1 Subcase 1.1: a Homogeneous Behaviour**

A behaviour can be homogeneous in one of two senses:

- The characteristics are invariant (constant) throughout the reference set, i.e. all references have the same characteristics.
- The characteristics change from one reference to another in a regular way, which is invariant throughout the reference set. For example, the value of a numeric attribute may increase over time at a constant rate.

For brevity, we shall refer to these meanings as “invariant characteristics” and “regular change”, respectively.

Of course, it rarely occurs in reality that any characteristics are absolutely invariant or that the changes of characteristics are absolutely regular. Rather, it is possible that an explorer will regard a certain degree of variation as negligible. Hence, it is better to say “nearly invariant” or “quasi-invariant”, and “nearly regular” or “quasi-regular”, respectively.

When an explorer finds that the behaviour under analysis may be regarded as a quasi-invariance of the characteristics, he/she characterises this behaviour by indicating its general character, i.e. quasi-invariance, and by specifying the characteristics associated with the reference set. In a case of real invariance rather than quasi-invariance, the specification of the characteristics consists of a single value of each attribute. In a case of quasi-invariance, the specification includes the subset of values of each attribute that occurs throughout the reference set. This may be supplemented with elementary statistics such as the value frequencies, the mean, the mode, etc. In Sect. 3.4.3, we have called this sort of pattern a “distribution summary”. The generation of such a pattern is supported by, in addition to the visualisation, querying or computational tools that may provide the necessary statistics.

When the explorer regards a behaviour as subject to a quasi-regular change, he/she characterises the behaviour by referring to its general char-

acter, as in the previous case, and specifying relevant characteristics of the change, such as

- the character of the change, for example an increase, decrease, oscillation, or repeated succession of states;
- a more detailed description of the change in terms of the attribute values involved, for example the starting and final values in the case of an increase or decrease, the minimum and maximum values in the case of an oscillation, the sequence of values in the case of a repeated succession of states;
- the rate/frequency/period/amplitude of the change;
- the direction of the change with respect to the reference set (in the case of the reference set that has no linear order), for example the direction in space from north to south.

This sort of pattern has been called an “arrangement” in Sect. 3.4.3. The tools needed for building such a pattern are visualisation, querying (including the measurement of distances and possibly other relations), and data transformation, in particular, computing changes.

It may be noted that a number of elementary subtasks are involved in building the pattern:

- direct lookup (i.e. ascertaining the attribute values associated with particular references) – in describing the change in terms of the attribute values involved, determining its rate or amplitude, etc.;
- inverse lookup (i.e. finding the references corresponding to specific characteristics) – in ascertaining the rate, frequency, or period of the change;
- direct comparison (i.e. determining the relations between characteristics) – in identifying the character of the change and ascertaining its rate or amplitude;
- inverse comparison (i.e. determining the relations between references) – in ascertaining the rate, frequency, or period of the change and its direction;
- relation-seeking (i.e. detecting references with characteristics related in a certain way) – in determining the period of oscillation (where does the increase change to a decrease and vice versa?) or of a repeated succession of states (where does the state  $S_n$  change back to state  $S_0$ ?).

The procedure for characterising a homogeneous behaviour is summarised in Table 5.6.

**Table 5.6.** Characterisation of a homogeneous behaviour (subcase 1.1)

Actions	Subtasks	Tools	Principles
Identify the character of the homogeneity: (A) invariant characteristics; (B) regular change.	Visualisation	See the whole	
(A) Get/produce a distribution summary	Elementary lookup (what values occur?)	Querying Computation (elementary statistics)	
(B.1) Specify the character of the change	Visualisation	See the whole	
(B.2) Determine various qualities and measures of the change	Elementary lookup, comparison, and relation-seeking	Querying Data transformation (e.g. change computing)	See in relation

### 5.5.1.2 Subcase 1.2: a Heterogeneous Behaviour

The general idea in characterising a heterogeneous behaviour is to divide the reference set into subsets such that the behaviour over each subset can be treated as homogeneous. The underlying principle is “divide and group”. In Sect. 5.4.3, we have considered various approaches to dividing/grouping and the tools that support this procedure. Since the behaviour over each reference subset resulting from the dividing/grouping is supposed to be homogeneous, the procedure suggested above for characterising homogeneous behaviours (subcase 1.1) may be applied to each of these behaviours. In the result, each part of the behaviour corresponding to a reference subset will be approximated by a separate pattern. These separate partial patterns, or subpatterns, must then be combined into a single overall pattern, i.e. the principle “establish linkages” must be applied to them.

However, the procedure of dividing the reference set into parts with internally homogeneous behaviours is not always appropriate. A behaviour may consist of two or more interrelated structural components, which cannot be separated by means of dividing the reference set because they overlap greatly or simply exist everywhere over the reference set. An example is the presence of linear and cyclic components in a time-based behaviour. In the subsection dealing with the principle “establish structure”, we have also given the example of the spatial behaviour of forest structure, which has a latitudinal and an altitudinal component.

We have explained that such structured behaviours can be analysed by means of splitting one referential component into two or more components or, as in the case of the forest structure, introducing additional referrers derived from other attributes or phenomena. This increases the dimensionality of the data; hence, the data resulting from the transformation need to be analysed by applying a procedure devised for multidimensional data, which will be described later.

Let us now return to the situation where a heterogeneous behaviour can be split into internally homogeneous parts by means of dividing the reference set. As we have described in the respective subsection, the major mechanism for linking subpatterns defined on the basis of dividing the reference set is to specify, with adequate precision, the validity domain of each subpattern, i.e. the reference subset on which it is based. We have noted that, in principle, this basic linkage may be sufficient; however, the explorer may wish or need to relate the subpatterns additionally by revealing their similarities and differences. We have also noted that the explorer may even try to go beyond deriving a merely descriptive pattern and discover possible interactions between the parts of the entire behaviour, such as correlations or influences. We have given an example of seeking such interactions in the case of climate data; however, the result of our investigation was negative rather than positive.

Table 5.7 summarises the procedure for characterising a heterogeneous behaviour through reference set partitioning and refers to the relevant sub-tasks, tools, and principles.

So, we have reviewed how to characterise a behaviour in the case where the dataset contains a single referential component and the corresponding characteristics can be visualised in such a way that a viewer perceives them as a unified whole, as an image of the behaviour. We have considered two subcases: first, when the behaviour can be treated as homogeneous, and second, when it is heterogeneous but can be divided into homogeneous parts. The analysis procedure applied in the second subcase includes the procedure for the first subcase.

Before the consideration of these subcases, we mentioned that various simplification measures may be applied to the display and/or to the data. Some of them, such as reordering, do not reduce the amount of information present in the display, while others, such as smoothing, aggregation, or outlier removal, involve information loss. If some information reduction has taken place in the initial stage of data analysis, it is appropriate, after the general pattern has been constructed, to pay attention to the information previously removed and to describe how it is related to the general pattern. In other words, after simplification, abstraction, and concentrating on generalities, it is time to “attend to “particulars”.

**Table 5.7.** Characterisation of a heterogeneous behaviour through reference set partitioning (subcase 1.2)

Actions	Subtasks	Tools	Principles
<p>Estimate the character of the heterogeneity:</p> <p>(A) the behaviour is divisible into parts based on different reference subsets;</p> <p>(B) the behaviour consists of two or more interrelated components</p>	<p>Relation-seeking: locating behaviour changes</p> <p>Behaviour/pattern comparison</p> <p>Pattern search: looking for recognisable pattern types</p>	<p>Visualisation</p> <p>Various tools for dividing/grouping: classification, querying, clustering</p> <p>Arranging (e.g. permutation, juxtaposing, overlaying)</p>	<p>See the whole</p> <p>Divide and group</p> <p>See in relation</p>
<p>(A.1) Divide the reference set into subsets so that the respective behaviours are homogeneous</p>	<p>See above<sup>a</sup></p>	<p>See above<sup>a</sup></p>	<p>See above<sup>a</sup></p>
<p>(A.2) Characterise each partial behaviour (go to <i>subcase 1.1</i>)</p>	<p>Behaviour characterisation (for each subset)</p>	<p>Focusing or query tools to focus on each partial behaviour</p> <p>See also <i>subcase 1.1</i></p>	<p>Zoom and focus</p>
<p>(A.3) Specify the validity domain of each subpattern</p>	<p>Inverse lookup: establish the limits of each subpattern in the reference set</p> <p>Inverse pattern comparison: position the reference subsets that the subpatterns are based on in relation to each other</p>	<p>Querying</p>	<p>Establish linkages</p>
<p>(A.4) Reveal the similarities and differences between the subpatterns</p>	<p>Behaviour/pattern comparison</p>	<p>Visualisation</p> <p>Display multiplication</p> <p>Arranging of the subpattern displays</p>	<p>See in relation</p>
<p>(A.5) Discover correlations and influ-</p>	<p>Connection discovery</p>	<p>Visualisation</p> <p>Display multiplication</p>	<p>Establish linkages</p>

Actions	Subtasks	Tools	Principles
ences between the subpatterns		Arranging of the subpattern displays	
(B.1) Split the structural components by introducing additional referrers		Querying Aggregation	Establish structure
(B.2) Characterise the resulting multi-dimensional behaviour (go to <i>case 2</i> )	Behaviour characterisation Connection discovery	See <i>case 2</i>	Establish structure Establish linkages

<sup>a</sup> The partitioning of the reference set is first done in a “trial” mode in order to estimate the feasibility of this approach. If the approach is judged appropriate (case A), a “conclusive” division takes place, which is referred to in item A.1.

If the display or the data have been simplified by means of smoothing, it is appropriate to compute the residuals, i.e. the differences between the original and transformed values. If all the residuals can be regarded as small, the explorer may leave the derived pattern as it is, or perhaps add some numeric measure of the imprecision of the pattern (for this purpose, statistical techniques can be applied). If some residuals are large, the explorer needs to consider them specially, according to the principle “attend to particulars”. The same applies to outliers previously removed to make the data display more expressive and to make the general character of the behaviour more prominent. The “particulars” need to be characterised and, whenever possible, explained. These characterisations and explanations complete the general pattern.

The procedure for analysing a behaviour over a reference set composed of values of a single referrer can be summarised as is shown in Table 5.8.

### 5.5.2 Case 2: Multiple Referrers

When two or more referential components are present in a dataset, it is often impossible to visualise the overall behaviour in such a way that it can be perceived as a unified whole. However, there are cases where this is possible. If, for example, we have the results of some measurements of the traffic density along a road over time, it is possible to represent the two referrers of this dataset, the position along the road and the time, by two spatial display dimensions. The values of the attribute, the traffic density, can be encoded in the sizes or degrees of darkness of marks drawn within

**Table 5.8.** Characterisation of a holistically representable behaviour over the value set of a single referrer (case 1)

Actions	Subtasks	Tools	Principles
Estimate whether fluctuations and outliers in data obstruct seeing the character of the behaviour: (A) yes; (B) no		Visualisation	See the whole
(A.1) Simplify the view, abstract from details and particulars	Smoothing Aggregation Outlier removal (focusing) Reordering	Simplify and abstract	
(A.2), (B.1) Estimate whether the behaviour is (C) homogeneous or (D) heterogeneous	Visualisation	See the whole	
(C) Characterise the behaviour as homogeneous (go to <i>subcase 1.1</i> )	Behaviour characterisation	See <i>subcase 1.1</i>	See <i>subcase 1.1</i>
(D) Characterise the behaviour as heterogeneous (go to <i>subcase 1.2</i> )	Behaviour characterisation	See <i>subcase 1.2</i>	See <i>subcase 1.2</i>
(A.3) Characterise the deviations from the general pattern	Elementary lookup and comparison	Computation (of residuals) Querying	Attend to particulars

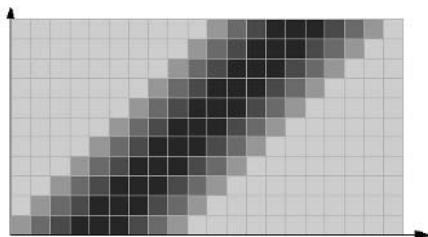
the coordinate system thus formed. We have also demonstrated some examples of the use of three-dimensional display space simulated in a perspective view; see Figs 5.22, 5.24, and 5.41. In these examples, two dimensions represent geographical space. Although the geographical space is considered as a single referrer, it has two dimensions of its own. For this reason, the visual representation of it consumes two display dimensions. The third display dimension in Figs 5.22 and 5.24 represents time, and in Fig. 5.41 it represents altitude.

In this subsection, we shall consider both the situation where a holistic representation of the entire behaviour is possible and the situation where this is impossible, since some of the steps and methods of the analysis are common to both situations. We shall start with the situation where a holistic view is possible.

### 5.5.2.1 Subcase 2.1: Holistic View Possible

The visual analysis of a data display where a combination of several referrers is mapped onto a unified display space differs from the consideration of a display of a dataset with a single referrer, even when the two displays have the same dimensionality and appear quite similar. Thus, a two-dimensional display representing the distribution of the traffic density along a road over time differs inherently from a two-dimensional display of the distribution of the values of a numeric attribute over a two-dimensional physical space. The analyst must always remember the meaning of each display dimension in order to interpret correctly any shape or structure visible in the display and, ultimately, to characterise the behaviour adequately.

Consider, for example, the display in Fig. 5.42, representing an artificial behaviour of a numeric attribute. The values of the attribute are portrayed by the degrees of darkness of the square marks drawn in the two-dimensional display space. If the display space reflects two-dimensional geographical space, the structure visible in the display can be interpreted as a zone of high values extended linearly in the north-east–south-west direction. If one of the display dimensions represents a one-dimensional space, such as the extent of a road, and the other dimension represents time, the structure means a zone of high values, the spatial position of which changes over time. Thus, if we assume that the display represents the variation of the traffic density along a road over a period of time, the structure may mean a traffic congestion zone that has formed around several slow vehicles moving together. The zone shifts gradually along the road as the vehicles move.



**Fig. 5.42.** The interpretation of a shape or structure visible in a data display depends on whether the display space represents a single referrer or a combination of two or more referrers

In considering the case of a single referrer, we pointed to dividing/grouping as the key approach to characterising a heterogeneous behaviour. In the case of several referrers, the idea of dividing the behaviour into

sufficiently homogeneous parts does not seem quite appropriate. In this case, the primary approach is a search for interpretable patterns, i.e. applying the principle “look for recognisable”. This requires, first of all, adequate visualisation and display manipulation tools. As we have mentioned in the subsection dealing with the principle “look for recognisable”, specific computational and querying tools may also be helpful.

As in the single-referrer case (case 1), the general character of a behaviour based on multiple referrers may be difficult to grasp from a visual display owing to outliers, excessive fluctuations in the data, or the absence of any apparent organisation among the display items. In such situations, it is appropriate to try to simplify the display, according to the principle “simplify and abstract”. We have previously mentioned reordering, smoothing, aggregation, and outlier removal as techniques that support simplification and abstraction. Outlier removal is done in the multidimensional case in basically the same way as in the one-dimensional case. Reordering, if allowed by the properties of the reference set, is done separately with respect to each display dimension. Recall that reordering of display items positioned along a display dimension is allowed when the component represented by this dimension has no inherent ordering among its elements. When two display dimensions represent two unordered referential components, i.e. the display has the form of a matrix filled with marks, it is possible to change the relative positions of the rows or columns of this matrix but not to arbitrarily move the individual marks within the matrix. Smoothing in a multidimensional space jointly formed by several referrers requires an appropriate definition of the neighbourhood of a given reference, which may be rather non-trivial, taking into account the different natures and properties of the referrers. Thus, the application of smoothing to spatially and temporally referenced data requires the definition of the spatio-temporal neighbourhood of an element of the reference set, which is a combination of a spatial location and a time moment.

If outlier removal or smoothing has been applied, the explorer needs to attend to the omitted information after characterising the general character and properties of the behaviour.

Besides joint consideration of the multiple referrers, it is usually also appropriate to focus on each referrer individually and study how the attributes behave with respect to it. This is done in the situation where the dimensionality of the data does not permit a holistic view of the overall behaviour. In such a situation, this is in fact the only possible approach. The exploration of aspectual behaviours, i.e. behaviours with respect to particular referrers, is considered below as subcase 2.2. Before that, we summarise what has been said concerning case 2.1 in Table 5.9.

**Table 5.9.** Characterisation of a behaviour based on a multidimensional reference set but still allowing a holistic view (subcase 2.1)

Actions	Subtasks	Tools	Principles
Estimate whether fluctuations and outliers in data obstruct seeing the character of the behaviour: (A) yes; (B) no		Visualisation	See the whole
(A.1) Simplify the view, abstract from details and particulars	Outlier removal Reordering Smoothing Aggregation	Simplify and abstract	
(A.2), (B.1) Detect and describe interpretable shapes and arrangements of display items	Pattern search	Visualisation Focusing Querying Computation	Look for recognisable
(A.3), (B.2) Compare the subpatterns detected	Direct pattern comparison: compare pattern properties Inverse pattern comparison: compare positions with respect to the reference set	Visualisation Querying Computation (summary statistics)	See in relation
(A.4) Characterise the deviations from the general pattern	Elementary lookup and comparison	Computation (of residuals) Querying	Attend to particulars
(A.5), (B.3) Characterise the aspectual behaviours (go to subcase 2.2)	Behaviour characterisation Connection discovery See also subcase 2.2	See <i>subcase 2.2</i>	Establish structure Establish linkages
(A.6), (B.4) Join the subpatterns and the aspectual patterns into a unified overall pattern	Pattern comparison Connection discovery	Visualisation Display coordination	Establish linkages See in relation

### 5.5.2.2 Subcase 2.2: Behaviour Explored by Slices and Aspects

When the behaviour with respect to a selected referrer is examined (we call such a behaviour “aspectual”), there are two opportunities concerning the other referrer(s):

- slicing: Specific values of the other referrer(s) are selected, and the behaviours of the corresponding characteristics over the referrer in focus are investigated;
- aggregation: The characteristics are aggregated over the other referrer(s), and the behaviour of the aggregate characteristics over the referrer in focus is explored.

Which of these techniques to apply depends on the nature of the data and the goals of the analysis. In some cases, it may be reasonable to apply one of them, and in other cases both.

Thus, in the case of the hypothetical traffic density data (which would not necessarily distributed in space and time as is shown in Fig. 5.42), it would be useful to consider the variation of traffic over time by fragments of the road as well as the temporal variation of the aggregate characteristics of the entire road. On the other hand, it might be interesting to consider the distribution of traffic along the road at different time moments as well as the aggregated characteristics of the traffic along the road for the entire time period.

An example where only one technique may be appropriate is the case of the monthly variation of the temperature in a particular place or set of places during a time period of several years. According to the principle “establish structure”, the exploration of this dataset involves splitting the temporal referrer into two referrers reflecting the linear and cyclic components of the time (see also subcase 1.2). It is reasonable then to investigate the variation of the temperature with respect to the linear component by means of slicing, for example to consider the long-term variation of January temperatures, February temperatures, and so on (see Fig. 5.39). It may be less meaningful to aggregate the temperatures for each year into, for example, the average or median yearly temperature and to consider the variation of this aggregated characteristic. The major reason is that the temperature usually varies greatly over a year. At the same time, the aggregation of January, February, ... temperatures over many years can make sense (see Fig. 5.40).

The exploration of behaviour slices in the case where the entire behaviour can be represented in a single display does not, in principle, require any additional tools. Thus, with a display organised like that in Fig. 5.42,

an explorer may focus his/her attention on any row or any column. Appropriate focusing or marking tools, such as “muting” of the irrelevant display items or drawing a frame around the slice currently being examined, can make concentration easier.

In the case where a holistic display of the entire behaviour is impossible, slicing can be done with the help of querying tools. Dynamic querying tools are typically not intended for such operations, and one needs to apply comprehensive tools with sufficiently powerful query languages. In some cases, slicing may also be done with a visualisation tool that allows the user to select a subset of the data for visualisation. However, visualisation tools are rarely flexible to such an extent that one may choose only data referring to a particular month over a sequence of years. Slicing of a dataset with linear and cyclic temporal components can be done by means of specialised query tools such as Time Wheel or “temporal focusing”.

Irrespective of whether slicing or aggregation is applied, the dimensionality of the data is reduced. If it has been reduced to a single dimension (i.e. a single referrer), the resulting selected or transformed data can be analysed as in the single-referrer case discussed earlier. If the data still contain multiple referrers, the procedure described in this section is applied recursively. However, individual behaviour slices do not usually undergo such deep and detailed examination as is suggested in Tables 5.6–5.9. It may be sufficient just to grasp the general character of the behaviour in each slice.

Generally, the exploration of a behaviour by slices does not suppose that each slice is considered in isolation from the other slices. It should not be forgotten that an explorer usually seeks to gain a general understanding of the entire behaviour, not just of individual slices. For this purpose, the explorer needs to merge his/her observations extracted from multiple partial views into a coherent mental image and, when needed, an explicit overall pattern. This task implies the grouping of the slices of the behaviour by similarity, taking into account the relations (ordering and distances) between the referrer values that they correspond to. For example, when the behaviour slices correspond to different time moments, the explorer groups similar slices referring to consecutive moments together, rather than slices referring to arbitrary moments. When the slices refer to different locations in space, the explorer tries to identify spatial clusters of similar local behaviours.

Comparison and grouping of behaviour slices may be supported by an appropriate arrangement of partial views. It is convenient when all these views can be present simultaneously on the screen – this greatly supports comparisons and the grasping of the general tendencies in the changes from slice to slice. The arrangements that support simultaneous visibility

of different slices are space partitioning (display juxtaposition, for example a sequence of maps for different time moments, as in Fig. 3.16), space embedding (for example a map with embedded time graphs, as in Fig. 3.13), and space sharing (for example overlaid time graphs in a common coordinate framework, as in Fig. 4.3). Simultaneous visibility of behaviour slices, however, is not always possible, since the partial views may be too numerous.

Another way of arranging partial views is to use the display time, where a view of one slice is replaced by a view of another slice. This may be done in an animation-like mode, which needs to be complemented by flexible controls that allow the user to choose any slice. Since the use of the display time is not an ideal solution from the perspective of supporting comparisons, it may be appropriate to combine it with display juxtaposition. For example, one may have a dynamic display where slices replace each other and, simultaneously, a static display showing a specific selected slice, which can thus be compared with all the other slices.

The outcome from exploring slices and aggregates of the data is a number of aspectual patterns, which need to be brought together, in accord with the principle “establish linkages”. If some initial general pattern has been derived at the stage of joint consideration of all the referrers, the compound pattern resulting from the study of the aspectual behaviours is attached to it as an extension, which adds relevant details and increases the precision of the overall pattern.

In Sect. 5.3 and Sect. 5.4.8, we have presented an approach to establishing linkages between aspectual patterns, which is based on cross-partitioning, i.e. transferring partitions made in the course of the exploration of one aspectual behaviour to the view(s) of the other aspectual behaviour(s). For this purpose, display coordination is the primary instrument. Besides responding to reference set division through multicolour marking or display multiplication, it may also be useful to allow the user to make sketches on top of displays and transfer these drawings to other displays, as is shown in Figs 5.4C and 5.5C.

Table 5.10 deals with the exploration of a single aspectual behaviour. Table 5.11 refers to the procedure specified in Table 5.10 as a sort of subroutine and describes the entire process of analysing a behaviour based on a multidimensional reference set by means of considering the aspectual behaviours.

**Table 5.10.** Characterisation of an aspectual behaviour in a behaviour based on a multidimensional reference set

Actions	Subtasks	Tools	Principles
<p>Overview the set of behaviour slices corresponding to different values of one of the referrers. Is the perception of the character of the behaviour obstructed by outliers or fluctuations?</p> <p>(A) yes; (B) no</p>		Visualisation: single display; spatial and/or temporal arrangement of multiple displays	See the whole
<p>(A.1) Simplify the view; in the case of multiple displays – consistently simplify all of them</p>	<p>Outlier removal Smoothing Aggregation Reordering; in the case of multiple displays, re-arrangement of the displays</p>	Simplify and abstract	
<p>(A.2), (B.1) Grasp the character of the behaviour and its properties in each slice, and their variation between the slices. Group the slices by similarity, taking into account the relations (ordering and distances) between the referrer values that they correspond to</p>	<p>Behaviour characterisation (for each slice) Pattern comparison (between the slices) Synoptic relation-seeking: detect significant pattern changes between the slices</p>	<p>Visualisation: single display, spatial and/or temporal arrangement of multiple displays Data standardisation<sup>a</sup> Display coordination Display grouping Aggregation (similar slices) Clustering (similar slices) Computing changes Overlaying</p>	See in relation Divide and group Zoom and focus (when looking at individual slices)
<p>(A.3) Characterise the deviations from the general pattern, both “elementary outliers” (individual values) and “behavioural outliers” (slices with uncommon behaviours)</p>	<p>Elementary lookup and comparison Pattern comparison</p>	<p>Computation (of residuals) Querying Display arrangement: juxtaposition, overlaying</p>	Attend to particulars See in relation

<sup>a</sup> For a better comparability; for example, transformation to  $z$ -scores.

**Table 5.11.** Characterisation of a behaviour based on a multidimensional reference set by means of exploring the aspectual behaviours (subcase 2.2)

Actions	Subtasks	Tools	Principles
For each referrer, explore the corresponding aspectual behaviour as specified in <i>Table 5.10</i>	Behaviour characterisation; see <i>Table 5.10</i>	See <i>Table 5.10</i>	See the whole
For selected referrers, whenever appropriate, explore the variation of the characteristics aggregated over the other referrers: depending on the dimensionality, apply the procedure for <i>case 1</i> or <i>case 2</i>	Behaviour characterisation See also <i>case 1</i> and <i>case 2</i>	Aggregation Querying See also <i>case 1</i> and <i>case 2</i>	See the whole
Join the aspectual patterns and the aggregated patterns (if any) into a unified overall pattern	Connection discovery Pattern comparison	Visualisation Display coordination	Establish linkages Establish structure See in relation

### 5.5.3 Case 3: Multiple Attributes

The presence of multiple attributes in a dataset does not necessarily exclude a holistic perception of their joint behaviour. In the part of Sect. 5.4.1 dealing with unification, we have given several examples of the visualisation of multiple attributes in a single display; see Figs 5.12C–5.14. In these examples, values of multiple attributes have been encoded in various visual properties of display elements or in components of structured marks (diagrams). A slightly different example can be seen in Fig. 5.19C, where several layers overlaid in a single display represent the behaviours of different attributes. Such displays are often quite difficult to interpret, but, after some training, an explorer can grasp the general character and essential features of the joint behaviour from them.

If a holistic visualisation of the joint behaviour of several attributes is possible, it can be analysed as described in the previous subsections dealing with case 1 and case 2. However, this possibility is quite limited; it can work only with a fairly small number of attributes. Moreover, increasing the dimensionality of the data (i.e. the number of referrers) reduces the possibilities for joint visualisation of several attributes.

In Sect. 5.4.8, we have discussed two approaches to the exploration of the joint behaviour of multiple attributes. The first approach is based on

visual or computational integration of the attribute values associated with each reference. The second approach involves a separate consideration of individual attributes or groups of attributes and establishing linkages between the patterns thus derived. These approaches were called Decomposition A and Decomposition B, respectively.

The example displays in Figs 5.12C–5.14 and 5.19C demonstrate the possibilities for visual integration of attribute values. Computational integration means producing a single attribute from several original attributes, as is described in Sect. 4.5.2. It has been noted that such a method for attribute integration is usually quite domain-specific, and its applicability is therefore rather limited. Besides, it involves tremendous information loss.

There is yet another method of integration, specifically, classification or clustering of the references according to the corresponding values of the multiple attributes. Typically, the explorer seeks to obtain a division of the reference set into subsets such that the corresponding parts of the overall behaviour can be regarded as (sufficiently) homogeneous. The process of classification or clustering may involve many trials until an appropriate result is obtained. The final result may be viewed as a new qualitative attribute, the values of which denote the classes or clusters.

With any sort of attribute integration, the subsequent analysis can be subsumed under either case 1 or case 2 discussed earlier, depending on the dimensionality of the reference set. With the separate analysis of individual attributes or groups (followed by pattern linking), each individual attribute is analysed as in case 1 or 2. Consideration of an attribute group means that the members of the group undergo visual or computational integration, as described above, and are then analysed as in case 1 or 2.

Establishing linkages between the behaviours of individual attributes or groups involves a comparison of these behaviours to detect similarities and corresponding features. For this purpose, the behaviours can be visualised in separate displays, or the visualisations may be combined (overlaid) in a single display. Sometimes, it is possible to transform the data so as to make the behaviours more comparable; the views of the transformed data can be manipulated through a common display manipulation tool. The approaches to behaviour comparison and the supporting tools for this are discussed in Sect. 5.4.4, which deals with the principle “see in relation”.

Besides behaviour comparison, establishing linkages may involve searching for correlations or typical associations of values of the attributes which correspond to either the same reference or neighbouring references. The appropriate tools, which are overviewed in Sect. 5.4.8, include various computational methods for correlation analysis, special visualisation techniques such as scatterplots and scatterplot matrices, display linking through selection or division, and overlaying of several visualisations.

It is not supposed that an explorer must choose a single approach, i.e. either Decomposition A or Decomposition B, and never try to go another way. For a comprehensive study, it is appropriate to explore both the joint behaviour of multiple attributes and the individual behaviours of the attributes.

**Table 5.12.** Characterisation of a joint behaviour of multiple attributes through value integration

<b>Actions</b>	<b>Subtasks</b>	<b>Tools</b>	<b>Principles</b>
Option A: visualise the joint behaviour in a single display in a way that promotes unification. Apply <i>case 1</i> or <i>case 2</i> , depending on the number of referrers.	Behaviour characterisation See also <i>case 1</i> and <i>case 2</i>	Visualisation	See the whole
Option B: integrate the attributes into a single attribute. Apply <i>case 1</i> or <i>case 2</i> to the resulting attribute.	Behaviour characterisation See also <i>case 1</i> and <i>case 2</i>	Data transformation (attribute integration) Visualisation, in particular for testing sensitivity to integration parameters	Simplify and abstract
Option C(1): by means of classification or clustering, divide the reference set into subsets according to the similarity of characteristics	Behaviour characterisation, pattern search, pattern comparison: estimate the goodness of division and/or interpret the results of clustering Elementary lookup and comparison: interpret the results of clustering	Classification Clustering Visualisation Display linking, querying, computing summary statistics to interpret the results of clustering	Divide and group See in relation Look for recognisable
Option C(2): regard the resulting classes or clusters as a new attribute with qualitative values. Apply <i>case 1</i> or <i>case 2</i> to explore its behaviour.	Behaviour characterisation See also <i>case 1</i> and <i>case 2</i>	Visualisation See also <i>case 1</i> and <i>case 2</i>	Simplify and abstract

The case of the analysis of multiple attributes is summarised in Tables 5.12 and 5.13. As in the previous subsection, the first of these tables specifies a kind of subprocedure, which is then referred to from the second table. Specifically, Table 5.12 deals with the attribute integration approach, which may be combined with Decomposition A, and Table 5.13 deals with the approach of separation and linking, i.e. Decomposition B.

**Table 5.13.** Characterisation of a behaviour consisting of values of multiple attributes (case 3)

<b>Actions</b>	<b>Subtasks</b>	<b>Tools</b>	<b>Principles</b>
Explore the behaviours of the individual attributes by applying <i>case 1</i> or <i>case 2</i>	Behaviour characterisation (individual behaviours)	See <i>case 1</i> and <i>case 2</i>	
Compare the behaviours of the individual attributes	Behaviour/pattern comparison	Visualisation Display arrangement: juxtaposition, overlay Data standardisation <sup>a</sup> Joint display manipulation Display linking	See in relation
If appropriate, form attribute groups for a joint study according to domain knowledge or similarity of behaviours. Characterise the joint behaviour of each group as specified in <i>Table 5.12</i>	Behaviour characterisation (groups of attributes)	See <i>Table 5.12</i>	See the whole
Establish linkages between individual attributes and attribute groups	Connection discovery	Computational methods; in particular, correlation analysis Specific visualisations, e.g. scatterplots Linked displays Overlaid visualisations	Establish linkages

<sup>a</sup> For better comparability; for example, transformation to *z*-scores.

### 5.5.4 Case 4: Large Data Volume

By “large data volume” we mean a large number of references, i.e. elements of the reference set, for which characteristics are specified in terms of attribute values. Leaving aside the technical problems of computer performance and memory capacity, the main problem that such data pose to an explorer is the impossibility of obtaining a clear view of the overall behaviour by applying the usual methods of visualisation. Visual displays of large datasets suffer from cluttering of marks and overlap of marks, or the marks have to be so small that one hardly sees them, or the display size is simply too small for all the data items to be fitted in. The functionality of all tools that suppose user interaction with the data display, such as dynamic querying and display manipulation, also deteriorates greatly.

There are two basic approaches to handling large amounts of data: selection and aggregation. The former approach means that the explorer selects subsets of the data and analyses those subsets. This approach conflicts with the principle “see the whole”: it is very difficult to gain a coherent picture of the overall behaviour in this way, or, formally speaking, to derive a unified pattern approximating the overall behaviour. The latter approach means that the original references are united into groups, these groups are treated as new references, and hence the number of different references decreases substantially and becomes manageable. The characteristics corresponding to the new references are derived from the original characteristics by means of statistical summarisation over the groups. This approach involves significant information loss and conflicts with the principle “attend to particulars”, since there is a risk of missing important deviations from what is standard and usual.

A feasible solution of the problem may lie in combining these two approaches. Some examples can be seen in Figs 5.25–5.28, where selected subsets of individual data items are represented together with aggregated data as additional layers superimposed upon the visualisation of the aggregated data. In order to detect particulars requiring the explorer’s attention, it may be recommended that one examines the statistical distribution of the attribute values within the aggregates, for example by using positional statistical measures. It is also possible to explore the general features of the overall behaviour by means of aggregation and then to apply zooming, focusing, and filtering in order to look at different subsets of the original data and detect and inspect various particular features that may occur in those subsets.

The main purpose of using aggregation is to reduce the size of the reference set of the data that it is applied to. However, aggregation may result in a much more serious transformation of the reference set than just size

reduction. Thus, we have discussed the fact that aggregation may be used as a tool for reducing the dimensionality of data (see Sect. 5.4.1): the data are aggregated through bringing together all values of a referrer, which excludes this referrer from further analysis. Aggregation may also increase the dimensionality of the data and/or replace the original referential components by completely new ones. In fact, when data are aggregated according to the values of an attribute, this attribute becomes a new referrer, while the original referrer may be ignored.

This occurs, for example, in aggregating earthquake data by territorial compartments and time intervals. Originally, the dataset has a single referrer of population type, specifically, the set of earthquakes. Among the attributes, there are the place and the time of the occurrence of the earthquake. To do the aggregation, one takes the value domains of these two attributes and divides them into subsets: equal-size rectangles for the spatial attribute and equal-length intervals for the temporal attribute. Then, for each rectangle–interval combination, the number of earthquakes is counted, and summary statistics of the corresponding values of attributes such as the magnitude or depth are computed.

The resulting dataset has two referrers, spatial and temporal, with value sets consisting of rectangular territorial compartments and regular time intervals, respectively. The values of the attributes refer to these compartments and intervals. The original referrer has been omitted from further consideration. Hence, as a result of such a transformation, one obtains, strictly speaking, a new dataset with its own behaviour, which is not the same as the behaviour of the original data.

Is this a valid substitution? We would prefer to say “no”, unless the transformation agrees with the goals of the exploration. In this particular case, the explorer may be specifically interested in investigating the spatial and temporal distribution of the number and characteristics of the earthquakes. Hence, transforming the original dataset into a dataset with space and time as referrers is convenient for the explorer and adequate for achieving the goal. At the same time, the transformed data are unsuitable for the task of detecting spatio-temporal clusters of earthquakes described in Sect. 5.4.5. That task has to be performed by means of selection.

Generally, aggregation is done by means of introducing equivalence classes of values of one or more data components, i.e. disregarding differences between some values and treating them as being the same. Unless the aggregation is done over the entire value set of a component, there has to be some basis for considering different values as equivalent. Typically, this basis is sufficient closeness of the values, which means that the value domain that they belong to must have *distances* (or the values may be semantically close; for example, both pine and spruce are coniferous trees).

A component used as a basis for data aggregation may be either a referrer or an attribute. If a referrer of a dataset has distances between values, it is generally preferable to use it as a basis for data aggregation, since this does not radically change the structure of the data, and the behaviour of the transformed data may be viewed as a “coarsened” version of the original behaviour rather than as a different behaviour. An example of this kind of aggregation may be seen in Figs 4.86C and 4.87C, in which we aggregated spatially referenced data by dividing the territory into regular compartments and averaging the values within the compartments. In the result, we obtained spatially referenced data again, i.e. the structure of the structure did not change.

In the case of the earthquakes, the reference set of the original data is the set of individual earthquakes, i.e. a population in the statistical sense, with no distances between members defined. Hence, the data cannot be aggregated on the basis of the referrer but only on the basis of one or more appropriate attributes, i.e. attributes with distances between their values. The date of occurrence of the earthquake and the location of the epicentre are, in this respect, suitable attributes. However, other attributes with distances, such as the magnitude, the depth, or the time of day, could, in principle, be chosen as well to be the basis for aggregation.

The goals of the analysis and/or the explorer’s domain knowledge may dispose him/her to choose particular attributes or attribute combinations. Thus, a typical preference in the exploration of a set of events is to aggregate them on the basis of their spatial and/or temporal attributes. Another criterion for choosing attributes to be the basis for aggregation is the statistical distribution of the attribute values. Thus, if some attribute values occur extremely frequently while others occur very rarely, such an attribute is hardly suitable as a basis for aggregation. Thus, in the earthquake dataset, 3588 of the 10 560 earthquakes (34%) have magnitudes in the range from 2.9 to 3.0 (more precisely, 1878 earthquakes have a magnitude of 2.9 and 1710 earthquakes have a magnitude of 3.0) and there are only 54 earthquakes with magnitudes in the range from 5.0 to 7.3, which is the maximum value in the dataset. Hence, the earthquakes cannot be divided on the basis of this attribute into groups of comparable size.

In cases where there are no well-grounded preferences, it is necessary to consider aggregation on the basis of all appropriate attributes and investigate how the resulting behaviours are related to each other. Such an investigation may be supported by linked displays, for example histograms that support selection of a subset of the data through clicking on bars and represent the selection by means of marking (see Fig. 4.99).

We would like to stress once again the importance of reaggregation (we did this in the Sect. 4.5.4 dealing with data aggregation), i.e. redefining the

value equivalence classes in different ways and checking the validity of the patterns perceived earlier.

The procedure for analysing a dataset with a large number of references is presented in Table 5.14. The major approach is data aggregation. Table 5.14 refers to cases 1, 2, and 3, which have been described earlier, as sub-procedures to be applied for the analysis of the results of data aggregation.

**Table 5.14.** Characterisation of the overall behaviour in the case of a very large reference set (case 4)

Actions	Subtasks	Tools	Principles
Can the data be aggregated on the basis of the referrer(s)? (A) yes; (B) no	(A.1) Aggregate the data by uniting close references and averaging their characteristics	Aggregation	See the whole Simplify and abstract
(B) Are there well-grounded preferences for choosing particular attributes as the basis for aggregation? (C) yes; (D) no	(C.1) Aggregate the data on the basis of the chosen attribute(s)	Aggregation	See the whole Simplify and abstract
(A.2), (C.2) Characterise the overall behaviour of the aggregated data by applying <i>case 1</i> , <i>case 2</i> , or <i>case 3</i>	Behaviour characterisation (aggregated data)	See <i>case 1</i> , <i>case 2</i> , and <i>case 3</i>	See the whole
(A.3), (C.3) Re-aggregate the data and check the validity of the pattern derived	Behaviour characterisation Pattern comparison	Visualisation Display arrangement: juxtaposition, overlay Joint display manipulation Display linking	See in relation
(D.1) Perform multiple aggregations of the data according to the	Aggregation	See the whole Simplify	

<b>Actions</b>	<b>Subtasks</b>	<b>Tools</b>	<b>Principles</b>
suitable attributes			and abstract
(D.2) Characterise the overall behaviour with respect to each attribute by applying <i>case 1</i>	Behaviour characterisation (aggregated data)	See <i>case 1</i>	See the whole
(D.3) Reaggregate the data and check the validity of the patterns derived	Behaviour characterisation Pattern comparison	Visualisation Display arrangement: juxtaposition, overlay Joint display manipulation Display linking	See in relation
(D.4) Compare the behaviours with respect to different attributes	Behaviour/ pattern comparison	Visualisation Display arrangement: juxtaposition, overlay Joint display manipulation Display linking	See in relation
(D.5) Establish linkages between the attributes	Connection discovery	Linked displays Computational methods; in particular, correlation analysis	Establish linkages
Detect outliers by considering the statistical distribution of attribute values within the aggregates or over the whole dataset. Examine the outliers found	Elementary lookup and comparison	Computation (elementary statistics; in particular, positional measures) Visualisation (distribution displays such as histograms or box-and-whiskers plot) Querying Combined visualisation of aggregated data and selected individual data items	Attend to particulars See in relation
Select various subsets of the original data and compare the individual data items with the aggregated characteristics. Detect and examine atypical values and value combinations, in	Elementary lookup and comparison	Querying Zooming and focusing Combined visualisation of aggregated data and selected individual data items Display coordination	Zoom and focus Attend to particulars See in relation

Actions	Subtasks	Tools	Principles
<p>particular, local outliers</p> <p>If the behaviour of the original data may have specific features of interest that could have been erased by the aggregation, explore the original behaviour by scanning data subsets</p>	<p>Pattern search Elementary lookup and comparison</p>	<p>Visualisation Zooming and focusing Querying Display coordination Specialised computational tools for pattern detection</p>	<p>Zoom and focus Look for recognisable Attend to particulars</p>

### 5.5.5 Final Remarks

This subsection has been conceived as a summary of what has been said earlier concerning tasks, tools, principles, and the relationships between these. We have tried to describe how to accomplish the primary task of exploratory data analysis, that is, characterisation of the overall behaviour of the phenomenon underlying a dataset. In the course of the analysis, this task is decomposed into subtasks and supporting actions (e.g. data transformation), which are performed by means of certain tools. The actions, the subtasks, and, hence, the tools differ depending on the structure and peculiarities of the data under analysis. These differences are reflected in Tables 5.6–5.14, which deal with four general cases:

1. Data with a single referrer and a single attribute or several jointly explored attributes that may be visualised holistically together.
2. Data with multiple referrers.
3. Data with multiple attributes.
4. Data with a large reference set.

For these cases, the tables show how the primary task is decomposed, and they relate the actions and subtasks involved to the appropriate tool categories and the appropriate general principles of data analysis and tool selection (the principle “involve domain knowledge” is not explicitly mentioned in any of the tables, but it is relevant virtually everywhere).

By looking through the tables, it can be noted that almost every table refers to the cases considered earlier as subprocedures to be applied after a data transformation, partitioning, or subset selection. Hence, if the dataset to be explored contains both multiple referrers and multiple attributes and, in addition, has a very large reference set, the whole procedure for analys-

ing this dataset is reflected in the entire collection of tables, which need to be viewed in a backward order, starting from the last one.

The tables contained in this section have been meant to establish explicit links between the possible exploratory tasks and the tools capable of supporting these tasks. We wrote at the beginning of this chapter that direct linking between task types and appropriate tools is hardly meaningful because of the very high level of generality of the task typology. Such linking cannot provide useful guidance for an explorer, whose tasks (questions) are very specific, in the sense of being formulated in terms of particular data components.

Therefore, although we have indicated what tool categories may be appropriate for various types of tasks, we give the primary role in tool selection to the general principles that have been formulated in this chapter. Accordingly, the tables also link the tasks to these principles. The general idea is that an explorer can find what principle(s) are relevant to the task category that his/her specific task belongs to and then apply these principle(s) to choose suitable analysis tools, taking into account the structure and properties of the actual dataset. For example, an explorer may learn from a table that the relevant principle for a certain kind of task is “see the whole”. This principle means that all referential and characteristic components need to be represented visually according to certain requirements, which are defined in Sect. 5.4.1, which deals with this principle. Knowing the structure and properties of the dataset at hand, the explorer can translate the general requirements into more specific ones, which can help to define the appropriate visualisation technique.

Moreover, the tables and the entire content of this section have been organised so that the explorer is not required to determine what general task category his/her particular question belongs to. The explorer does not even need to have any explicit question (in Sect. 3.8, we have mentioned that an explorer may be unaware that his/her actions in the course of data analysis are actually aimed at finding answers to certain questions) and, consequently, does not need to look through all the tables to try to find the applicable row containing the guiding information. The organisation of the material supposes a different way of using the tables.

It was essential that we did not just list the task types in an arbitrary order or arrange them according to their formal properties; instead, we had to define their places in the common context of exploratory data analysis, i.e. organise the tasks into a hierarchical system where less general tasks appear as subtasks of more general ones. Besides tasks in the sense of questions that need to be answered, we have also mentioned various supporting actions, which do not involve seeking answers to questions but prepare the data for further analysis.

With this organisation, the intended use of this section may be described as follows. The explorer determines what referential and characteristic components exist in the data that he/she needs to analyse, and finds out which of the four cases is applicable to the data. Then, the explorer looks at the corresponding table, which simultaneously suggests an appropriate procedure for data exploration, refers to the relevant general principles, and mentions the categories in which to look for suitable tools. Hence, it is not required that the explorer explicitly considers any tasks, either specific or general.

The various task types mentioned in the columns entitled “Subtasks” are not intended in fact for explorers and are not intended to play any role in the process of tool selection and data analysis. Instead, these links to our task typology may be helpful for tool designers and developers, who are also regarded as potential users of the results of our study. Unlike explorers, tool designers do not seek answers to exploratory questions concerning data. However, they do need to anticipate the questions that may arise so that the tools they design can really support finding the answers. Correspondingly, the references to the task types are meant to guide tool designers in identifying the questions that potential tool users may have. The corresponding principles can help in defining the essential requirements to the tools. Designers do not necessarily need to look in the column listing the relevant tool categories. They can, in principle, try to invent something completely new, not fitting into the current tool classification. On the other hand, they could use the existing approaches as a basis, and modify the existing techniques or use them as building blocks for new tools.

We have thus summarised our dual-use theory and indicated how it can be applied in the practice of exploratory data analysis and that of designing tools for EDA. In the next section, we demonstrate the first type of use with an example. In the concluding part of the book we present some ideas intended to promote the application of our theory.

## 5.6 Applying the Scheme (an Example)

Let us now briefly demonstrate how the suggested scheme can be applied to a particular dataset. As an example, we shall take the data mentioned earlier about earthquake occurrences. Recall that the dataset consists of 10 560 records of earthquakes that occurred in western Turkey and the surrounding area during the period from 1 January 1976 to 30 December 1999. The reference set here is the set of all earthquakes, which may be described as a statistical population, i.e. a set without ordering or distances.

It may be objected that the earthquakes are ordered according to the time of their occurrence and that distances exist between their epicentres in geographical space. However, this ordering and these distances are not the properties of the set of earthquakes as such but are “borrowed” from the attributes of the earthquakes, specifically, the date of occurrence and the epicentre location. Besides these two attributes, the dataset specifies the magnitudes of the earthquakes, their depths, and the times of day when they occurred.

The major complexity of this dataset is the large size of the reference set; hence, of the cases described above, case 4 applies to it. According to the recommended procedure, we need to aggregate the data. As we have already mentioned, the referrer of this dataset is not suitable as a basis for aggregation: the absence of distances between the values does not provide a reasonable basis for introducing equivalence classes of values. Consequently, the aggregation has to be done on the basis of some attribute(s). We have noted that the date of earthquake occurrence and the location of the epicentre are suitable candidates. This corresponds to our major interest concerning the data: we would like to know, first of all, how the earthquakes are distributed in space and time. Of the other attributes, the magnitude and the depth are completely unsuitable as bases for aggregation because of the peculiarities of their statistical distributions: there are extremely many occurrences of small values and quite few cases of high values. The time of day has a more even distribution of values; so, we may try this as well. However, let us consider aggregation on the basis of the spatial and temporal attributes first.

For the aggregation, we divide the value domain of the spatial attribute, i.e. the territory of western Turkey and its neighbourhood, into regular spatial compartments, namely rectangular cells, as has been shown in Figs 4.81–4.83 and 4.85C. Simultaneously, we divide the value domain of the temporal attribute, i.e. the time period from 1 January 1976 to 30 December 1999, into regular intervals. The length of an interval is chosen to be one year. Then, we apply the available software to count, for each combination of a spatial compartment and a time interval, the number of earthquake occurrences in this compartment during this interval. The software also computes summary statistics of the magnitudes and depths of the earthquakes aggregated in this way, specifically the minimum, maximum, and median.

In the result of the aggregation, we obtain a dataset with two referrers: a spatial referrer, the values of which are the spatial compartments, and a temporal referrer, the values of which are the time intervals. The attributes are the earthquake count, and the minimum, maximum, and median magnitude and depth.

This is a case of multiple attributes, i.e. case 3. We do not know a good way of simultaneous visualisation of all the attributes so that an effect of unification can be achieved. We also do not know a reasonable method for integrating all the attributes into a single attribute. Hence, according to Table 5.13, we need to explore the behaviour of each attribute individually.

Let us start with the earthquake count. The behaviour of this attribute is based on a two-component reference set, where one of the referrers is time and the other is geographical space, which is, in turn, two-dimensional. Hence, three display dimensions are necessary for the representation of the reference set, i.e. we need one more dimension in addition to the usual two spatial dimensions of a display. A simulation of the third spatial dimension, as in the space-time cube in Fig. 5.22, cannot help in this case, since the cube would have to be filled with marks representing the attribute values, and very many marks would be hidden behind marks positioned in front of them. Consequently, we need to involve either the display time or an arrangement in order to represent all the dimensions of the reference set. As we have discussed in Sect. 5.4.1, neither of these solutions supports the perception of the resulting display as an integral space, in a single glance. Hence, this is not a case where a holistic view is possible (case 1), but rather a case where the behaviour of an attribute splits into aspectual behaviours, i.e. case 2.

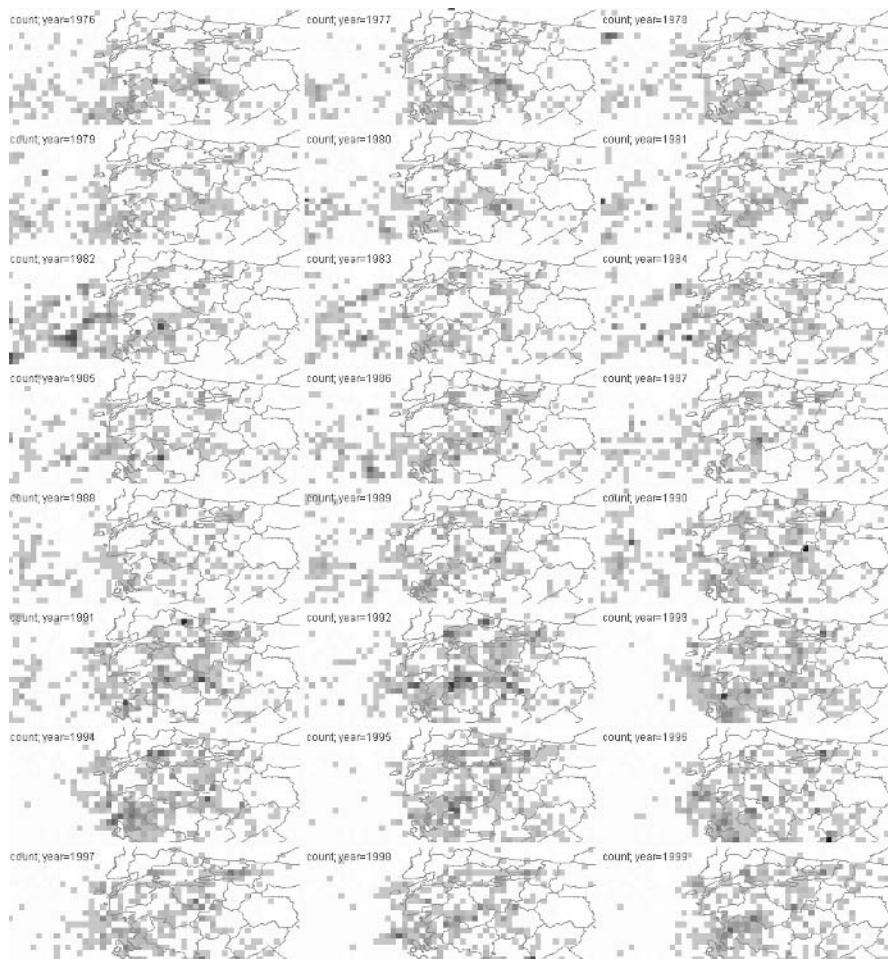
According to Table 5.11, we need to explore each aspectual behaviour. In this particular case, we have two aspectual behaviours:

- the temporal variation of the distribution of the earthquake frequency over the territory;
- the spatial distribution of the local temporal behaviours, i.e. the variations of the earthquake frequency at different spatial locations.

This is quite analogous to the case of the behaviour of the burglary rate over the territory of the USA and the time period from 1960 to 2000, which we considered earlier.

As in the case of the burglary rate data, the first aspectual behaviour can be explored by means of a series of maps, where each map represents one slice of the overall behaviour corresponding to one fixed value of the temporal behaviour, in our case corresponding to one year. Such a map series is shown in Fig. 5.43. Since there are 24 maps corresponding to the 24 years from 1976 to 1999, each of the maps has to be quite small. Of course, the maps are not suitable for retrieving any detailed information, but are intended mostly to give an overall impression of the spatial distributions of the earthquake frequency in the respective years. The frequencies are represented by shading the rectangles corresponding to the space

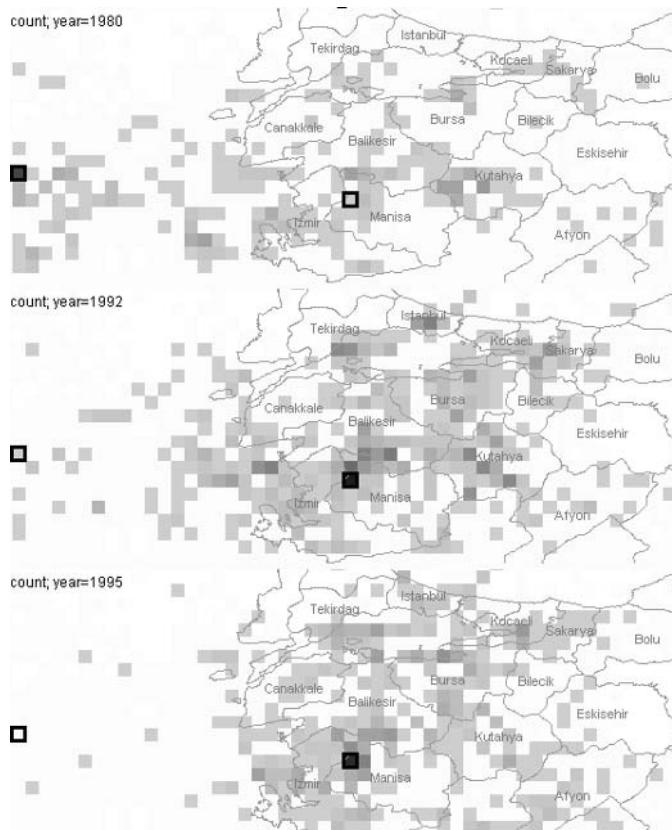
compartments with varying degrees of darkness: the higher the frequency, the darker the colour.



**Fig. 5.43.** The “small multiples” here represent one of the aspectual behaviours of the earthquake frequency, specifically, the temporal variation of the spatial distribution

Actually, what can be seen in Fig. 5.43 is not the original appearance of the “small multiples” display. The original view was insufficiently expressive owing to a few outliers, which were represented by the darkest shades while the remaining rectangles were light. Therefore, according to the recommendation given in Table 5.10, we have simplified the view by removing the outliers. Specifically, we have applied a focusing tool and limited

the range of values to be represented by shading to values from 0 to 15. This manipulation has been consistently applied to all 24 maps. So, the highest value represented by shading in Fig. 5.43 is 15 earthquake occurrences per compartment. By means of focusing, we have removed three outliers: 18, 23, and 21 earthquakes per compartment, which occurred in the years 1980, 1992, and 1995, respectively. Figure 5.44 shows the positions of the outliers on maps corresponding to these years. The compartments where the high values were attained are marked by thick black boundaries. In 1980, a frequency of 18 earthquakes was attained at the western edge of the territory under study. In the years 1992 and 1995, frequencies of 23 and 21 were attained in one and the same compartment in the centre of the territory, on the border between the districts of Izmir and Manisa.



**Fig. 5.44.** The outliers that have been removed from the previous picture: 18 in 1980 (on the western edge), 23 in 1992 (in the district of Manisa), and 21 in 1995 (in the same place)

Let us now attend to the “small multiples” in Fig. 5.43 and try to grasp the character and major properties of each behaviour slice and the variation of these properties between the slices, as Table 5.10 recommends. We have to admit that what we see does not immediately produce an impression of a consistent development of a prominent spatial pattern. However, certain general observations can be made. Thus, we can note that in the period from 1976 to 1990 there were many more earthquakes in the west outside the territory of Turkey than later, especially after 1992. The frequency of the earthquakes in the territory of Turkey, in contrast, increased after 1990. We can see that the part of the Turkish territory affected by earthquakes increased in size as compared with the beginning of the period under investigation. The “shakiest” area is in the south-west of Turkey, in the district of Izmir and its neighbourhood.

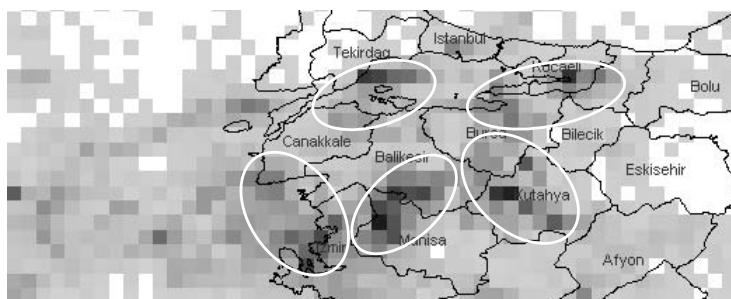
The slice corresponding to the year 1982 looks like a “behavioural outlier” in the sequence of slices for the years from 1976 to 1989, which are quite similar to each other. From 1990 to 1992, the character of the spatial behaviour changes: the earthquake-affected area moves to the east and north-east. The patterns that can be perceived starting from the year 1993 appear as a result of this movement.

In order to obtain a kind of general view of the character of the behaviour during each of the three periods that we detected (i.e. 1976–1989 with the exception of 1982, 1990–1992, and 1993–1999), we have summed the frequencies over these periods and visualised the sums thus obtained in three maps. For better comparability, we have transformed the computed values into  $z$ -scores, i.e., roughly speaking, deviations from the mean frequency for the respective periods (the mean frequency has been computed individually for each period). The result can be seen in Fig. 5.45C. Shades of green correspond to frequencies lower than the means, and shades of red to frequencies higher than the means. The difference between the characters of the behaviour in the first and the third period is clearly seen. The behaviour in the second period looks intermediate between the former and the latter.

So, we have managed in a sense to grasp the character and properties of each behaviour slice and the variation of the character and properties over time, as is recommended in Table 5.10. We have also mentioned the deviations from the general pattern: first of all, the three outliers that were removed from the initial view (see Fig. 5.44), and second, the “behavioural outlier” – the spatial distribution in the year 1982. According to Table 5.10, we need to characterise these deviations; however, we shall not do this now, hoping instead that the book contains a sufficient number of examples of how to “attend to particulars” and “see in relation”. As a re-

minder, we shall note only that the elementary outliers can be investigated by means of querying tools, and the behavioural outlier by visualising the uncommon behaviour and comparing it with the typical behaviours. In Fig. 5.46C, the spatial behaviour of the earthquake frequency in 1982 is visualised in the same manner as for the three summarised behaviours in Fig. 5.45C. We have transformed the original values into  $z$ -scores for a more convenient comparison of the untypical behaviour with the typical behaviours.

According to Table 5.11, we may try to aggregate the data over the temporal referrer and look at the spatial distribution of the summary earthquake frequencies over the entire period. In fact, we did such an aggregation earlier to produce Fig. 4.81; however, for a more convenient comparison, we reproduce the result of this aggregation once again in Fig. 5.47, which shows the same territory as in the “small multiples” in Fig. 5.43 (for better visibility, we have slightly zoomed into the part of the territory mostly affected by earthquakes).



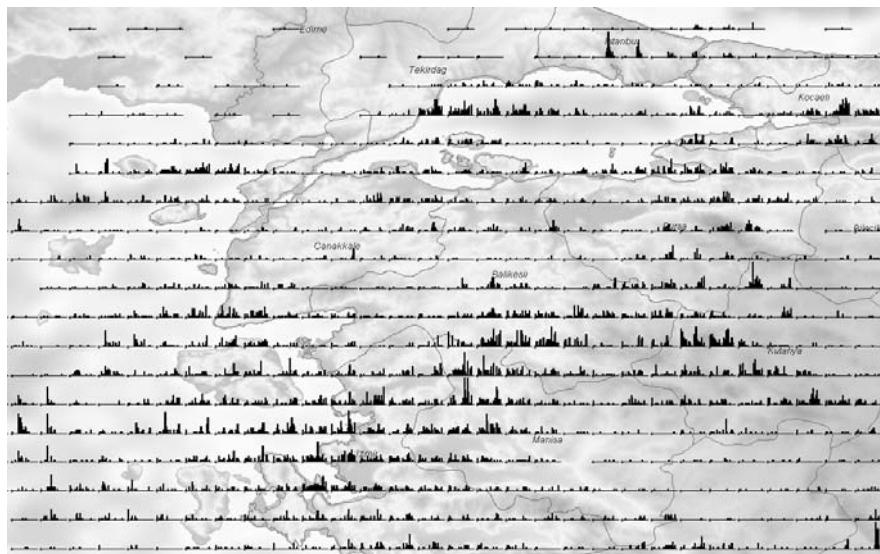
**Fig. 5.47.** The total earthquake number aggregated over all 24 years; the maximum count is 90

As is stated in Table 30, the behaviour of the aggregated data should be characterised by applying case 1 or case 2. Since we now have a single referrer (specifically, a spatial referrer), case 1 is applicable. The behaviour may be described as heterogeneous (subcase 1.2) and characterised through dividing the reference set into subsets, as is suggested in Table 5.7.

To characterise the behaviour represented in Fig. 5.47, we can divide the reference set, i.e. the territory under investigation, into a “background” with relatively low earthquake frequencies and a number of “spots” of high and very high frequencies. The largest “spots” are roughly outlined in Fig. 5.47. The “background” can be characterised, according to Table 5.6, as having relatively invariant characteristics. We need to apply appropriate querying and statistical tools to specify the range of frequencies in the

“background”, the mean and median frequencies, the variance, etc. Before doing that, we used the manual classification tool mentioned in Sect. 5.4.3 to separate the “spots” from the “background”.

For each of the “spots”, we need to specify its location and extent, and characterise its internal behaviour by a suitable partial pattern. By specifying the location and extent, we define the applicability domain of the partial pattern, as is required in Table 5.7. We shall not now characterise every “spot”, but shall give a rough example of how this may be done. Thus, the spot marked in the lower left can be described as extending from the south-west of the district of Canakkale to the north-west of the district of Izmir, including a coastal part of the Turkey and the nearby part of the sea. The behaviour inside this area can be characterised as a behaviour of increasing frequency in the directions from the centre (which is in the sea) towards the north-west and the south-east. The rate of increase in the latter direction is higher than in the former. Hence, we have characterised the partial behaviour in the Canakkale–Izmir spot as a more or less regular change. By means of querying, we can obtain various numeric characteristics of the change, as is recommended in Table 5.6. The behaviour in the nearby “spot”, which could be called Balikesir–Manisa spot, could be described as one of decreasing frequency in the direction from the central “core” to the periphery.



**Fig. 5.48.** The local behaviours of the earthquake frequency (after removing the outliers  $>15$ )

After describing the entire aggregated behaviour, we must attend to the second aspectual behaviour, i.e. the spatial distribution of the local temporal behaviours of the earthquake frequency. This aspectual behaviour can be visualised as is shown in Fig. 5.48. The local behaviours in the spatial compartments are represented in a map by bar charts with bars corresponding to the years from 1976 to 1999. As previously, in Fig. 5.43, the outliers have been removed from the representation by focusing on the value range from 0 to 15. According to Table 5.10, we need to grasp the character and properties of each behaviour slice (i.e. each local behaviour in this case) and their variation over the territory, and to group the local behaviours by similarity, taking spatial distances into account. It should be noted, however, that the visualisation in Fig. 5.48 is not very supportive of data exploration, because the diagrams are very numerous and very small. It is much more difficult to group the local behaviours by similarity than in the case of the burglary rates in the USA, where we have only 51 states.



**Fig. 5.49.** The earthquake counts summed over the three periods (1976–1989, 1990–1992, and 1993–1999) are represented here by bar charts, after having been transformed into z-scores

We can try to simplify the view a little more by applying the results of our previous exploration of the other aspectual behaviour, specifically, the division into three time periods with different characters of behaviour. In Fig. 5.49, we have simplified the bar charts by representing the local earthquake frequencies in only the three periods. Like previously in Figs

5.45C and 5.46C, we have applied a standard normal transformation to the computed sums, so the bars represent z-scores.

This picture is easier to deal with than the previous one. We are now better able to detect clusters of compartments with similar profiles of the earthquake frequency in the three periods. As with Fig. 5.47, we can divide the territory into a “background” where the earthquake frequencies are low, and several relatively small “spots” with higher frequencies. Concerning these spots, it should be noted that the profiles of the earthquake frequencies differ quite a lot even between neighbouring locations, but there are also quite a few cases of neighbouring locations with earthquake frequencies above the mean that have similar profiles. Some of these cases are marked in Fig. 5.49.

Alternatively or additionally to the visual exploration of the spatial distribution of the local behaviours and “manual” grouping, we can try to apply a clustering tool and look at how it groups the local behaviours. We should remember, however, that the clustering tool will not take into account the spatial distribution of the local behaviours and the distances between them. In Fig. 5.50C, we see the result of applying a clustering tool to the local behaviours. The tool has built six clusters (we tried different numbers of clusters but found the result with six clusters to be the most appropriate). In Fig. 5.51C, we see six time graphs showing the envelopes of the lines belonging to each of the clusters.

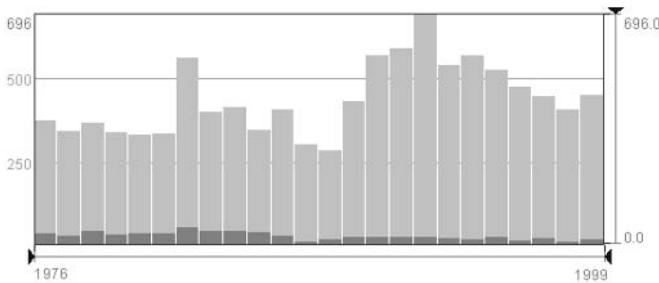
It cannot be said that the clusters are nicely shaped on the map – the compartments included in each of the clusters are quite scattered. The general character of the behaviours in each group can be understood from the time graphs. Thus, the cluster shown in the lightest shade of pink consists of the places where the earthquake frequencies are the lowest, and the cluster shown in the darkest red contains the “shakiest” places. The behaviour envelope represented in the middle of the right side of Fig. 5.51C is rather interesting: the earthquake frequencies were quite high until 1992 but then dramatically decreased. The major part of the respective cluster is situated in the south-western corner of the territory under study. This corresponds to our earlier observation concerning the time period from 1993 to 1999, when quite a few earthquakes occurred in the south-west.

The upper right time graph shows a quite opposite behaviour: low earthquake frequencies from the beginning until 1989 and then an increase to rather high values. The corresponding compartments are very scattered but are mostly located in the centre and the east of the territory. The locations are in agreement with our earlier observation concerning the increase in the earthquake-affected area in the centre and the east after 1989.

This is more or less the way in which the second aspectual behaviour might be explored and characterised. It may be noted that, simultaneously

with exploring and characterising the local behaviours and their spatial distribution, we have been establishing linkages with the pattern derived as a result of the exploration of the first aspectual behaviour, i.e. the temporal variation of the spatial distribution. Thus, we have used the division made earlier of the entire time period into three intervals for the simplification of the view of the local behaviours. We have also compared the behaviours in the clusters with this division and linked the difference between the spatial distributions in 1976–1989 and 1993–1999 to the peculiarities of the local behaviours in two groups of territory compartments. Furthermore, the cartographic representation of the clustering results can be compared with the maps of the earthquake frequency that we used for the exploration of the first aspectual behaviour, for example, the maps showing the summed frequencies for the three time periods (Fig. 5.45C). All such linkages and comparisons contribute to joining the aspectual patterns into an overall pattern, as is suggested in Table 5.11.

As in our exploration of the spatial aspect of the overall behaviour, when we aggregated the data over the entire temporal dimension, we can also aggregate the data over the entire value set of the spatial referrer and consider the temporal variation of the earthquake frequency for the whole territory. The result of the aggregation can be represented in a histogram display as is shown in Fig. 5.52 (the dark segments at the bottom of the histogram will be discussed a little later).



**Fig. 5.52.** Earthquake frequency by year for the entire territory. The proportions of earthquakes that have magnitudes of 4 or higher are highlighted

The data resulting from the aggregation have a single temporal referrer and hence correspond to case 1. The behaviour of the data can be explored as is suggested in Table 5.8, which, in turn, refers to Tables 5.6 and 5.7. We shall not discuss this behaviour in much detail but can note that, disregarding the outlier corresponding to the year 1982, it can be divided into three roughly homogeneous parts based on the intervals from 1976 to 1989, from 1990 to 1992, and from 1993 to 1999. These partial behaviours

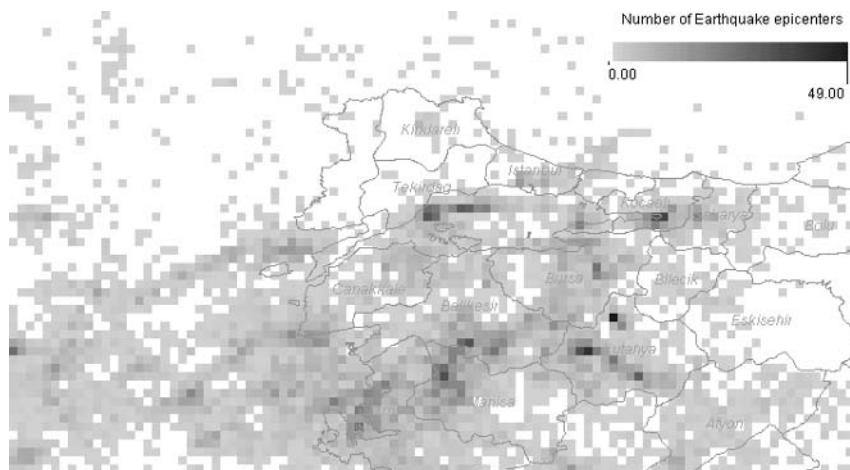
can be described as showing relatively stable frequencies in the range from 285 up to 417 earthquakes per year (except for the outlier of 564 in 1982) in the first interval, a rapid increase to 696 earthquakes per year in 1992 in the second interval, and an initial drop followed by a gradual decrease and stabilisation in the third interval. We would like to point out that this characterisation corresponds to our previous observations made with the use of the “small multiples” (Fig. 5.43): we noted then that the year 1982 was quite particular and divided the whole time period into almost the same intervals. The only difference is in the break between the first and the second interval. In fact, the situation in the year 1989, as may be seen from the “small multiples”, is similar both to that in 1988 and to that in 1990, and it is possible to regard the year 1989 both as the end of the first interval and as the beginning of the second interval.

We have now more or less finished with the exploration of the behaviour of the attribute “earthquake frequency” resulting from the aggregation of the data by space compartments and time intervals (years). According to Table 5.11, we need to join the aspectual and aggregate patterns derived thus far into a unified pattern. In fact, we have been establishing linkages between the patterns in the course of the exploration by noting commonalities between various findings. The job of bringing all the observations together is quite technical, and we do not think that we need to do this now.

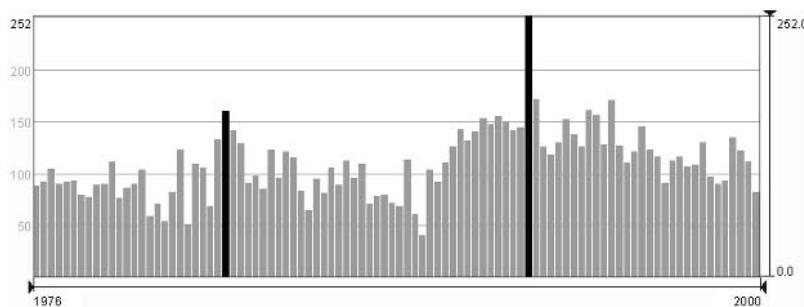
The attribute “earthquake frequency” is not the only attribute whose behaviour we need to explore and characterise. According to Table 5.13, we need to repeat the exploration procedure for the other attributes, specifically, the aggregated magnitudes and depths. Since our major goal here is to demonstrate the procedure, not to do a full analysis, we shall not reapply the procedure to the other attributes now. We shall only make a general note that each attribute behaves quite differently from the others, and it does not make much sense to join them into groups and consider the behaviours of the groups, as is suggested in Table 5.13 (with the reservation “if appropriate”). Concerning establishing linkages between the attributes, we could not detect any correlations or other indications of possible connections except for a slight positive correlation between the earthquake frequency and the maximum magnitude, which is demonstrated in a scatterplot in Fig. 4.84.

Let us assume that we have finished with Table 5.13 and return to Table 5.14, from which we started. According to Table 5.14, we need now to reaggregate the data and check the validity of the observations made earlier. In the present case, we can choose larger or smaller spatial compartments and longer or shorter time intervals.

We shall not do a detailed analysis with a different degree of data aggregation now but shall only note that the observations made earlier remain generally valid. As an illustration, we include here a visualisation of the spatial distribution of the earthquake frequency aggregated over the entire temporal referrer but by smaller territorial compartments (Fig. 5.53) and a display of the temporal behaviour of the earthquake frequency aggregated over the whole territory but by time intervals with an approximate length of one-quarter of a year (Fig. 5.54). These displays can be compared with those in Figs 5.47 and 5.52, respectively. It is interesting that the histogram in Fig. 5.54 reveals two outliers, which are marked in black. The left one corresponds to the second quarter of the year 1982, which agrees with the outlier for 1982 visible in Fig. 5.52. The right outlier corresponds to the second quarter of the year 1992. In Fig. 5.52, the bar for the year 1992 is the highest, but it does not look very much higher than the neighbouring bars. So, we have an opportunity to refine our earlier observations concerning the times of maximum seismic activity and the behaviours before and after those times. Analogously, we can define more precisely the time of the minimum seismic activity: the lowest bar in Fig. 5.54 corresponds to the last quarter of the year 1988. This bar looks much more unusual than the bar for the year 1988 in Fig. 5.52. It becomes clear from Fig. 5.54 that the earthquake frequency for the entire year 1988 does not look so low in Fig. 5.52 because the small number of earthquakes in the fourth quarter of this year has been summed with the quite large number in the second quarter of the same year.



**Fig. 5.53** The earthquake frequencies have been counted here for smaller territorial compartments



**Fig. 5.54.** The earthquake frequencies have been counted here for smaller time intervals (quarters of years). The highlighted bars correspond to the second quarter of the year 1982 (left) and the second quarter of the year 1992 (right)

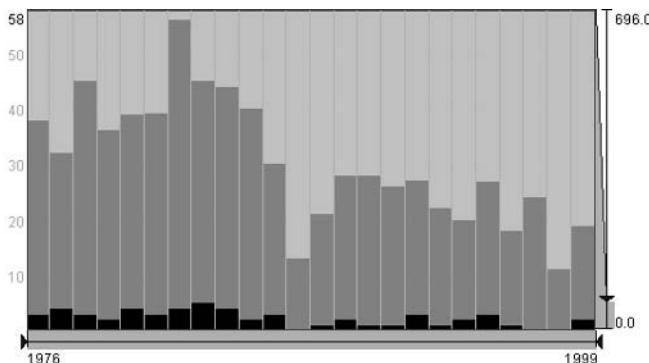
After trying various degrees of data aggregation, we can feel quite confident concerning what we have learned about the general character and major features of the behaviour of the dataset. Now, as Table 5.14 suggests, it is time to attend to various particulars: global and local outliers, atypical value combinations, and data subsets of special interest. We have detected some outliers in the course of the previous exploration; now, we can explore and characterise them in more detail. Thus, we can use querying and display linking in order to see the spatial distributions of the earthquakes in the shakiest period, the second quarter of 1992, and the quietest period, the fourth quarter of 1988, and to find out how the earthquake magnitudes varied in those periods (in fact, there were no strong earthquakes during the period of the highest earthquake frequency; the highest magnitude in that period was 4.5).

Besides attending to the unusual values and behaviours that have been discovered, it is also appropriate to do this for expectable atypical characteristics. In the earthquake dataset, high earthquake magnitudes are quite atypical. Therefore, this is an interesting target for a special investigation.

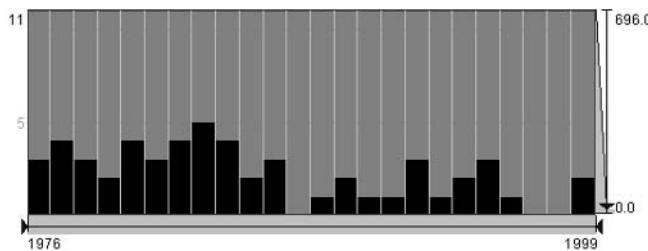
In the histogram in Fig. 5.52, the dark segments of the bars show the proportion of earthquakes with a magnitude of 4 or higher in all the earthquakes that occurred in the respective year. In Fig. 5.55, the histogram has been zoomed to make these segments more clearly visible. The proportions of earthquakes with magnitudes of 5 or higher can now be seen in black. Figure 5.56 shows the result of another zooming operation, which has made the black segments more clearly visible.

It can be noted that stronger earthquakes occurred more frequently in the period from 1976 to 1986 than after 1986. On a higher-granularity histogram (by quarters of years) in Fig. 5.57, we can see in more detail when the earthquakes with magnitudes of 5 or more occurred. In Fig. 5.58, we can see where these earthquakes occurred. By exploiting a dynamic link

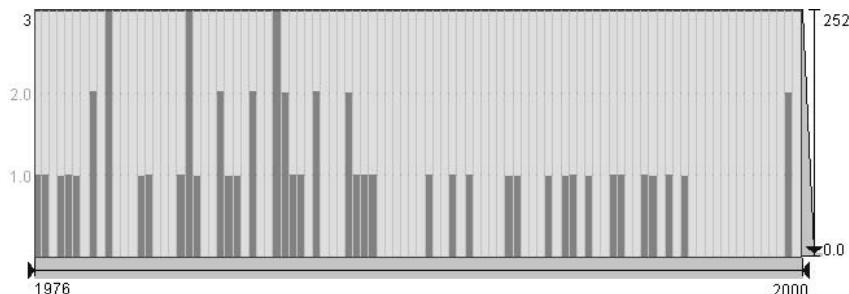
between the histogram and the map, we can find the locations of the strong earthquakes that occurred in particular time intervals. Using querying tools, we can access detailed data about any specific earthquake.



**Fig. 5.55.** The histogram from Fig. 5.52 has been zoomed here for better visibility of the numbers of earthquakes with magnitudes of 4 or higher (grey bars). The black segments show the proportions of earthquakes with magnitudes of 5 or higher



**Fig. 5.56.** The histogram has been zoomed once again here so that the numbers of earthquakes with magnitudes of 5 or higher can be seen better



**Fig. 5.57.** On this histogram display with bars corresponding to quarters of years, the dark segments show the numbers of earthquakes with magnitudes of 5 or higher. The display has been strongly zoomed into make the segments visible



**Fig. 5.58.** This map shows the locations of earthquakes with magnitudes of 5 or higher

In this way, by applying zooming and filtering, querying, and display coordination, we can explore various particulars as is suggested in Table 5.14. This table also says that we need to think whether the aggregation that we applied earlier could hide some potentially interesting features of the overall behaviour and, if so, we should explore the original behaviour (i.e. not aggregated) by manageable parts in order to detect such features. In the present case, the aggregation hides potentially existing spatio-temporal clusters of earthquake occurrences, i.e. sequences of earthquakes that occurred shortly one after another in the same place. So, we need to look for such clusters. In Sect. 5.4.5, we have described some methods for doing this, specifically, a visual search using a space–time cube display, and special computational techniques. Therefore, we shall not describe this part of the analysis here.

We have also tried to aggregate the earthquake data by the time of day. The result of this aggregation may be explored using a histogram display, as with aggregation by years or quarters of years. Therefore, we shall not describe the process of analysing this variant of aggregated data but shall note only that we did not detect anything interesting, except for a lower than usual earthquake frequency for times between 6 and 7 a.m.

In the course of our analysis, we have gone through many tables, corresponding to different cases in terms of the structure and properties of the portion of the data under analysis. Table 5.15 provides an overview of our route and summarises the actions taken and the tools applied.

**Table 5.15.** A schematic view of the process of analysing the earthquake dataset

Tables	Actions	Tools
Table 5.14	Aggregate data by spatial grid cells and time intervals	Aggregation
Table 5.13	Choose an attribute (frequency)	
Table 5.11	Choose behaviour aspect: temporal variation of spatial distribution	
Table 5.10	Explore the aspectual behaviour (temporal variation of spatial distribution)	Multiple maps Outlier removal Grouping/dividing Aggregation Querying
Table 5.11	Aggregate the data over the whole time by grid cells	Aggregation
Table 5.8	Explore the spatial behaviour of the aggregated data	Map
Table 5.7	Divide the behaviour into homogeneous parts	Map Marking, sketching
Table 5.6	Characterise the partial behaviours	Map Querying
Table 5.7	Define the applicability domains of the partial patterns, compare the patterns	Map Querying
Table 5.11	Choose another behaviour aspect: spatial distribution of local temporal behaviours	
Table 5.10	Explore the aspectual behaviour (spatial distribution of local temporal behaviours)	Map with diagrams Outlier removal Aggregation Clustering Aggregation
Table 5.11	Aggregate the data over the whole territory by time intervals	
Tables 5.8, 5.7, 5.6	Explore the temporal behaviour of the aggregated data	Time-based histogram Querying
Table 5.11	Establish linkages between the aspectual and aggregate patterns	Visualisation Display linking Querying
Table 5.13	Choose another attribute	
Tables 5.11, 5.10, ...	Repeat the analysis procedure for the other attribute	
Table 5.13	Compare the behaviours of the different attributes; establish linkages between the attributes	Visualisation Display linking Scatterplots
Table 5.13	Re-aggregate and validate patterns	Aggregation Visualisation
Table 5.14	Examine outliers and other particulars	Display linking Zooming and filtering Space-time cube
	Detect spatio-temporal clusters	

We would like to point out the wide range of exploratory tools that we have applied in the course of our example analysis. One could hardly find a single tool that, alone, could support every step and task of our explora-

tion. We would also like to note that the suggested analysis scheme does not impose a very strict sequence of actions. Of course, there are cases where some action creates prerequisites for another action and, hence, must be performed before that other action. For example, data aggregation must precede the exploration of the aggregated data, which is quite natural. In cases where there is no logically determined ordering, actions can be done in any sequence or even in parallel. Thus, the action “explore the original behaviour by scanning data subsets” at the end of Table 5.14 does not mean that this action must necessarily be the last in the analysis process – it is possible to do it right at the beginning, before any aggregation, or after the exploration of one of the aspects of the behaviour of the aggregated data. The investigation of outliers and other particulars need not necessarily be done after the observation of the general character and features of a behaviour – it may be more convenient to do these actions in parallel, as we actually did in our example. Some of the suggested actions may be skipped as being irrelevant to a particular case. For example, we did not try to discover correlations or influences between the spatial clusters of high earthquake frequencies, since we did not expect that such links might exist.

Unfortunately, we have no appropriate domain knowledge in order to judge whether our observations are really meaningful and interesting. As we said earlier, the use of domain knowledge is very welcome in data analysis. Most probably, if we were seismologists, we would not have undertaken such a broad investigation but would have focused from the very beginning on specific aspects, features, and subsets of interest. It is also probable that we would have used some domain-specific analysis methods and tools. It might also be that we would not have even tried to explore these data, knowing in advance that they alone could not tell us anything interesting and that additional data would need to be used.

So, our analysis should be regarded as just a demonstration of how the scheme suggested in the previous section can be applied. It also demonstrates that the scheme is not an algorithm that can be executed formally and mindlessly. Nevertheless, the scheme can provide useful guidance to those who might need it, and this is what we wanted to achieve.

## Summary

The principal objective of this chapter has been to relate exploratory tools to the tasks that they can support. Recall that we have used the word “tasks” to refer to various questions concerning data that a data analyst may seek answers to. According to our idea, the links should be directed

not from tools to tasks but from tasks to tools. We believe that this direction is more practical as a basis for guiding data analysts in choosing right tools for their tasks.

Although the idea of creating an instruction book that recommends appropriate tools for all possible tasks may seem quite straightforward, there is an inherent problem that makes it hard to achieve. The real tasks arising in the course of data analysis are too specific: they are always formulated in terms of particular data components or even particular data items. For this reason, the possible tasks are countless and hence cannot be listed in any book. The idea of considering task categories instead of tasks is also ineffective: the task categories are too broad, in the sense of embracing a mixture of tasks that differ substantially from each other owing to the differences in the structure and properties of the respective datasets. Hence, any tool suitable for one group of tasks may turn to be completely inappropriate for other tasks belonging to the same task category. Each task category has therefore to be linked to a long list of tools or to broad tool categories. This sort of linkage cannot be regarded as providing good guidance.

As a way to overcome this difficulty, we arrived at the idea of formulating general principles for tool selection that could help analysts in finding the right tools for their own tasks whatever data they might need to analyse. It was clear from the very beginning that the content of such principles would necessarily extend beyond just tool selection. Choosing a tool means, in fact, deciding on a particular approach to processing and analysing the data. Moreover, this implies a particular attitude to the data, a particular way of treating it. Hence, the principles have to deal with approaches and attitudes as much as with tools. They teach us, in quite a broad sense, how to do exploratory data analysis, which involves adopting a certain attitude of mind, adhering to a certain philosophy, paying attention to certain aspects, organising the work in a certain way, and, as a consequence of all these considerations, choosing certain approaches and certain kinds of tools.

The principles that we have formulated are strongly related to the tasks. They have been derived from an examination of our experience: how, being equipped with a variety of tools, we usually deal with a new dataset in accordance with the general philosophy of exploratory data analysis. We have taken the task “characterise the overall behaviour of the characteristics over the entire set of references” as the primary task of EDA and considered how we approach it and decompose it into subtasks, what we look for, how we bring together the bits and pieces of information gained, and, of course, what tools we use for the decomposition, characterisation, and synthesis.

At the same time, these principles correspond greatly to the ideas of other researchers in the areas of visualisation, data analysis, systems analysis, and cognitive psychology. We have mentioned the relation of our principles to Ben Shneiderman's "Information Seeking Mantra", to Jacques Bertin's image theory and the primacy of the overall level of information processing, and to gestalt psychology and Rudolf Arnheim's ideas about "visual thinking". We have not explicitly mentioned the relation of our ideas concerning reference-invariant depiction of a behaviour and concerning approaches to the decomposition of a complex behaviour to George Klir's general theory of system analysis, and would like to acknowledge here that we were greatly influenced by this theory.

So, we have presented each of the ten principles that we have arrived at with many examples of their operation. At the end, we have provided a sort of guide through the principles and the corresponding tasks and tools. For this purpose, we have taken the general task "characterise the overall behaviour" and considered how to perform it in various situations, depending on the dimensionality and size of the reference set and the number of attributes. We did not consider the possible types of data components; this needs to be done specifically for each dataset. An analyst is thus expected to apply these general procedures and principles to his/her specific case, as we did in the example of the earthquakes. We hope that the multitude of examples provided in this and the previous chapter can help analysts to make the right choice of tools and approaches.

## References

- (Arnheim 1997) Arnheim, R.: *Visual Thinking* (University of California Press, Berkeley 1969, renewed 1997)
- (Bertin 1967/1983) Bertin, J.: *Semiology of Graphics. Diagrams, Networks, Maps* (University of Wisconsin Press, Madison 1983). Translated from Bertin, J.: *Sémiologie graphique* (Gauthier-Villars, Paris 1967)
- (Fredrikson et al. 1999) Fredrikson, A., North, C., Plaisant, C., Shneiderman, B.: Temporal, geographic and categorical aggregations viewed through coordinated displays: a case study with highway incident data. In: *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation (NPIVM'99)* (ACM Press, New York 1999) pp. 26–34.
- (Furnas 1986) Furnas, G.W.: Generalized fisheye views. In: *Proceedings of CHI'86* (ACM, New York 1986) pp. 16–23
- (Gatalsky et al. 2004) Gatalsky, P., Andrienko, N., Andrienko, G.: Interactive analysis of event data using space–time cube. In: *IV 2004: 8th International Conference on Information Visualization, Proceedings*, ed by Banissi, E. et al,

- London, July 2004 (IEEE Computer Society, Los Alamitos 2004) pp. 145–152
- (Hägerstrand 1970) Hägerstrand, T.: What about people in regional science? *Papers, Regional Science Association* **24**, 7–21 (1970)
- (Hedley et al. 1999) Hedley, N.R., Drew, C.H., Arfin, E.A., Lee, A.: Hägerstrand revisited: interactive space–time visualizations of complex spatial data. *Informatica: An International Journal of Computing and Informatics* **23**(2), 155–168 (1999)
- (Keogh and Kasetty 2003) Keogh, E.J., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration, *Data Mining and Knowledge Discovery* **7**(4), 349–371 (2003)
- (Klir 1985) Klir, G.J.: *Architecture of Systems Problem Solving* (Plenum, New York 1985)
- (Kraak 2003) M.-J. Kraak, The space–time–cube revisited from a geovisualization perspective. In: *Proceedings of the 21st International Cartographic Conference (ICC)*, Durban, South Africa, August 2003, pp. 1988–1996
- (MacEachren 1995) MacEachren, A.M.: *How Maps Work: Representation, Visualization, and Design* (Guilford, New York 1995)
- (Mackinlay et al. 1991) Mackinlay, J.D., Robertson, G.G., Card, S.K.: The perspective wall: detail and context smoothly integrated. In: *Proceedings of the Conference on Human Factors in Computing Systems CHI'91* (ACM Press, New York 1991) pp. 173–179
- (Miller and Han 2001) Miller, H.J., Han, J.: Geographic data mining and knowledge discovery: an overview. In: *Geographic Data Mining and Knowledge Discovery*, ed. by Miller, H.J., Han, J. (Taylor & Francis, London 2001) pp. 3 –32.
- (Openshaw and Openshaw 1997) Openshaw, S., Openshaw, C.: *Artificial Intelligence in Geography* (Wiley, Chichester 1997)
- (Shneiderman 1996) Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages* (IEEE Computer Society Press, Piscataway 1996) pp. 336–343
- (Spence 2001) Spence, R.: *Information Visualisation* (Addison-Wesley, Harlow 2001)
- (Spence and Tweedy 1998) Spence, R., Tweedy, L.: The attribute explorer: information synthesis via exploration, *Interacting with Computers* **11**, 137–146 (1998)
- (StatSoft 2004) StatSoft, Inc.: *Electronic Statistics Textbook* (StatSoft, Tulsa 2004), <http://www.statsoft.com/textbook/stathome.html>. Accessed 28 Mar 2005
- (Tufte 1983) Tufte, E.R.: *The Visual Display of Quantitative Information* (Graphics Press, Cheshire CT, 1983)
- (Tufte 1990) Tufte, E.R.: *Envisioning Information* (Graphics Press, Cheshire CT, 1990)

## 6 Conclusion

Throughout our professional life, we have been involved in various projects in which some partners had data and wanted to understand what these data meant and how they could be used. These partners often complained that, despite having tools and even instructions on how to use the tools, they could not figure out how to apply these tools to their own data in a sensible way. Usually, we proposed that they sent us their data so that we could “play” with those data and thereby find the right ways to handle them. As a result of such “playing”, we produced collections of screenshots demonstrating the observations we had made and the features we had discovered. We also made suggestions concerning further analyses and the possible uses of the findings.

Our partners have often asked us, “How do you know what tool or method to apply in this or that situation?” The answer is obvious: this comes from experience, from regular “playing” with various data and various tools, which converts one’s knowledge of the tools and methods from theoretical to practical, from declarative to procedural.

However, just as the data owners and domain specialists encountered difficulties in choosing appropriate approaches and tools, we also encountered difficulties in analysing the data because of our lack of domain knowledge and insufficient understanding of the semantics of the data provided to us. Because of this, it turned out sometimes that our findings were meaningless or trivial to domain experts. In fact, the most effective and fruitful cases of data exploration took place when we had an opportunity to sit together with domain experts. The experts formulated their questions concerning the data, and we chose the appropriate visualisations, data transformations, divisions, computations, etc., which helped the experts to find the answers.

This means that what is needed is to find a way of combining domain knowledge and expertise in using tools. It is hardly possible to supply every domain expert with a professional analyst who knows the tools that exist and is experienced in using them, although some high-level specialists and decision makers enjoy such a privilege.

Some of our partners have asked us, “Can you teach us to analyse our data by ourselves, without calling for your help?” We had to answer that

this was not an easy task. Experience is not a set of verbal statements but something tacit, probably subconscious, and significant effort is needed to externalise it, systemise it, and make it comprehensible to others. However, we understood that we ought at least to try to do this. So we did, and the result is this book.

We hope that the book will be useful. However, we also understand that reading it will not guarantee that one will know how to approach this or that particular dataset. The reason is simple: theoretical knowledge, which can be acquired by reading this book, is not the same as practical experience. One needs sufficient practice in order to turn the theory into one's own skills. Hence, even if we assume that every person who is going to explore data will read this book (which is a rather bold assumption), the desired synergy of domain knowledge and tool-related experience will not be reached.

Is it possible to cope with this problem? We believe that good solutions will eventually be found. As computer scientists, we can for the present suggest a solution, which seems to us feasible and worth trying out. The idea is to put knowledge about tools and methods for analysing data on the computer side so that the computer could act as an intelligent assistant to a domain expert who is exploring data. The assistance might include

- suggestion of appropriate procedures and methods for data analysis;
- selection of relevant tools from a tool kit;
- automatic application of the tools to the data;
- help in the operation of the tools;
- instruction concerning the interpretation and use of the outcomes of using the tools.

For such intelligent functioning, the computer, of course, needs to be knowledgeable about the capabilities of the tools and their applicability domains, i.e. what data they allow as their input. It needs to be aware of the typical tasks of exploratory data analysis and the approaches to performing them. It needs to know how to handle various complex cases by decomposing them into simpler ones. In short, the content of this book plus information about the specific tools available in the tool kit need to be put in the computer's "mind", which is quite possible.

This is not enough, however. In order to make apt suggestions, the computer must understand the meaning of the data under analysis, exactly like a human analyst. Thus, we have sometimes received data with components with names such as "QN", "TNN", and "TMM", or with names in other languages that we could not understand. Without having our partners' explanations of what the components meant, trying to apply any tool or

method to the data would be completely pointless. The computer also needs the user, i.e. the domain expert, to state the meaning of the data components. However, a true “understanding” of the semantics of the data is hardly achievable and, in fact, is not necessary. In our case, we visualise the data, look at the displays, and try to interpret what we see. In the case of a computer cooperating with a domain expert, it is supposed that it is the expert who looks and interprets, whereas the computer only visualises, transforms, and computes. The limited understanding of the data required by the computer includes knowing

- which data components are referential and which are characteristics;
- the types of the components: spatial, temporal, population, numeric, ordinal, or nominal;
- some other characteristics, in particular, the meaning of spatial references: whether they are discrete locations, sample locations in which a continuous phenomenon is observed and measured, discrete spatial objects, or territorial divisions.

This information needs to be provided to the computer at the beginning of the data analysis. A good human–computer interface design is certainly needed in order to make the procedure of informing the computer about the data quick and easy for anyone who wishes to utilise the intelligent services. Another design problem is to make the computer assistant supportive but not annoying, advising but not prescriptive, and informing but not boring.

We believe that these design problems are solvable, and true human–computer partnership in exploratory data analysis will soon be achieved. We hope that we shall be able to contribute to this.

# Appendix I: Major Definitions

## I.1 Data

### Data

Data are viewed abstractly as a set of records with a common structure, each record being a sequence of elements, such as numbers or strings, which either reflect the results of some observations or measurements or specify the context in which the observations or measurements were obtained. The context may include, for example, the place and the time of observation or measurement, and the object or group of objects characterised.

A dataset reflects characteristics of a certain phenomenon. By means of data analysis, an explorer gains knowledge about that phenomenon.

The elements that a data record consists of are called *values*. Values that reflect results of measurements or observations are also called *characteristics*. Values that reflect the context of the observation or measurement are called *references*.

For example, to make a study of climate, one measures various properties of the climate such as the air temperature and the wind direction in various places and at various time moments. Each combination of measured values of the air temperature and wind direction refers to a particular place and a particular time moment, which are indicated in the corresponding data record. The measured values of the air temperature and wind direction are characteristics. The data elements indicating the places and time moments are references.

### Structure and Components of Data

All records of a dataset are assumed to have a common structure, with each position having its specific meaning, which is common to all values appearing in it. These positions may be named to distinguish between the positions. The positions are usually called *components* of the data.

A component may correspond to a certain measured or observed property of the phenomenon reflected in the data, for example air temperature or wind direction, or may reflect a certain aspect of the context in which the observations or measurements were made, for example the geographical location or time moment.

All possible elements that can potentially appear in data as values of a particular component constitute the *value domain* of that component. If the value domain consists of numbers, it is often viewed as a *value range*, defined by specifying the minimum and maximum possible values of the component.

### **Attribute, or Characteristic Component**

A data component corresponding to a measured or observed property of the phenomenon reflected in the data is called a *characteristic component*, or *attribute*. Some examples of attributes are air temperature or wind direction, which reflect properties of the climate. Values of attributes are also called *characteristics*.

### **Referrer, or Referential Component**

A data component reflecting an aspect of the context in which the observations or measurements were made is called a *referential component*, or *referrer*. For example, the geographical location and the time moment are referrers for measurements of properties of the climate such as air temperature or wind direction.

The most frequently occurring types of referrers are *space*, *time*, and a (statistical) *population*, i.e. a collection of items or, more generally, any referrer with a value domain that has no ordering and no distances between the elements.

### **Reference**

The value of a single referrer or the combination of values of several referrers that fully specifies the context of some observation(s) or measurement(s) is called a *reference*, or, more specifically, the reference of the characteristic(s) obtained in this context.

## Reference Set

The set consisting of all references occurring in a dataset is called the *reference set* of this dataset. The reference set of a subset of data consists of all references occurring in this subset.

## Multidimensional Data

A dataset with two or more referrers is called *multidimensional*. A dataset with two referrers may be called two-dimensional, a dataset with three referrers three-dimensional, and so on. Attributes are not counted as dimensions of a dataset.

It may also be said about a particular dataset that it has a *multidimensional reference set* (two-dimensional, three-dimensional, etc.)

## Independent and Dependent Data Components

Referrers are regarded as *independent* data components, since the context for making observations or measurements, i.e. the times, places, objects to be observed, etc., may be usually chosen arbitrarily, and the choice concerning any particular aspect of the context, such as time or space, may be made independently of the other aspects.

A particular choice of the context fully determines the characteristics obtained in that context. Hence, characteristics (attribute values) depend on references, and attributes are therefore *dependent* components. Attribute values are always associated with particular references and have no meaning separately from the references. For example, a particular value of the air temperature is meaningless if the place and time of its measurement are unknown.

It may also be said that references are *subjective*, since an observer may choose them more or less arbitrarily, and characteristics are *objective*, since they reflect something measured or observed rather than arbitrarily chosen.

## Data Function, Functional Data Model

Data may be viewed formally as a function, in the mathematical sense, with the referrers as independent variables and the attributes as dependent variables. The function, which is called the *data function*, defines the correspondence between the references (combinations of values of the refer-

fers) and the characteristics (combinations of values of the attributes). For each combination of values of the referential components, there is no more than one combination of values of the attributes.

A data function may be represented by the formula (see Chap. 3)

$$\mathbf{f}(x_1, x_2, \dots, x_M) = (y_1, y_2, \dots, y_N) \quad (3.2)$$

where  $M$  is the number of referrers in the dataset,  $N$  is the number of attributes,  $\mathbf{f}$  is the function symbol, the independent variables  $x_1, x_2, \dots, x_M$  stand for the referrers, and the dependent variables  $y_1, y_2, \dots, y_N$  stand for the attributes.

Any attribute of a dataset may be considered independently of the other attributes. This allows the formula (3.2) to be split into an equivalent set of expressions:

$$\begin{aligned} f_1(x_1, x_2, \dots, x_M) &= y_1 \\ f_2(x_1, x_2, \dots, x_M) &= y_2 \\ &\dots \\ f_N(x_1, x_2, \dots, x_M) &= y_N \end{aligned} \quad (3.3)$$

Each of the  $N$  functions  $f_1, f_2, \dots, f_N$  represents one of the attributes.

## Ordered Component, Ordering of Values

*Ordering* is defined mathematically as a binary relation, i.e. a relation between two items, which has the following properties (in the expressions below, the symbol “ $\leq$ ” denotes the ordering relation):

1. *Antisymmetry*: For any two items  $a$  and  $b$ , if  $a \leq b$  and  $b \leq a$ , then  $a = b$ .  
This means that if the items  $a$  and  $b$  are different, the statements  $a \leq b$  and  $b \leq a$  may not be true simultaneously.
2. *Transitivity*: For any three items  $a$ ,  $b$ , and  $c$ , the truth of the statements  $a \leq b$  and  $b \leq c$  implies that  $a \leq c$ .

A data component is called *ordered* if an ordering relation exists between at least some elements of its value domain. The ordering among the elements of a value domain is called *linear* or *total* if, for any pair of elements  $a$  and  $b$  from this domain, either  $a \leq b$  or  $b \leq a$ ; otherwise, the ordering is called *partial*.

Some examples of linearly ordered value domains are the set (range) of values of a numeric data component, and a set of time moments.

## Set (Value Domain) with Distances

*Distance* is a numeric measure defined for pairs of elements of a set (a value domain) and has the following properties (in the statements below, the expression  $d(a, b)$  denotes the distance between the elements  $a$  and  $b$ ):

1.  $d(a, b) \geq 0$  (the distance between any two elements is non-negative).
2.  $d(a, a) = 0$  (the distance from an element to itself equals zero).
3.  $d(a, b) = d(b, a)$  (the distance from  $a$  to  $b$  is the same as the distance from  $b$  to  $a$  for any two elements  $a$  and  $b$ ).
4.  $d(a, c) \leq d(a, b) + d(b, c)$  (for any three elements  $a$ ,  $b$ , and  $c$ , the distance between any two of them is not more than the sum of the distances from each of them to the third element).

A set or value domain is considered as a *set with distances* if it is possible to determine the distance between any two elements.

Some examples of sets with distances are space, time, and the value range of a numeric data component. Distances between numeric values are usually defined as the arithmetic differences between them. Distances between time moments are the lengths of the time intervals between them. A distance in space is often defined as the length of the straight line connecting the pair of locations. However, it is possible to define distances differently. For example, in geographical space, distances may take account for the Earth's curvature and/or relief or be measured along roads.

## Continuous Set (Value Domain)

A set with distances is *continuous* if, for any element  $a$  and any number  $D > 0$  (which may be arbitrarily small but not equal to zero), there is another element  $b$  in this set,  $b \neq a$ , such that  $d(a, b) < D$  (the distance from  $a$  to  $b$  is less than  $D$ ).

## I.2 Tasks

### Behaviour

The *Behaviour* of a data function (or of an attribute or group of attributes) over a set of references is the particular configuration (arrangement) of the characteristics corresponding to all the elements of this reference set, taken

together, and considered together with the relations that exist between references.

Thus, the behaviour of a function over a linearly ordered reference set is the particular sequence of characteristics that corresponds to the sequence of ordered references. The behaviour of a function over a space (which is an unordered set with distances) is the distribution of the characteristics over this space. This includes relations between various characteristics with respect to the distances between the corresponding references, for example whether the changes between neighbouring locations are smooth or abrupt. The same can be said concerning a behaviour over any reference set with distances. A behaviour over a population, i.e. a reference set without any relations between the elements and without distances, may be viewed as the frequency distribution of the various value combinations.

When we are considering the behaviour of a data function (or attribute or attribute group) with respect to some reference set, this reference set may be called the *base* of this behaviour.

## **Pattern**

A *pattern* is a construct that reflects essential features of a behaviour in a parsimonious manner, i.e. in a substantially shorter and simpler way than specifying every reference and the corresponding characteristics. The construct may be a description in some language (natural, formal, or graphical) or a mental image of the behaviour. Some examples of patterns are an increasing or decreasing trend of a numeric attribute over time, a spatial cluster of events or of high attribute values, and a skewed frequency distribution of attribute values over a statistical population.

A pattern derived by means of observing or analysing a behaviour is said to *approximate* this behaviour. Different patterns may approximate one and the same behaviour.

## **Compound Pattern**

A *compound pattern* approximating some behaviour is a combination of two or more patterns such that each of these patterns approximates only a part of the behaviour but all the patterns jointly approximate the entire behaviour.

## Aspectual Behaviour, Aspectual Pattern, Behaviour Slice

When the base of some behaviour is a multidimensional reference set, it is possible to choose a specific value of one referrer and consider the behaviour of the attributes with respect to the other referrers, i.e. over the subset composed of all references containing the chosen referrer value. The latter behaviour may be called a *slice* of the entire behaviour.

Choosing different values of a referrer gives different slices of the entire behaviour. The particular arrangement of the slices with respect to the whole set of values of this referrer and relations between them is called an *aspectual behaviour*.

For a two-dimensional dataset, i.e. a dataset with two referrers  $\mathbf{R}_1$  and  $\mathbf{R}_2$ , there are two aspectual behaviours of the attributes:

1. The behaviour (arrangement) of the behaviour slices based on  $\mathbf{R}_1$  over the set of values of  $\mathbf{R}_2$ .
2. The behaviour (arrangement) of the behaviour slices based on  $\mathbf{R}_2$  over the set of values of  $\mathbf{R}_1$ .

For example, for a behaviour with a base formed by a spatial and a temporal referrer, the partial behaviours are the following:

1. The behaviour of the spatial behaviour over time, i.e. how the spatial distribution changes over time.
2. The behaviour of the temporal behaviour over space, i.e. how the local temporal behaviours (behaviours at individual locations) are distributed over space.

For a dataset with  $N$  referrers, the number of different aspectual behaviours is  $N!$  ( $N$  factorial), i.e.  $N \cdot (N-1) \cdot (N-2) \cdots 2 \cdot 1$ . For the case of three referrers, this yields six aspectual behaviours. For example;

1. The behaviour (arrangement) over the set of values of  $\mathbf{R}_1$  of the aspectual behaviour over the set of values of  $\mathbf{R}_2$  formed by the behaviour slices based on  $\mathbf{R}_3$ .
2. The behaviour (arrangement) over the set of values of  $\mathbf{R}_1$  of the aspectual behaviour over the set of values of  $\mathbf{R}_3$  formed by the behaviour slices based on  $\mathbf{R}_2$ .
3. The behaviour (arrangement) over the set of values of  $\mathbf{R}_2$  of the aspectual behaviour over the set of values of  $\mathbf{R}_1$  formed by the behaviour slices based on  $\mathbf{R}_3$ .

And so on.

## Task, Exploratory Task, Data Analysis Task

A *task* is a question concerning data that can be answered on the basis of the information contained in the data, for example “What characteristics correspond to this reference?” or “What is the behaviour of this attribute (or group of attributes) over this reference set?”

A task is viewed as consisting of two parts: a target, i.e. what information needs to be obtained, and the constraints, i.e. what conditions this information needs to fulfil. The target and constraints can be understood as unknown and known (specified) information, respectively; the goal is to find the initially unknown information corresponding to the specified information.

In the task “What characteristics correspond to this reference?”, the specified reference is the constraint and the corresponding characteristics are the target. The constraint specifies, besides the reference, its relation to the characteristics specified by the target: these characteristics must correspond to the reference (as defined by the data function).

In the task “What is the behaviour of this attribute (or group of attributes) over this reference set?”, the specified reference set, including all its elements and the relations between them, if any, is the constraint. Another constraint is the specification of the attribute or group of attributes and the requirement that the behaviour must be based on the specified reference set. The target is some pattern that approximates this behaviour appropriately.

## Elementary Task, Elementary Level of Analysis

An *elementary task* is a task stated in terms of individual elements, i.e. individual references and characteristics, for example “What characteristics correspond to this reference?” An elementary task may involve two or more references and/or characteristics, which are dealt with as individual items rather than as a unified whole, for example “Compare the characteristics corresponding to these two references”.

The *elementary level of analysis* is the finding of answers to various elementary tasks. Elementary tasks play a marginal role in exploratory data analysis.

## Synoptic Task, Synoptic Level of Analysis

*Synoptic tasks* are tasks stated in terms of sets of references and the corresponding behaviours of attributes or attribute groups, for example “What is

the behaviour of this attribute (or group of attributes) over this reference set?" A set of references in a synoptic task is considered as a unified whole.

The *synoptic level of analysis* is the finding of answers to various synoptic tasks. Synoptic tasks play the primary role in exploratory data analysis.

## **Comparison, Comparison Task**

*Comparison* is understood in a broad sense as identifying the relations existing between two or more items, which may be individual references or characteristics, sets of references, or behaviours. The types of relations that may be of interest to an explorer include

- “same” or “different” (between any items);
- order (between values of an ordered component or sets of such values);
- distance (between values of a component with distances);
- “including”, “overlapping”, or “not overlapping” (between sets);
- “similar”, “dissimilar”, or “opposite” (for behaviours).

## **Behaviour characterisation**

A *behaviour characterisation* task is a task that may be stated in the form “What is the behaviour of this attribute (or group of attributes) over this reference set?” or, in slightly other words, “What pattern can adequately approximate the behaviour of this attribute (or group of attributes) over this reference set?”

The exploratory analysis of a particular dataset may be viewed as finding the answer to the overall behaviour characterisation task, i.e. the question “What pattern can adequately approximate the overall behaviour of all the attributes over the entire reference set?” In the course of the analysis, this general task is decomposed into smaller, less general tasks of various types.

## **I.3 Tools**

### **Visualisation**

*Visualisation* is understood as the representation of data in a visual form, i.e. creating various pictures from data: graphs, plots, diagrams, maps, etc.

For this purpose, elements of data are translated into graphical features such as positions within a display, colours, sizes, or shapes.

## Marks

*Marks* are any distinguishable visual items that can appear in a display, such as dots, lines, or shapes.

## Display Dimensions

The *display dimensions* provide a set of positions at which marks or groups of marks can be placed. The *primary display dimensions* are:

- two spatial dimensions, width and height, which are available in any display medium, including computer screens and paper;
- the third spatial dimension, depth, which is available in some technologically advanced media and can be simulated on a two-dimensional medium;
- the temporal dimension, i.e. the display time, which may be available in a computer-based visualisation and means that the content of the display changes over time.

Besides the primary dimensions, there are also secondary dimensions called *arrangements*, which use the primary dimensions as a basis.

## Arrangements

*Arrangements* are used for the following purposes:

- to provide positions for multiple displays (rather than elementary marks), which represent certain parts of the data and may have their own, internal dimensions;
- to change the perceived properties of the display space, for example to mitigate the perception of continuity of the space.

The most frequently used arrangements are:

- *space partitioning*: The display space is divided into compartments, in which multiple displays may be put;
- *space embedding*: Positions in the space of one display are used for placing other displays, which are superimposed on it;
- *space sharing*: Overlaying multiple displays within the same space;

- *space transformation*: Changing the perceived properties of the space and introducing specific relations between positions.

Arrangements utilise the primary display dimensions and are, therefore, secondary with respect to them.

## Retinal Variables

The *retinal variables* are abstractions of various visual properties of marks: colours, shapes, sizes, etc. Thus, for shapes, the retinal variable “shape” can be introduced, and all possible shapes are considered as its *values*. The most frequently used visual variables are colour hue, colour brightness (or darkness), colour saturation, size (which may be subdivided into length, height, area, and volume), texture, shape, and orientation.

## Visual Encoding Function

A *visual encoding function* is a mechanism specifying the conversion of data items into values of display dimensions and visual variables, i.e. it defines how each data item is represented in a display. Such a mechanism may have the form of a rule or a set of rules, a formula or a set of formulae, etc.

A visual encoding function typically involves items that may be chosen arbitrarily from a range of options, for example particular colours or the maximum size of a mark. The function can thus be generalised by substituting variables for such arbitrary items and defining the domains of admissible values for these variables, which are called *parameters* of the visual encoding function. By assigning various values to the parameters, one can obtain a family of specific visual encoding functions.

## Display Manipulation

*Display manipulation* includes various interactive operations that change the values of the parameters of the visual encoding function applied in the display and thereby modify the appearance of the display. *Dynamic display manipulation* means that the display reacts immediately to any change in the parameters of the visual encoding function by updating the picture in accordance with the new parameter values.

In exploratory data analysis, it is not any kind of display manipulation that is of interest, but only such manipulation that can facilitate or prompt

the analysis; for example, it may allow the answers to various questions to be found faster or make a pattern “pop up” from the display.

Display manipulation should be distinguished from *data manipulation*: display manipulation does not change the data but changes only the visual representation of the data, while data manipulation modifies the data, one result of which may be a change in the appearance of visual displays of those data.

## **Appendix II: A Guide to Our Major Publications Relevant to this Book**

In this book, we have omitted many details of our work that are relevant to the exploratory analysis of spatio-temporal data. In this appendix, we would like to provide references to some of our papers that extend the material presented in this book.

In recent decades, we have made several attempts to use expert knowledge of data visualisation principles for the automated generation of data displays. Thus, in [1], we describe how various thematic mapping techniques can be chosen automatically, depending on the characteristics of the data, such as the number and types of attributes and the semantic relationships between them. To enable knowledge-based design of thematic maps, it is necessary to describe the semantics of the data. We discuss the relevant aspects of the semantics of data in [2]. Reference [3] proposes a dialogue procedure for acquiring such information from domain experts. This procedure is an adaptation of our previous work on knowledge engineering and expertise transfer [4]. In our later papers, we extend the idea of knowledge-based user support from automated visualisation design to helping users to choose and apply various tools for exploratory data analysis. In [5], we discuss what categories of knowledge are needed for an intelligent software assistant. Reference [6] describes how knowledge-based visualisation and intelligent guidance can support data analysts and decision makers.

In parallel to our research on knowledge-based visualisation design and user guidance, we have developed a concept of interactive maps that change their appearance in response to manipulation by the user [1]. This concept was later extended to dynamic classification maps [7] and to techniques for the exploration of raster data [8]. Reference [9] reports the results of our study of the usability of interactive maps. Our general experience is that new users must first learn and “feel” the high interactivity of the novel tools with some examples. A short introduction of 30 to 60 minutes and some hands-on experience should generally induce a sufficient sense of fun and sufficient courage that users can continue with their own exploration of further tools.

The next group of publications relates to our contribution to forming the research agenda for geovisualisation and computer cartography. Reference [10] demonstrates the need for extending cartographic knowledge to interactive and dynamic maps. Two collective papers consider the research agenda in cartographic representation [11] and in the design of geovisualisation tools [12]. References [13, 14] present some steps towards the implementation of the research agenda in the area of the visualisation and interactive exploration of spatio-temporal data. Writing these papers initiated our thinking about this book.

The next series of publications presents the tools and techniques that we have designed to support the exploratory analysis of various categories of spatio-temporal data: exploration of object movement [15], detection of changes and analysis of variance in spatially distributed time series data [16, 17], characterisation and comparison of spatial development scenarios [18], and analysis of point events [19].

Several publications deal with the use of interactive statistical graphics. Thus, in [20], we have suggested a procedure of classification according to the dominant attribute. In [21], we have considered several different ways of scaling the axes of parallel-coordinates displays with the aim of supporting particular types of tasks. Reference [22] proposes an extension of the parallel-coordinates technique to large data sets. In [23], we introduce our extension of the cumulative-curve technique that generalises the ideas of histograms and the Lorenz curve.

The next group of publications reflects our work on visual data mining – the combination of interactive visualisation with computational methods of data analysis. We have proposed some methods for the visualisation of data-mining outputs and for the use of various data-mining techniques in combination with thematic maps [24]. In [22], we suggest some specific visualisation enhancements for cluster analysis. Reference [25] describes the integration of two software research prototypes: Descartes for geographic visualisation, and Kepler for data mining.

A significant part of our research relates to multicriteria decision analysis. We have highlighted the importance of visualisation for this kind of activity and proposed several visualisation techniques supporting various computational optimisation methods [26], as well as purely visual and interactive decision support methods that suit a variety of individual decision-making styles [27]. In [28], we discuss the value of display coordination for making informed, well-grounded spatial decisions.

In several publications, we describe the application of our tools and data exploration methods in various domains: simulation modelling [29, 30], forestry [31], seismology [32], and official statistics [23, 33].

## References

1. Andrienko, G., Andrienko, N.: Interactive maps for visual data exploration. *International Journal of Geographical Information Science* **13**(4), 355–374 (1999)
2. Andrienko, G., Andrienko, N.: Data characterization schema for intelligent support in visual data analysis. In: *Spatial Information Theory - Cognitive and Computational Foundations of Geographic Information Science, COSIT'99*, ed. by Freksa, C., Mark, D., Lecture Notes in Computer Science, Vol. 1661 (Springer, Berlin, Heidelberg 1999) pp. 349–366
3. Andrienko, G., Andrienko, N.: Knowledge engineering for automated map design in DESCARTES. In: *Advances in Geographic Information Systems*, ed. by Medeiros, C.B., 7th International Symposium ACM GIS'99, Kansas City, November 1999 (ACM Press, New York 1999) pp. 66–72
4. Andrienko, G., Andrienko, N.: AFORIZM approach: creating situations to facilitate expertise transfer. In: *EKAW'94: a Future for Knowledge Acquisition*, ed. by Steels, L., Schreiber, G., Van de Velde, W., Lecture Notes in Artificial Intelligence, vol. 867 (Springer, Berlin, Heidelberg 1994) pp. 244–261
5. Andrienko, G., Andrienko, N.: Making a GIS intelligent: CommonGIS project view. In: *AGILE'99 Conference*, Rome, April 1999, pp. 19–24
6. Andrienko, N., Andrienko, G.: Intelligent support for geographic data analysis and decision making in the Web. *Journal of Geographic Information and Decision Analysis* **5**(2), 115–128 (2001)
7. Andrienko, G., Andrienko, N., Savinov, A.: Choropleth maps: classification revisited. In: *Proceedings of ICA 2001*, Beijing, Vol.2, pp. 1209–1219 (2001)
8. Andrienko, G., Andrienko, N., Gitis, V.: Interactive maps for visual exploration of grid and vector geodata. *ISPRS Journal of Photogrammetry and Remote Sensing* **57**(5–6), 380–389 (2003)
9. Andrienko, N., Andrienko, G., Voss, H., Bernardo, F., Hipolito, J., Kretchmer, U.: Testing the usability of interactive maps in CommonGIS. *Cartography and Geographic Information Science* **29**(4), 325–342 (2002)
10. Andrienko, G., Andrienko, N.: Computer cartography and cartographic knowledge. In: *Diskussionsbeiträge zur Kartosemiotik und zur Theorie der Kartographie*, Heft 4, ed. by Wolodtschenko, A., Schlichtmann, H. (Selbstverlag der Technischen Universität Dresden, Dresden 2001) pp. 7–14
11. Fairbain, D., Andrienko, G., Andrienko, N., Buziek, G., Dykes, J.: Representation and its relationship with cartographic visualization: a research agenda. *Cartography and Geographic Information Science* **28**(1), 13–28 (2001)
12. Andrienko, G., Andrienko, N., Dykes, J., Gahegan, M., Mountain, D., Noy, P., Roberts, J., Rodgers, P., Theus, M.: Creating instruments for ideation: software approaches to geovisualization. In: *Exploring Geovisualization*, ed. by Dykes, J., MacEachren, A., Kraak, M.-J. (Elsevier, Oxford 2005) pp 103–125
13. Andrienko, N., Andrienko, G., Gatalsky, P.: Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing* **14**(6), 503–541 (2003)

14. Andrienko, N., Andrienko, G., Gatalsky, P.: Impact of data and task characteristics on design of spatio-temporal data visualization tools. In: *Exploring Geovisualization*, ed. by Dykes, J., MacEachren, A., Kraak, M.-J. (Elsevier, Oxford 2005) pp. 201–222
15. Andrienko, N., Andrienko, G., Gatalsky, P.: Supporting visual exploration of object movement. In: *Proceedings of the Working Conference on Advanced Visual Interfaces AVI 2000*, ed. by Di Gesù, V., Levialdi, S., Tarantino L., Palermo, May 2000 (ACM Press, New York 2000) pp. 217–220
16. Andrienko, N., Andrienko, G., Gatalsky, P.: Exploring changes in census time series with interactive dynamic maps and graphics. *Computational Statistics* **16**(3), 417–433 (2001)
17. Andrienko, N., Andrienko, G.: Interactive visual tools to explore spatio-temporal variation. In: (*Ed.*) *Proceedings of the Working Conference on Advanced Visual Interfaces AVI 2004*, ed. by Coastabile, M.F., Gallipoli, May 2004, (ACM Press, New York 2004) pp. 417–420
18. Andrienko, N., Andrienko, G., Gatalsky, P.: Tools for visual comparison of spatial development scenarios. In: *IV 2003. 7th International Conference on Information Visualization, Proceedings*, ed. by Banissi, E., London, July 2003 (IEEE Computer Society, Los Alamitos 2003) pp. 237–244
19. Gatalsky, P., Andrienko, N., Andrienko, G.: Interactive analysis of event data using space–time cube. In: *IV 2004. 8th International Conference on Information Visualization, Proceedings*, ed. by Banissi, E., London, July 2004, (IEEE Computer Society, Los Alamitos 2004) pp. 145–152
20. Andrienko, G., Andrienko, N.: Exploring Spatial Data with Dominant Attribute Map and Parallel Coordinates. *Computers, Environment and Urban Systems* **25**(1), 5–15 (2001)
21. Andrienko, G., Andrienko, N.: Constructing parallel coordinates plot for problem solving. In: *1st International Symposium on Smart Graphics*, ed by Butz, A., Krüger, A., Oliver, P., Zhou, M., Hawthorne, NY, March 2001, (ACM Press, New York 2001) pp. 9–14
22. Andrienko, G., Andrienko, N.: Blending aggregation and selection: adapting parallel coordinates for the visualisation of large datasets. *The Cartographic Journal* **42**(1), 49–60 (2005)
23. Andrienko, N., Andrienko, G.: Cumulative curves for exploration of demographic data: a case study of northwest England. *Computational Statistics* **19**(1), 9–28 (2004)
24. Andrienko, N., Andrienko, G., Savinov, A., Voss, H., Wettschereck, D.: Exploratory analysis of spatial data using interactive maps and data mining. *Cartography and Geographic Information Science* **28**(3), 151–165 (2001)
25. Wrobel, S., Andrienko, G., Andrienko, N., Luthje, A.: Kepler and Descartes. In: *Handbook of Data Mining and Knowledge Discovery*, ed. by Kloesgen, W., Zytkow, J. (Oxford University Press, New York 2002) pp. 576–583
26. Jankowski, P., Andrienko, N., Andrienko: G., Map-centered exploratory approach to multiple criteria spatial decision making. *International Journal of Geographical Information Science* **15**(2), 101–127 (2001)

- 
27. Andrienko, G., Andrienko, N., Jankowski, P.: Building spatial decision support tools for individuals and groups. *Journal of Decision Systems* **12**(2), 193–208 (2003)
  28. Andrienko, N., Andrienko, G.: Informed spatial decisions through coordinated views. *Information Visualization* **2**(4), 270–285 (2003)
  29. Chertov, O., Komarov, A., Andrienko, G., Andrienko, N., Gatalsky, P.: Integrating forest simulation models and spatial-temporal interactive visualisation for decision making at landscape level. *Ecological Modelling* **148**(1), 47–65 (2002)
  30. Chertov, O., Komarov, A., Mikhailov, A., Andrienko, G., Andrienko, N., Gatalsky, P.: Geovisualization of forest simulation modelling results: a case study of carbon sequestration and biodiversity. *Computers and Electronics in Agriculture* **49**(1), 175–191 (2005)
  31. Schuck, A., Andrienko, G., Andrienko, N., Folving, S., Kohl, M., Miina, S., Paivinen, R., Richards, T., Voss, H.: The European Forest Information System – an Internet based interface between information providers and the user community. *Computers and Electronics in Agriculture* **47**(3), 185–206 (2005)
  32. Gitis, V., Andrienko, G., Andrienko, N.: Exploration of seismological information in analytical Web-GIS. *Izvestya Physics of the Solid Earth* **40**(3), 216–225 (2004)
  33. Andrienko, G., Andrienko, N., Voss, H., Carter, J.: Internet mapping for dissemination of statistical information. *Computers, Environment and Urban Systems* **23**(6), 425–441 (1999)

## **Appendix III: Tools for Visual Analysis of Spatio-Temporal Data Developed at the AIS Fraunhofer Institute**

In this appendix, we present the history of the development of CommonGIS – a software system for interactive visual analysis of spatially and temporally referenced data – and give an overview of its functionality.

The roots of our approach to building interactive systems for visual data analysis originate from the software system IRIS (Information Retrieval Intelligent System), which was developed for Windows in the early 1990s (Andrienko and Andrienko, 1997). IRIS was implemented in C++. IRIS realised several innovative ideas:

1. The concept of interactive maps that change their appearance in real time upon activation of interactive manipulators by the user.
2. A knowledge-based approach to the automated selection of map symbolism depending on the characteristics of the data and the user's needs.

The development of IRIS was continued by applying the Java programming language and environment designed for the Internet. IRIS, renamed Descartes, became one of the first interactive mapping systems available on the Internet (Andrienko and Andrienko, 1999). As early as September 1996, it was included in the list of the Top 1% Web applets and top ten Web applets by the independent Java Applet Rating Service (<http://www.jars.com/>). In Descartes, we implemented dynamic linking between maps and statistical graphic displays (brushing).

In 1998–2001, further development continued within the framework of ESPRIT Project 28983 called CommonGIS (Andrienko et al. 2003), which was proposed and coordinated by AIS. In the course of the project, the software was renamed CommonGIS.

CommonGIS is unique among both commercial and research software systems as being composed of well-integrated tools, which can complement and enhance each other, thus allowing sophisticated analyses. The system includes various methods for cartographic visualisation; non-spatial graphs; tools for querying, search, and classification; and computa-

enhanced visual techniques. A common feature of all the tools is their high user interactivity, which is essential for exploratory data analysis.

The main features of CommonGIS are the following:

1. A variety of interactive mapping techniques, statistical graphic displays, and computational methods.
2. Comprehensive tools for analysis of spatial time series, including animated maps, and time-aware map visualisations.
3. Novel information visualisation tools (dynamic query tools, a table lens, parallel-coordinates plots, etc.).
4. Tools for interactive multicriteria decision-making and sensitivity analysis for individuals and small groups of decision makers, supporting various styles of and procedures for informed decision-making.
5. A possibility to complement interactive visual data analysis with mathematical methods of statistics and data mining.
6. A prototype of intelligent user guidance that helps users to follow problem-solving scenarios and utilise all tools for selected data-analysis and decision-making problems.
7. Space–time cube display for analysis of spatio-temporal events.
8. Tools for interactive aggregation of raster data, tightly coupled to dynamic visualisation of the results.

The system integrates all visualisation techniques via multiple mechanisms of coordination and linking: dynamic highlighting and selection, queries, synchronised zooming etc.

A commercial version of the CommonGIS software has been released by SPADE, the spatial decision support department of the AIS Fraunhofer Institute; see [www.commongis.com](http://www.commongis.com) for details. Universities and schools can order free licences from the same site for research and educational use.

## References

1. Andrienko, G., Andrienko, N.: Intelligent cartographic visualization for supporting data exploration in the IRIS system. *Programming and Computer Software* **23**(5), 268–282 (1997)
2. Andrienko, G., Andrienko, N.: IRIS: a tool to support data analysis with maps. In: *Interoperating Geographic Information Systems*, ed. by Goodchild, M., Egenhofer, M., Fegeas, R., Kottman, C. (Kluwer, Boston 1999) pp. 221–234
3. Andrienko, G., Andrienko, N., Voss, H.: GIS for everyone: the CommonGIS project and beyond. In: *Maps and the Internet*, ed. by Peterson, M. (Elsevier, Amsterdam 2003) pp. 131–146

# Index

abstraction, 135, 144, 168, 260, 293, 337, 452, 481, 486, 506, 514, 591, 596  
aggregation, 11, 29, 96, 97, 111, 164, 293, 333, 335, 372, 398, 401, 424, 430, 436, 486, 499, 507, 511, 520, 543, 558, 574, 582, 591, 598, 606, 614, 658  
reaggregation, 327, 334, 434, 435, 487, 608, 626  
animation, 163, 179, 292, 485, 506, 528, 562, 600  
animated display, 438  
animated map, map animation, 106, 199, 265, 291, 438, 488, 531, 562, 658  
Arnheim, Rudolf, 144, 149, 151, 167, 172, 452, 480, 483, 495, 506, 518, 632  
arrangement (of the display space), 183, 185, 648  
juxtaposition, space partitioning, 189, 196, 309, 482, 485, 497, 600, 648  
overlay, space sharing, 189, 190, 197, 491, 497, 523, 600, 648  
space embedding, 186, 189, 191, 482, 488, 600, 648  
space transformation, 189, 484, 649  
space behaviour, *see* behaviour, aspects of a behaviour  
attribute (characteristic component), 7, 17, 18, 640  
Attribute Explorer, 359, 368, 373, 559

bagplot, 321, 558  
bar chart, 191, 205, 207, 226, 233, 242, 253, 412, 439, 449, 496, 621  
behaviour, 8, 48, 86, 91, 99, 643  
aspects of a behaviour, aspectual behaviour, 9, 103, 111, 125, 135, 150, 462, 466, 472, 482, 491, 552, 574, 581, 596, 598, 615, 645  
local behaviour, 9, 85, 473, 491, 536, 572, 621  
mutual (relational) behaviour, 126, 128, 159, 557  
overall behaviour, 8, 103, 111, 136, 150, 152, 202, 219, 252, 461, 466, 468, 472, 482, 486, 510, 552, 574, 581, 584, 593, 596, 603, 606, 611, 615, 631, 647  
slices of a behaviour, 105, 492, 529, 553, 581, 599, 618, 645  
behaviour characterisation tasks, 85, 86, 107, 109, 131, 134, 150, 157, 297, 464, 465, 470, 510, 586, 647  
Bertin, Jacques, 10, 33, 47, 49, 51, 60, 81, 85, 120, 152, 171, 183, 189, 198, 207, 214, 226, 480, 482, 492, 496, 506, 513, 632  
box plot, *see* box-and-whiskers plot  
box-and-whiskers plot, 96, 317, 398, 411, 442, 520  
bivariate box plot, 321  
brushing, 368, 373, 394, 435, 517, 574, 657, 658  
cartographic research method, 169

- change map, 263  
characteristic component, *see* attribute  
characteristic set, 22, 639, 640  
choropleth map, 257, 567  
  classified, 217, 252, 381  
  unclassified, 217, 248, 264, 373, 503, 526  
classification, 157, 173, 219, 258, 415, 430, 436, 449, 498, 507, 511, 520, 556, 582, 588, 603, 657  
cross-classification, 227, 513, 523  
dominance classification, 224, 490, 513, 652  
classification tree, 402, 415, 423, 430  
clustering, 157, 407, 424, 426, 430, 436, 505, 517, 520, 556, 603, 622  
cognitive psychology, 144, 461, 494, 632  
comparable attributes, 335, 444, 513, 526, 618  
  making attributes comparable, *see* standardisation of attribute values  
comparison tasks, 73, 74, 78, 79, 108, 115, 124, 127, 142, 147, 151, 154, 341, 351, 394, 468, 521, 546, 647  
  direct comparison tasks, 67, 75, 79, 124, 150, 394, 589  
  inverse comparison tasks, 67, 75, 78, 79, 112, 124, 150, 589  
compound tasks, 63, 74, 78, 109, 119, 142  
conditioned maps, 362  
connection discovery tasks, 10, 127, 134, 141, 150, 157, 174, 466, 468, 584, 586  
constraint (of a task), 47, 53, 60, 61, 70, 73, 77, 80, 86, 116, 120, 139, 154, 158, 336, 342, 481, 585, 646  
correlation, 126, 128, 132, 134, 141, 143, 173, 208, 222, 292, 314, 356, 373, 381, 399, 430, 435, 477, 520, 553, 591, 603, 624  
correlation coefficient, 399  
spatial autocorrelation, 292  
cumulative curve, 316, 319, 334, 543, 652  
cumulative frequency curve, 303, 313, 525  
enhanced cumulative curve, 313, 513  
data, 17, 18, 27, 639, 641  
data analysis tasks, 6, 8, 17, 22, 47, 63, 139, 393, 584, 646  
data cube, 27, 30  
data function, 21, 22, 61, 72, 75, 78, 119, 139, 149, 152, 552, 641, 643, 646  
data mining, 14, 85, 157, 164, 383, 396, 397, 401, 406, 407, 415, 423, 426, 430, 508, 517, 536, 558, 583, 652, 658  
data, properties of, 27, 31  
continuity, 27, 29, 30, 31, 187, 190, 288, 293, 465  
distances, 27, 28, 62, 66, 89, 93, 97, 111, 140, 186, 187, 190, 272, 288, 294, 298, 303, 312, 343, 373, 465, 483, 529, 585, 587, 607, 643, 644, 647  
ordering, 27, 61, 66, 89, 93, 111, 140, 186, 190, 291, 294, 312, 465, 483, 547, 585, 587, 642  
smoothness, 28, 30, 289, 465, 644  
data, structure of, 6, 18, 27, 34, 120, 154, 465, 510, 608, 639  
  functional view, 21, 29  
decile, *see* positional measures  
decision making, 281, 652, 658  
decomposition, 466, 554, 581, 587, 631  
depth (a display dimension), 178, 183, 197, 484, 492, 496, 648  
dimensionality reduction, 488, 516, 528, 543, 563, 599, 607  
discretising of a continuous set, 29, 584  
dispersion graph, 368

- display coordination, 448, 452, 548, 556, 600, 603, 608, 626, 652
- display dimension, 163, 183, 190, 192, 196, 208, 242, 257, 301, 316, 438, 450, 483, 496, 505, 528, 532, 552, 561, 573, 585, 615, 648, 649
- display space, 183, 194, 196, 208, 307, 313, 449, 452, 484, 594, 595, 648
- display time, 163, 183, 196, 441, 484, 505, 600, 615, 648
- integrative display dimensions, 492, 496
- display linking, *see* display coordination
- display multiplication, 442, 449, 515, 519, 520, 561, 600
- distance
- Euclidean, 62, 374, 383
  - in space, 374, 431, 535, 562, 621
  - Manhattan, 375
  - Minkowski, 375
- distribution
- (statistically) normal distribution, 96, 263, 278, 333, 398
  - (statistically) skewed distribution, 96, 244, 258, 333, 334, 359, 398, 559, 644
  - spatial distribution, 86, 89, 96, 100, 149, 195, 219, 250, 258, 264, 327, 404, 473, 503, 523, 530, 615, 645
  - statistical distribution, 95, 244, 263, 305, 381, 525, 606, 608, 614
  - temporal distribution, *see* temporal variation
- diverging, or double-ended, colour scale, 249, 264, 330, 527, 539
- midpoint, 249, 527
- domain knowledge, 125, 133, 231, 462, 471, 508, 510, 531, 547, 554, 579, 586, 608, 611, 630, 635
- domain of values, *see* value domain
- dot plot, 236, 245, 261, 353, 366, 398, 439, 520
- double-ended colour scale, *see* diverging, or double-ended, colour scale
- drilling, drill up, drill down, 4, 30, 332, 509, 543
- dynamic attribute, 277, 279, 281
- Dynamic Query, 338, 346, 350, 353, 368, 373, 385, 435
- dynamic tools, 396, 433
- dynamic query tools, 12, 164, 351, 352, 355, 368, 373, 381, 389, 396, 433, 436, 464, 517, 524, 559, 583, 599, 606
- elementary level of analysis, *see* levels of analysis, elementary
- elementary tasks, 8, 9, 47, 60, 61, 75, 86, 107, 112, 113, 115, 119, 151, 158, 242, 257, 293, 382, 464, 468, 479, 482, 486, 546, 585, 646
- equivalence class, 29, 294, 584, 607, 614
- Euclidean distance, *see* distance, Euclidean
- filtering, 352, 353, 359, 368, 371, 378, 381, 383, 394, 412, 418, 424, 435, 449, 508, 519, 521, 532, 541, 559, 574, 606
- focusing, 82, 234, 244, 249, 258, 261, 312, 430, 436, 448, 507, 517, 519, 540, 559, 582, 588, 599, 606, 616
- focus plus context, 541
- frequency histogram, *see* histogram
- generalisation, 157, 258, 260, 337, 507
- map generalisation, 214, 588
- geographical objects (features), 188, 195, 196, 232
- geographical space, *see* space, geographical
- gestalt principles, 90, 174, 494, 506, 514, 530, 632

- graduated circles, 101, 249, 261, 268  
Green, Mark, 174, 183, 206
- heterogeneity, 96, 195, 512  
of geographical space, 195, 196
- histogram, 300, 312, 316, 334, 353, 363, 372, 392, 398, 408, 435, 505, 520, 525, 543, 556, 559, 608, 652
- holistic visualisation, holistic perception, *see* unification (perceptual)
- image, 50, 86, 163, 173, 183, 197, 207, 214, 228, 258, 472, 482, 485, 492, 495, 515  
image theory (Bertin), 173, 496, 632  
mental image, 85, 166, 252, 448, 503, 599, 644
- InfoCrystal, 372
- information loss, information reduction, 216, 226, 250, 258, 276, 293, 332, 483, 486, 498, 507, 591, 603, 606
- Information Seeking Mantra, 4, 15, 146, 149, 483, 540, 541, 632
- interpolation, 30, 154, 164, 288, 335, 401, 430, 582
- invariance, reference-invariant, 552, 588, 632
- isomorphism principle, 168, 187, 195
- Klir, George, 3, 4, 19, 27, 49, 152, 480, 632
- levels of data analysis  
elementary, 81, 115, 125, 219, 464, 521, 646  
synoptic, 48, 115, 120, 125, 382, 521, 647
- levels of measurement (of attributes), 33  
interval, 33, 172, 299, 312
- nominal (qualitative), 32, 33, 172, 190, 207, 276, 294, 299, 300, 333, 375, 430, 444, 513, 526, 559, 637
- ordinal, 32, 33, 172, 299, 333, 637
- quantitative, or numeric (combined interval and ratio), 19, 27, 32, 33, 62, 172, 190, 295, 298, 333, 375, 444, 513, 526, 558, 637
- ratio, 33
- levels of reading (Bertin), 47, 49, 52, 81, 120, 154, 171, 173, 202, 226, 483
- linked displays, *see* display coordination
- lookup tasks, 61, 73, 78, 113, 115, 140, 151, 341, 351, 468  
direct lookup tasks, 61, 66, 74, 78, 107, 342, 394, 464, 546, 589  
inverse lookup tasks, 61, 67, 74, 78, 79, 107, 113, 342, 394, 546, 589
- MacEachren, Alan, 153, 178, 183, 191, 197, 504
- Manhattan distance, *see* distance, Manhattan
- map animation, *see* animation, animated map
- map as a model of the world, 169
- map layers, 217, 441, 524
- marking (of display items), 352, 359, 363, 371, 389, 394, 435, 437, 449, 541, 548, 599, 608  
multicolour marking, 352, 368, 369, 372, 377, 381, 408, 424, 436, 449, 561, 600
- marks (in visualisation), 163, 171, 182, 190, 194, 197, 207, 231, 253, 257, 302, 321, 331, 483, 496, 498, 507, 593, 595, 602, 606, 615, 648, 649
- mean, *see* statistical mean

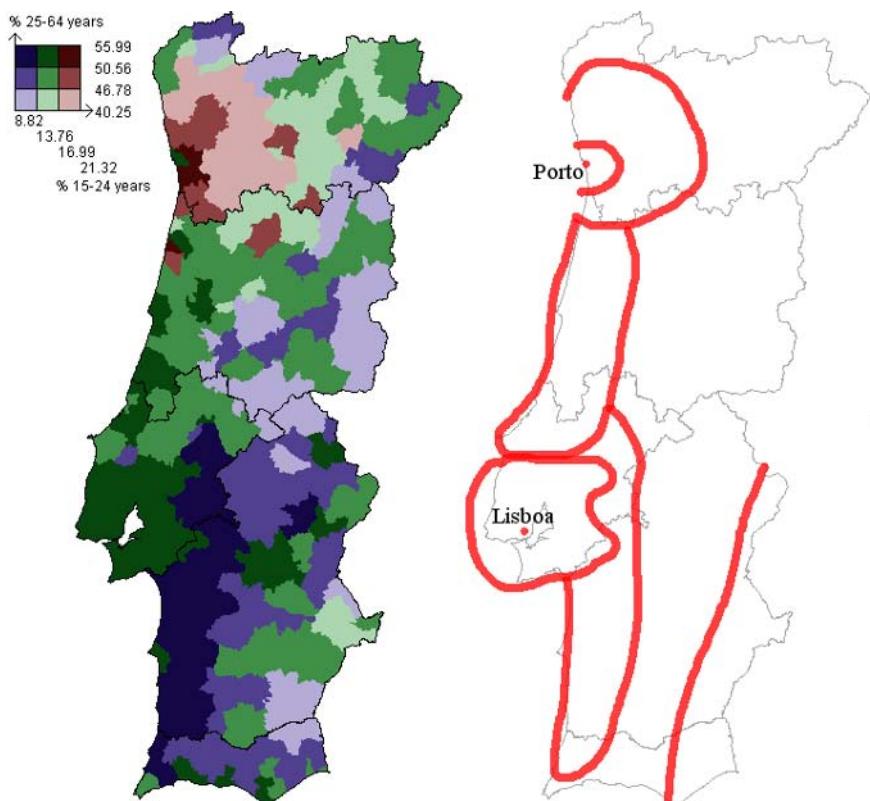
- median, 96, 263, 274, 299, 316, 331, 334, 398, 411, 432, 516, 541, 574, 614  
 metric relations, 195, 343, 348, 373  
 Minkowski distance, *see* distance,  
   Minkowski  
 model, 76, 105  
   behaviour model, 85, 134, 135, 149, 426, 466  
   data model, 26, 75, 139, 158  
   mathematical model, 288  
   mental model, 149, 335, 465, 479  
   simulation model, 132  
   task model, 139, 154, 158, 584  
 mosaic plot, 307, 543  
 multidimensional data, 9, 16, 27, 105, 124, 136, 150, 452, 463, 466, 472, 487, 508, 516, 528, 552, 586, 591, 596, 641, 645
- neighbour, neighbourhood, 89, 93, 97, 129, 146, 158, 182, 218, 250, 259, 269, 288, 335, 349, 403, 430, 468, 531, 544, 557, 582, 596, 603, 622, 644
- Occam's razor (principle of parsimony), 90, 174  
 ogive, *see* cumulative curve  
 OLAP, 332  
 outlier, 91, 93, 95, 97, 151, 157, 210, 235, 263, 274, 299, 321, 333, 334, 399, 432, 475, 486, 507, 544, 559, 587, 596, 616  
   outlier removal, 236, 249, 261, 507, 550, 591, 621
- parallel coordinates, 202, 207, 221, 323, 331, 376, 398, 412, 424, 427, 430, 432, 442, 449, 505, 514, 519, 521, 541, 544, 556, 558, 652, 658  
 parameter (of an analysis or transformation method), 281, 335, 446  
   changing the tool parameters, 406, 408, 427, 433, 448, 451, 507, 517  
   sensitivity to parameters, *see* sensitivity analysis  
   parsimony, principle of, *see* Occam's razor  
 pattern, 8, 85, 86, 91, 644  
   arrangement pattern, 91, 94, 131, 135  
   association pattern, 91, 97, 131, 135, 219, 539  
   compound pattern, 88, 96, 97, 510, 554, 600, 644  
   differentiation pattern, 91, 93, 131, 135  
   distribution summary pattern, 91, 95, 131, 135, 588  
   pattern comparison tasks, 138, 468, 510, 603  
   pattern definition tasks, *see* behaviour characterisation tasks  
   pattern search tasks, 107, 113, 135, 138, 150, 157, 394, 467, 510, 530, 536  
   percentile, *see* positional measures  
   perception, 90, 146, 167, 173, 452, 481, 483, 494  
   perceptual integration, *see* unification (perceptual)  
   permutation (in a matrix or table), 208, 505, 513  
   Peuquet, Donna, 32, 153  
   phenomenon, 19, 25  
     abrupt, 29, 31, 644  
     continuous, 28, 32, 582, 637  
     discrete, 28  
     smooth, 28, 30, 31, 288, 644  
   pie chart, 239, 302, 313, 331, 340, 445, 449, 488, 497, 577  
   population (statistical population),  
     type of referrer, *see* statistical population  
   positional measures, 299, 316, 321, 323, 332, 334, 336, 398, 486, 507

- quartile, *see* positional measures
- query language, 12, 337, 349, 394, 464, 583, 599
- spatial query language, 343
- SQL, 337, 343
- visual query language, 337, 344, 350, 394
- raster data model, 32, 272, 292, 296, 330, 423, 523, 532, 651
- redundant use of visual variables, 180, 191, 199
- reference (in data), reference set, 7, 16, 17, 18, 19, 639, 640, 641, 643
- referrer, referential component, 16, 17, 18, 22, 640
- types of referrers, 17, 25, 26, 28, 640
- regionalisation, 214, 219, 512
- relational tasks, 62, 141
- relation-seeking tasks, 64, 69, 78, 80, 115, 135, 140, 150, 341, 351, 394, 468, 469, 510, 546, 589
- residuals, 274, 593
- retinal variables, 163, 171, 182, 190, 193, 196, 213, 217, 258, 484, 496, 552, 649
- robustness analysis, *see* sensitivity analysis
- Salichtchev, Konstantin, 169, 181
- sample, 400, 426
- sampling, 29, 478
- sample locations, 32, 291, 582, 637
- scatterplot, 126, 145, 265, 330, 353, 363, 398, 427, 432, 435, 514, 520, 542, 544, 558, 603, 624
- binned scatterplot, 302, 542, 558
- scatterplot matrix, 427, 432, 445, 558, 603
- segmented bars, 302, 313, 316, 409, 497, 564
- sensitivity analysis, 281, 335, 359, 407, 427, 658
- Shneiderman, Ben, 4, 15, 146, 156, 338, 383, 480, 483, 540, 547, 632
- similarity, 374, 383
- grouping by similarity, 378, 407, 427, 505, 507, 510, 568, 621
- measuring (dis)similarity, 109, 380, 394
- simplification, 89, 174, 207, 214, 217, 250, 257, 260, 270, 293, 338, 486, 506, 514, 571, 580, 588, 591, 596, 623
- slices of a behaviour, *see* behaviour, slices of a behaviour
- slider, slider bar, slider line, 250, 280, 338, 344, 346, 350, 353, 368, 394
- small multiples, 185, 197, 199, 265, 491, 497, 522, 562, 616
- smoothing, 214, 215, 258, 270, 335, 387, 401, 445, 506, 588, 591, 596
- space
- absolute and relative view of space, 26
  - as a type of referrer, 18, 19, 341, 487, 496, 529, 530, 555, 576, 623, 640
  - as an attribute, 20, 26, 44, 342, 529, 530, 561, 607, 614
  - dual treatment of space, 26, 113
  - geographical, 29, 62, 89, 116, 188, 194, 196, 374, 532, 555, 561, 577, 595, 614, 643
- space–time cube, 196, 532, 615, 628, 658
- spatial relations, 113, 343, 403
- standard deviation, 263, 298, 331, 398, 411, 516
- standard normal transformation, *see* z-score
- standardisation of attribute values, 263, 375, 401
- statistical mean, 92, 96, 97, 263, 270, 298, 329, 332, 334, 336, 398, 411, 424, 432, 486, 507, 516, 588, 618

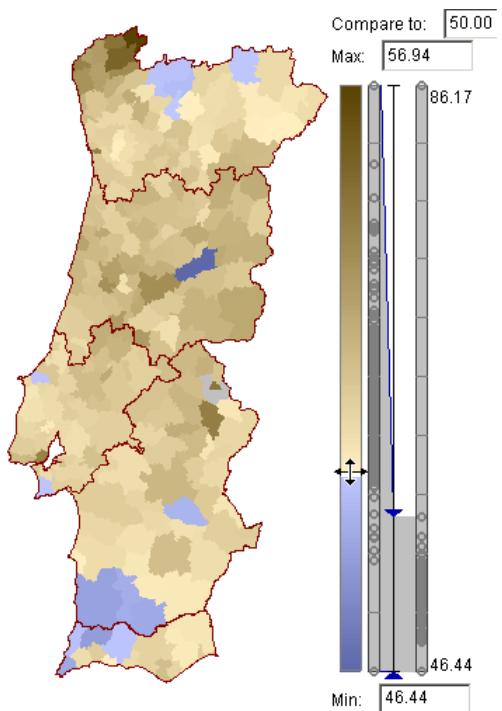
- statistical population (type of referrer), 18, 19, 26, 27, 86, 89, 186, 202, 327, 392, 400, 491, 512, 525, 531, 558, 607, 613, 637, 640, 644
- statistics, 1, 3, 14, 85, 95, 132, 164, 214, 274, 303, 359, 397, 403, 406, 426, 430, 558, 572, 588, 658
- descriptive statistics, 398, 426, 430, 508
- inferential statistics, 398, 426
- summary statistics, 411, 432, 583, 607, 614
- subtask, 63, 74, 75, 119, 135, 142, 150, 157, 158, 461, 466, 510, 584, 586, 591, 611, 631
- synoptic level of analysis, *see* levels of analysis, synoptic
- synoptic tasks, 8, 9, 47, 61, 81, 119, 127, 134, 135, 141, 157, 158, 173, 219, 257, 293, 381, 465, 469, 482, 486, 494, 506, 508, 510, 557, 585, 646
- target (of a task), 47, 53, 57, 61, 73, 77, 86, 109, 120, 139, 154, 158, 336, 344, 481, 585, 646
- task (of data analysis), *see* data analysis tasks
- temporal relations, 113, 346
- temporal variation, 100, 258, 464, 531, 574, 598, 607, 615, 623
- time
- absolute and relative view of time, 26
  - as a type of referrer, 19, 89, 195, 196, 213, 346, 484, 555, 573, 598, 619, 640, 645
  - as an attribute, 20, 25, 26, 44, 229, 295, 346, 513, 561, 607, 614
  - cyclic and linear time, 32, 213, 347, 462, 512, 555, 572, 580, 590, 598
- time graph, 82, 99, 166, 186, 213, 214, 241, 266, 270, 321, 340, 382, 393, 398, 436, 467, 482, 491, 496, 507, 521, 530, 541, 544, 561, 562, 574, 600, 622
- time series, 154, 181, 263, 323, 381, 382, 400, 427, 436, 536, 652
- Time Wheel, 347, 350, 517, 574, 599
- TimeSearcher, 383
- treemap, 307, 313, 543
- trend, 8, 48, 50, 82, 86, 91, 94, 97, 107, 113, 131, 133, 150, 154, 157, 166, 214, 219, 236, 273, 293, 322, 383, 400, 468, 530, 537, 576, 579, 644
- triad model of spatio-temporal data, 153
- Tufte, Edward R., 181, 185, 497
- Tukey, John, 3, 5, 96, 148, 316, 450
- unification (perceptual), 257, 463, 472, 482, 485, 494, 508, 522, 552, 573, 585, 602, 615
- value (of a data component), 7, 18, 639, 640, 641, 642, 649
- value domain, 19, 21, 59, 66, 69, 110, 190, 196, 584, 607, 614, 640, 642, 643
- vector data model, 31
- Venn diagram, 372
- visual comparison (display manipulation technique), 248, 259, 261, 330, 374, 443, 474, 507, 514, 526, 539, 582
- reference value, 249, 259, 262, 330, 514, 526, 582
- visual differentiation, 191, 194, 197
- visual encoding function, 164, 242, 250, 257, 429, 441, 526, 541, 559, 588, 649
- linear, 244, 258, 559
- logarithmic, 244, 258, 559
- non-linear, 244, 259
- parameterised, 247, 649
- visual linking, 191, 194, 197, 311, 448

- visual variables, 163, 171  
perceptual properties, 171, 180
- visualisation, 163, 166, 647  
basic principles, 10, 189, 316  
cartographic, 181, 194, 204, 330
- visualisation-based research  
method, 169
- weighted linear combination,  
weighted sum, 279, 280, 335, 433
- weighting, weights, 272, 279, 280,  
290, 298, 311, 335, 433
- Weka, 407, 408, 415
- Wilkinson, Leland, 182, 185, 321,  
333
- zooming, 82, 221, 231, 258, 312,  
430, 436, 449, 532, 540, 550, 582,  
606, 626, 658
- z-score, 263, 278, 280, 527, 618

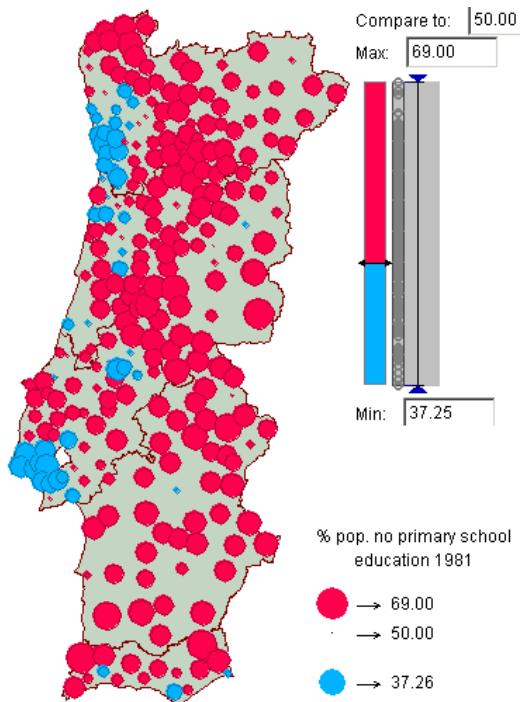
## Colour Plates



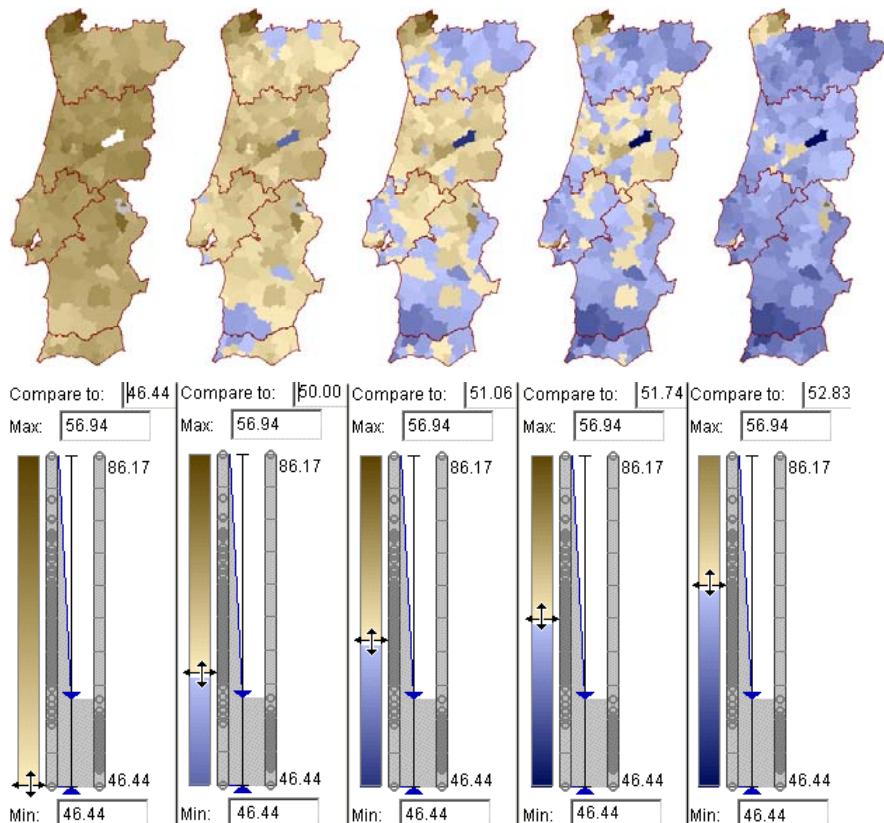
**Fig. 4.23C.** Cross-classification of the districts of Portugal according to the values of the attributes “% 25–64 years” and “% 15–24 years”. The division into classes and the assignment of colours to the classes are schematically shown in the top left corner. To divide the value range of each attribute into three subintervals, the algorithm for statistically optimal classification was applied (Sect. 4.4.3)



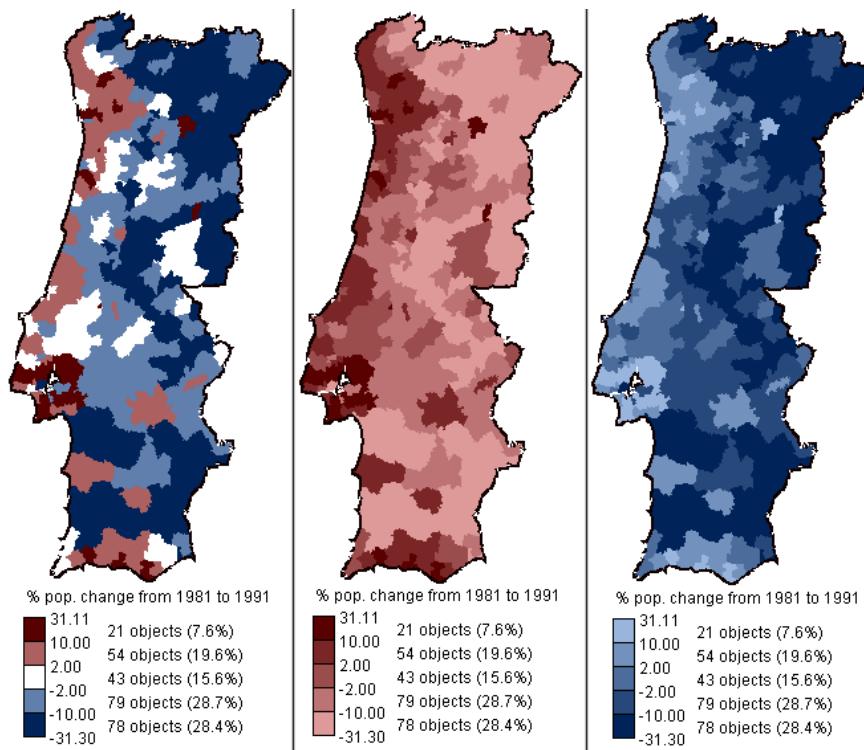
**Fig. 4.34C.** Visual comparison in an unclassified choropleth map. The values of the attribute “% female 1991” in the districts of Portugal are compared with 50%: values below 50 are represented by shades of blue, and values over 50 – by shades of brown. The outlier 86.17 has been excluded from the visualisation by means of focusing (Sect. 4.4.6)



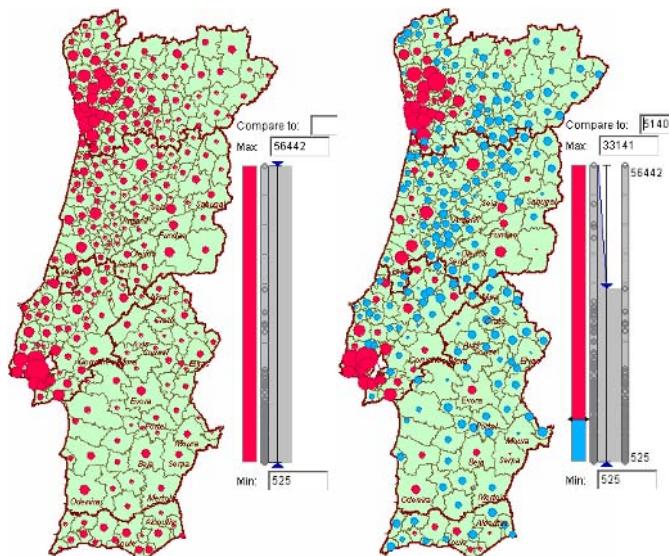
**Fig. 4.35C.** Visual comparison on a map using graduated circles. The circles represent the values of the attribute “% pop. no primary school education 1981” compared with 50%: the size (area) of a circle is proportional to the difference between the corresponding attribute value and 50, a cyan colour of the circle indicates that the value is less than 50, and a red colour indicates that the value is greater than 50 (Sect. 4.4.6)



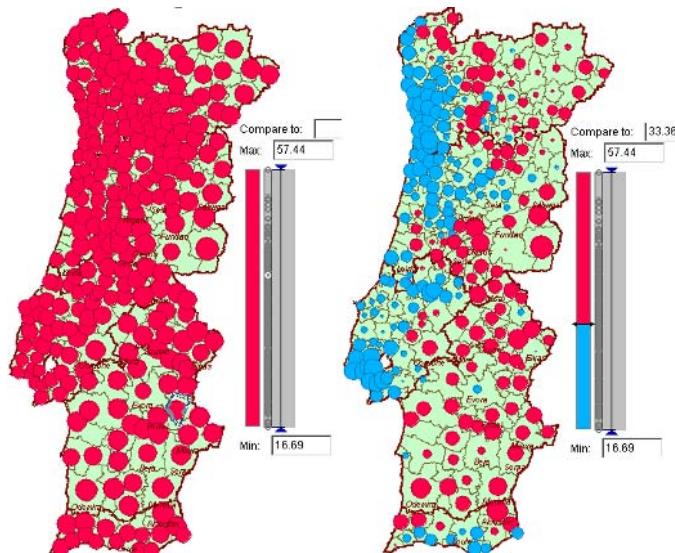
**Fig. 4.36C.** By gradually changing the reference value (midpoint) of a diverging colour scale, one can observe various spatial patterns formed by the visual association of districts coloured in the same hue (Sect. 4.4.6)



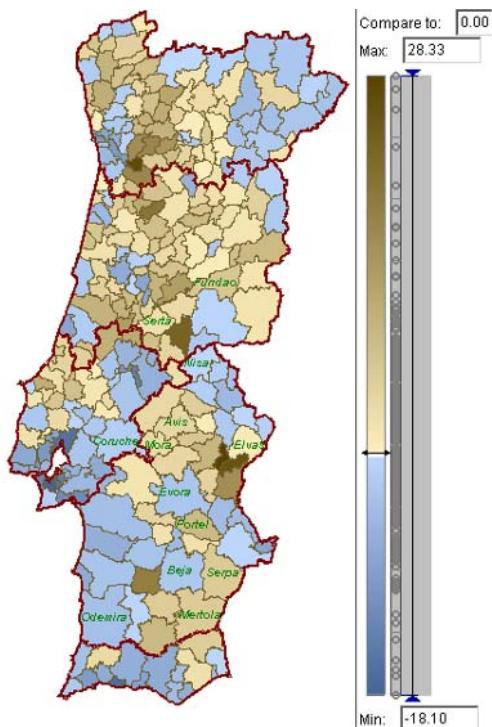
**Fig. 4.38C.** Left: a classified choropleth map with a diverging colour scale. Centre and right: the same classification is represented by single-hue colour scales with increasing and decreasing darkness, respectively (Sect. 4.4.6)



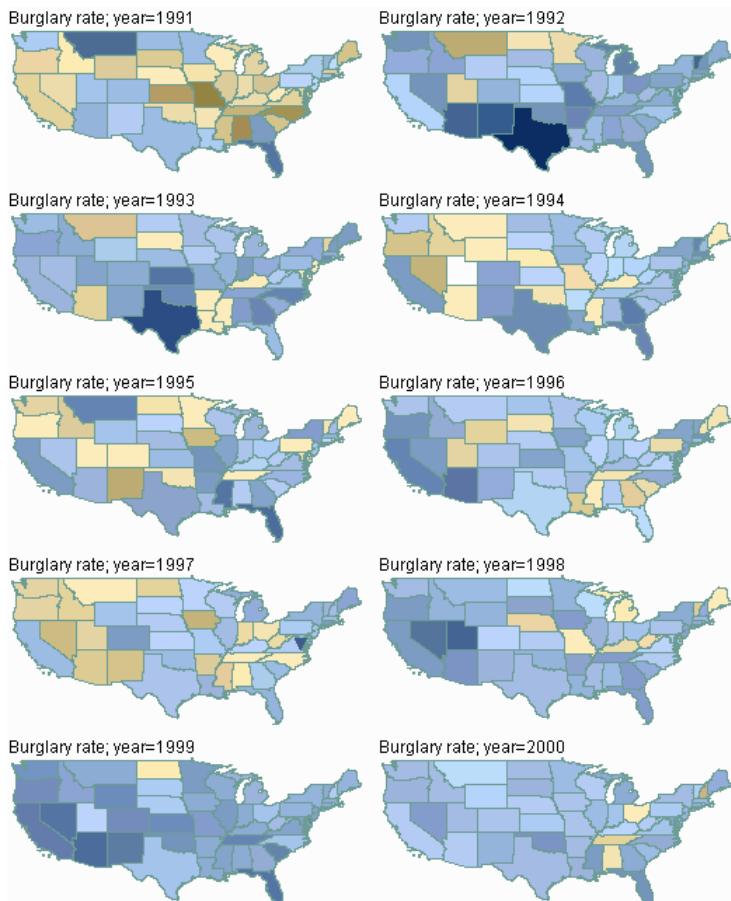
**Fig. 4.41C.** The distribution of the absolute numbers of people without primary school education over Portugal. Left, original view; right, after removing the outlier 56 442 and applying visual comparison with the country mean, 5140. The blue circles correspond to values below the country mean (Sect. 4.5.1.1)



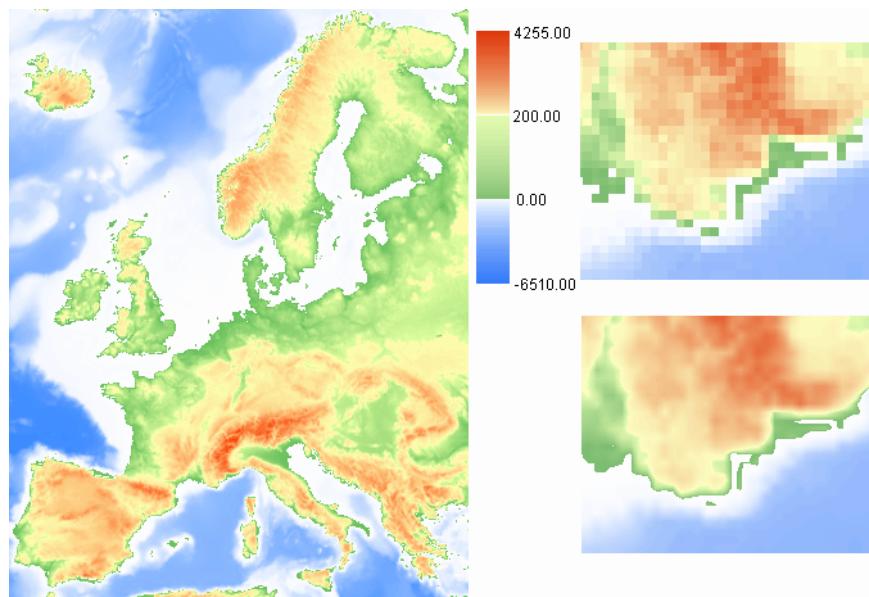
**Fig. 4.42C.** The distribution of the percentage of people without primary school education in the entire population of a district, over Portugal. Left, original view; right, after applying visual comparison with the country mean, 33.36%. The blue circles correspond to values lower than the mean (Sect. 4.5.1.1)



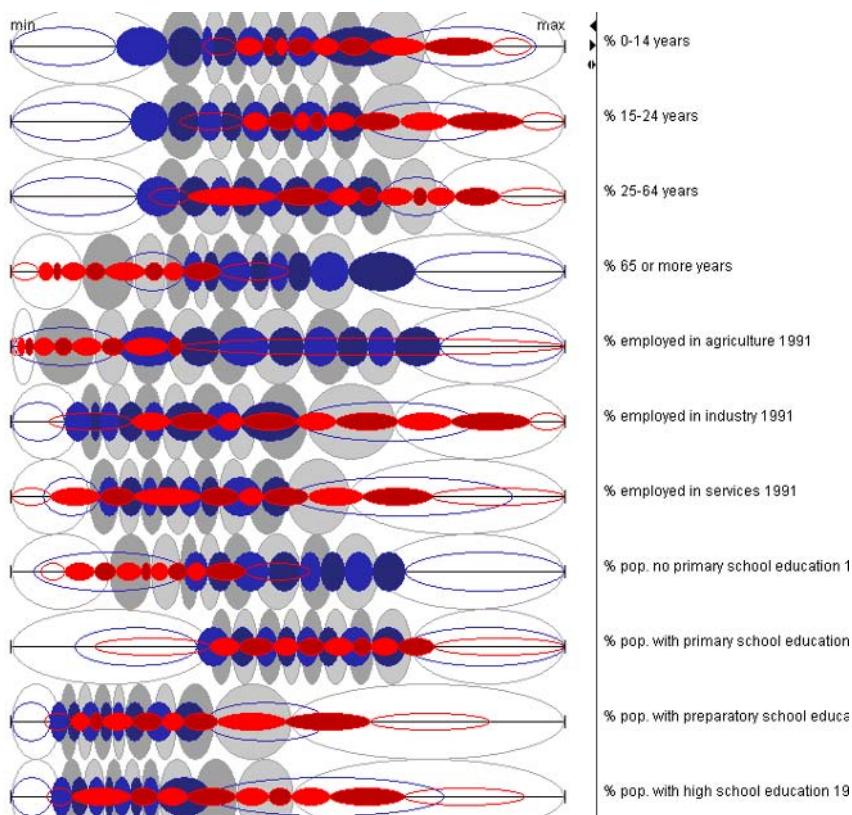
**Fig. 4.43C.** A change map representing the changes in the proportions of people employed in industry from the year 1981 to the year 1991. Positive changes (i.e. increases) are encoded by shades of brown, and negative changes (i.e. decreases) by shades of blue (Sect. 4.5.1.2)



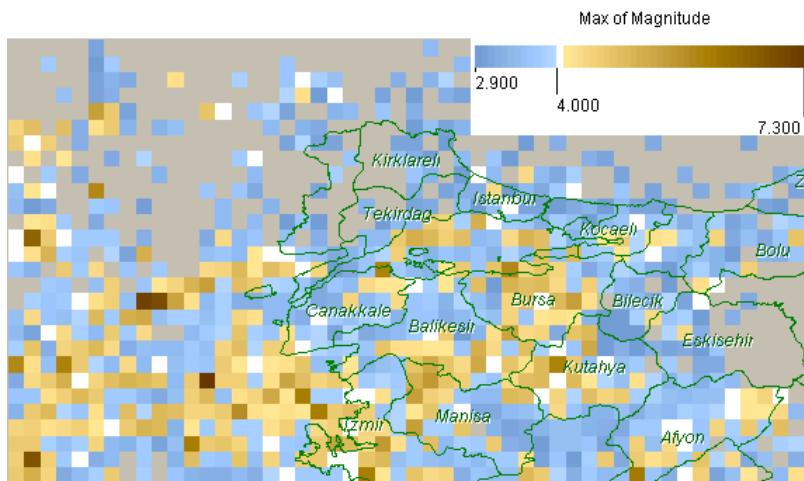
**Fig. 4.46C.** A series of change maps showing how the burglary rates changed from year to year over the states of the USA during the period from 1991 to 2000. Each map represents the differences between the burglary rates in the year indicated above the map and in the previous year. Brown shades encode positive differences, i.e. an increase in the burglary rate, and blue shades correspond to negative differences, i.e. a decrease in the burglary rate (Sect. 4.5.1.2)



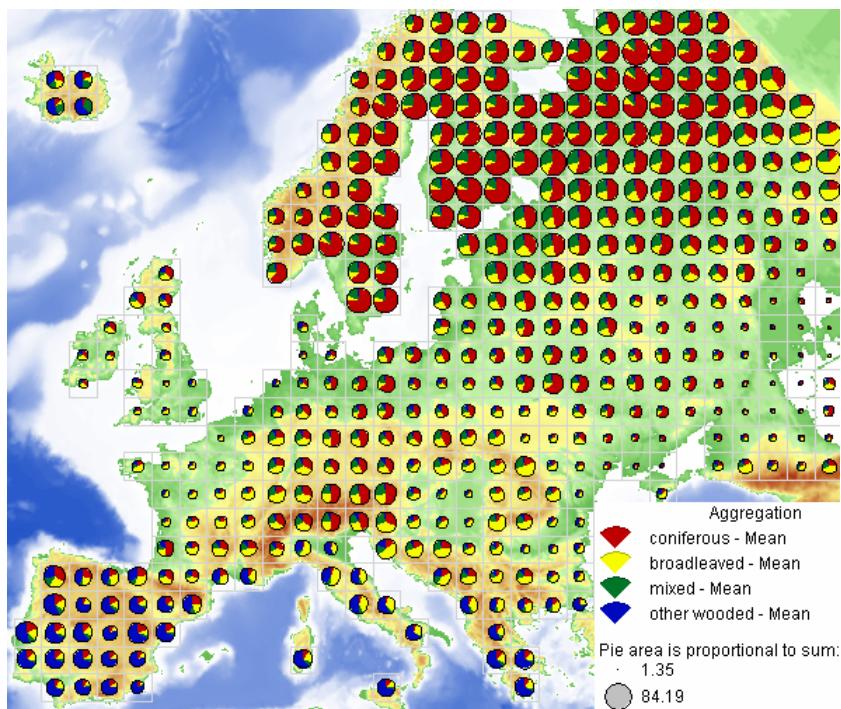
**Fig. 4.60C.** Visualisation of data specified in a raster format. On the left, the relief of Europe is visualised by encoding altitudes by colours of screen pixels. On the right, the effect of interpolation is demonstrated. The image fragment at the top right has been produced without interpolation. At the bottom right, the same data are visualised using linear interpolation (Sect. 4.5.3)



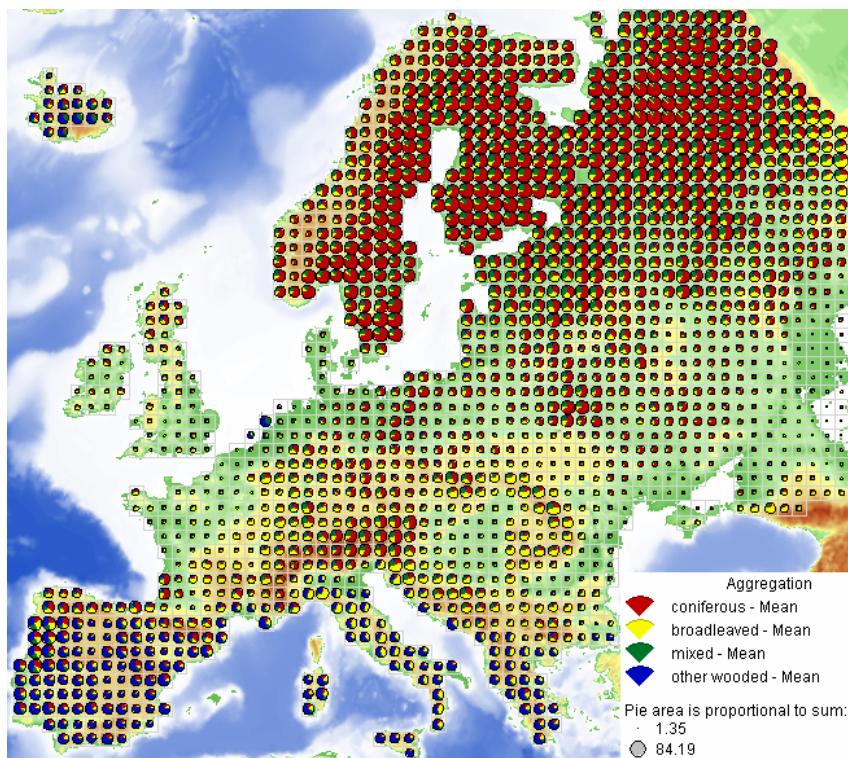
**Fig. 4.79C.** Comparison of aggregate characteristics of three sets: the set of all districts of Portugal (grey), the set of districts with a population decrease (population change between  $-31.3$  and  $-3\%$ ; blue), and the set of districts with a population increase (population change between  $3$  and  $31.11\%$ ; red) (Sect. 4.5.4.5)



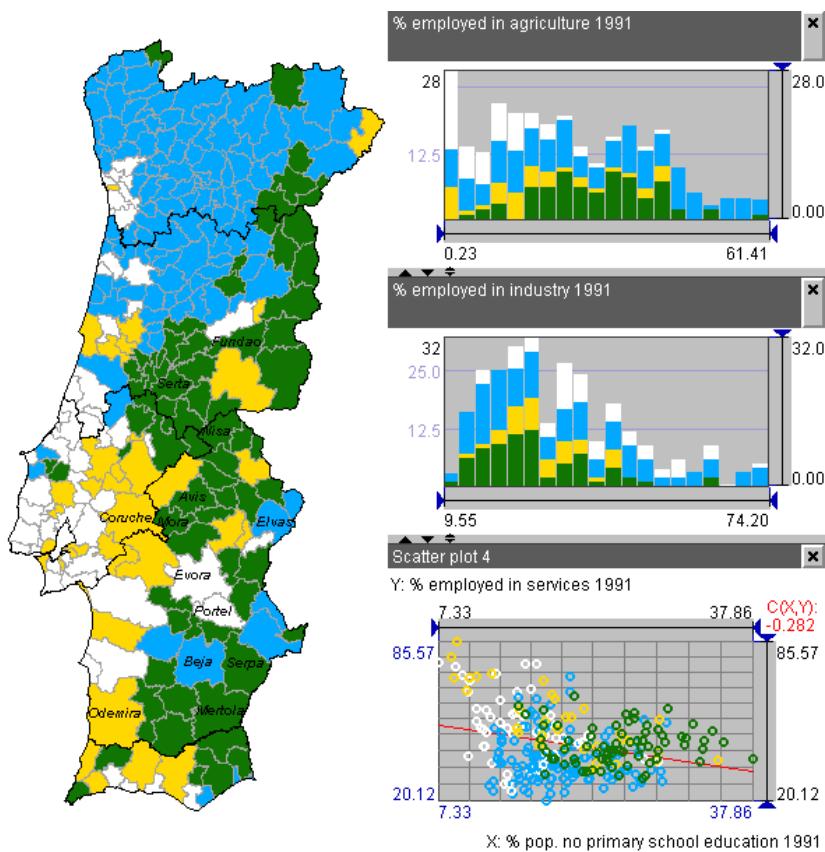
**Fig. 4.85C.** Application of a visual comparison technique to the visualisation of the maximum earthquake magnitudes. Blue shades represent values below 4, and brown shades values over 4. Cells where no earthquake occurrences were recorded are shaded in grey (Sect. 4.5.4.6)



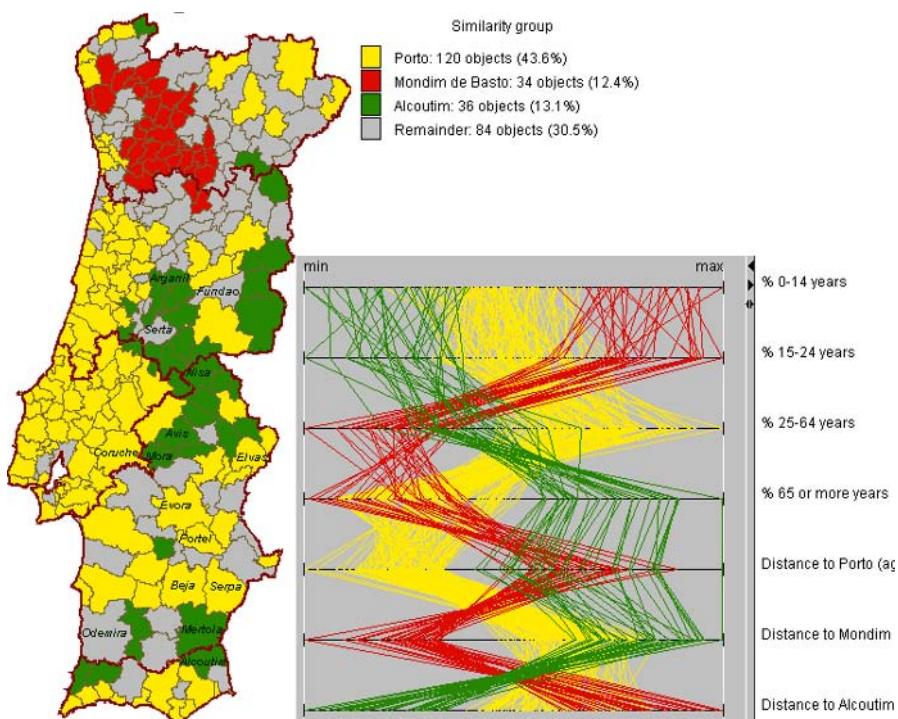
**Fig. 4.86C.** The forest data have been aggregated here by the cells of a regular rectangular grid. The pie charts represent the mean percentages of different types of forest: coniferous, broadleaved, mixed, and other wooded land. The sizes of the pies are proportional to the sums of these values, and hence show the approximate proportions of forest-covered land in the cells of the grid (Sect. 4.5.4.6)



**Fig. 4.87C.** The resolution of the aggregating grid has been increased as compared to Fig. 4.86C. The aggregate characteristics have been recomputed, and the same visualisation technique as before has been applied to the new data (Sect. 4.5.4.6)



**Fig. 4.104C.** Colour coding is used here to denote the satisfaction of different query conditions. For a query with two conditions, four different colours are needed (Sect. 4.6.2.2)



**Fig. 4.107C.** Groups of similar districts defined on the basis of their distances in terms of attribute values to Porto, Mondim de Basto, and Alcoutim. Districts similar to each of these three districts are coloured in yellow, red, and green, respectively. The districts dissimilar to any of these three districts are shown in grey (Sect. 4.6.2.4)

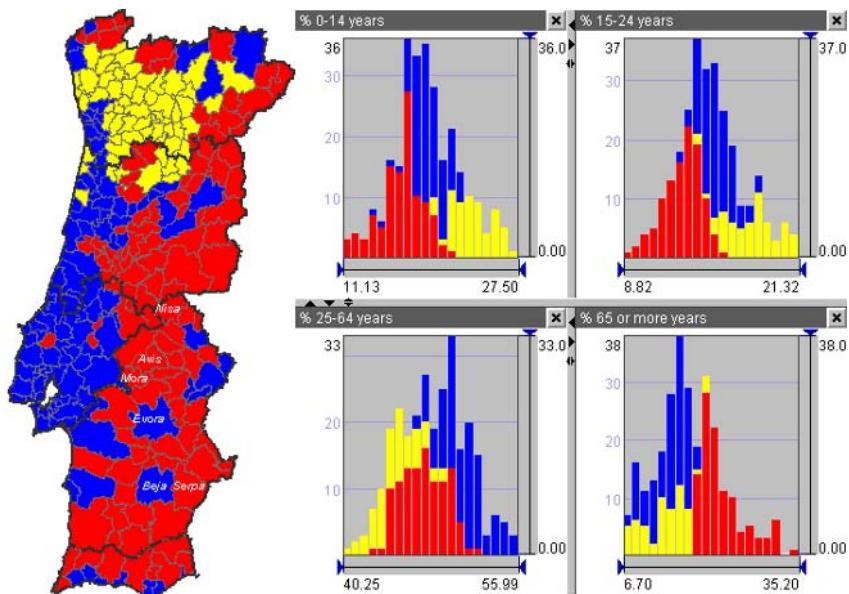


Fig. 4.120C. The result of dividing the districts into three clusters by the method of “simple k-means” (Sect. 4.7.4)

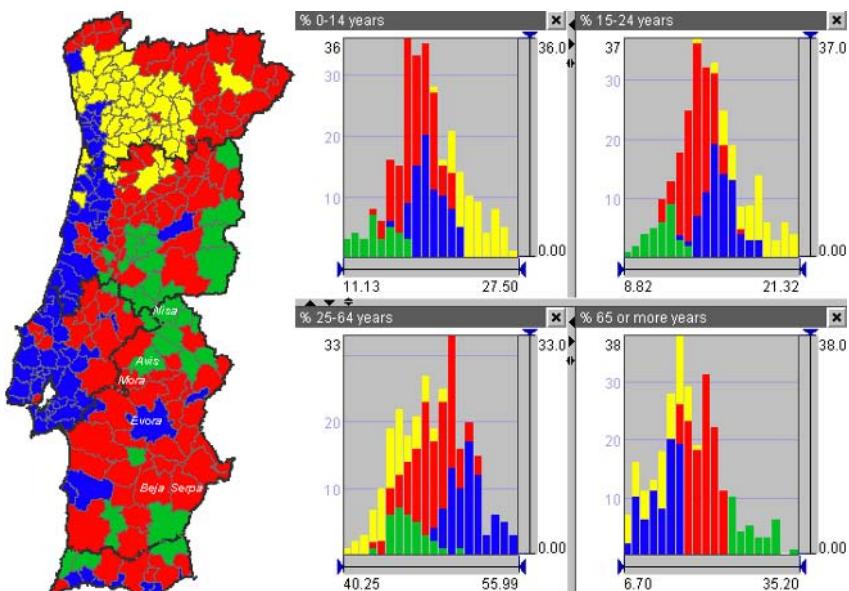
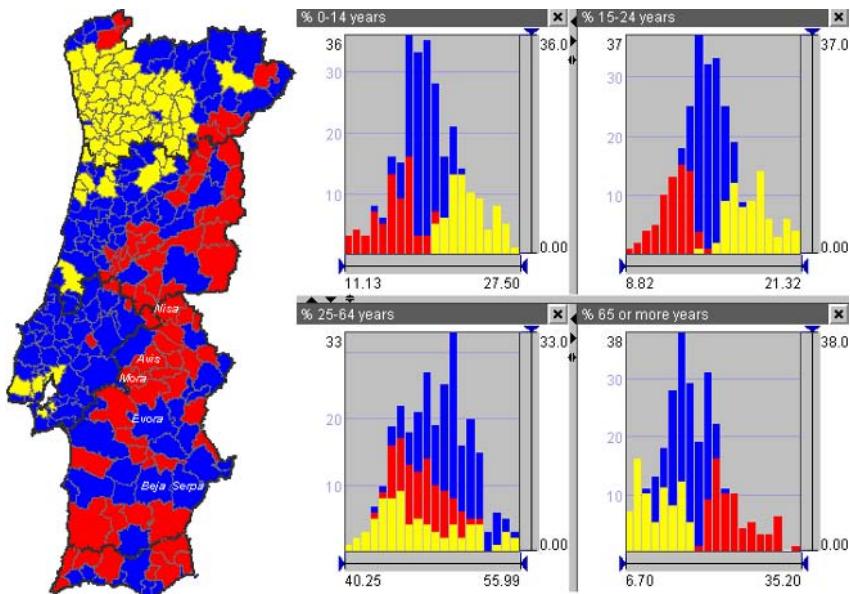
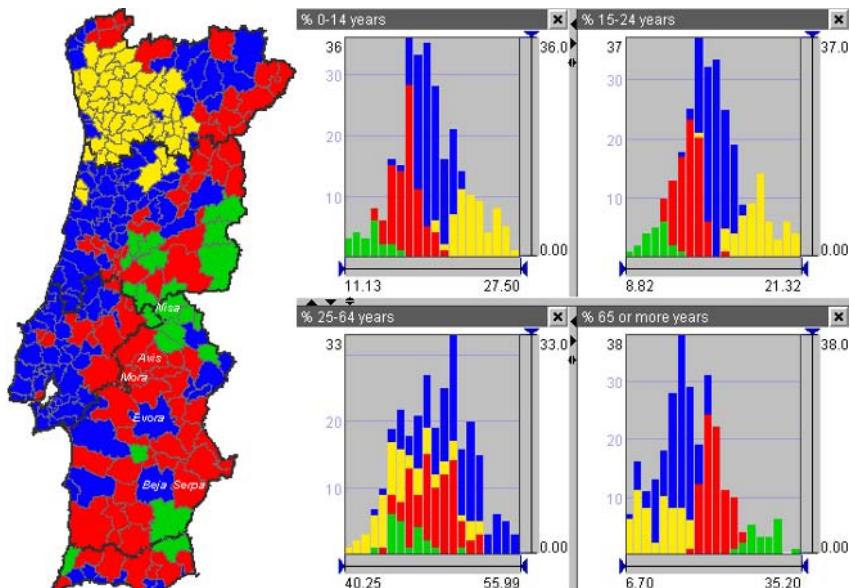


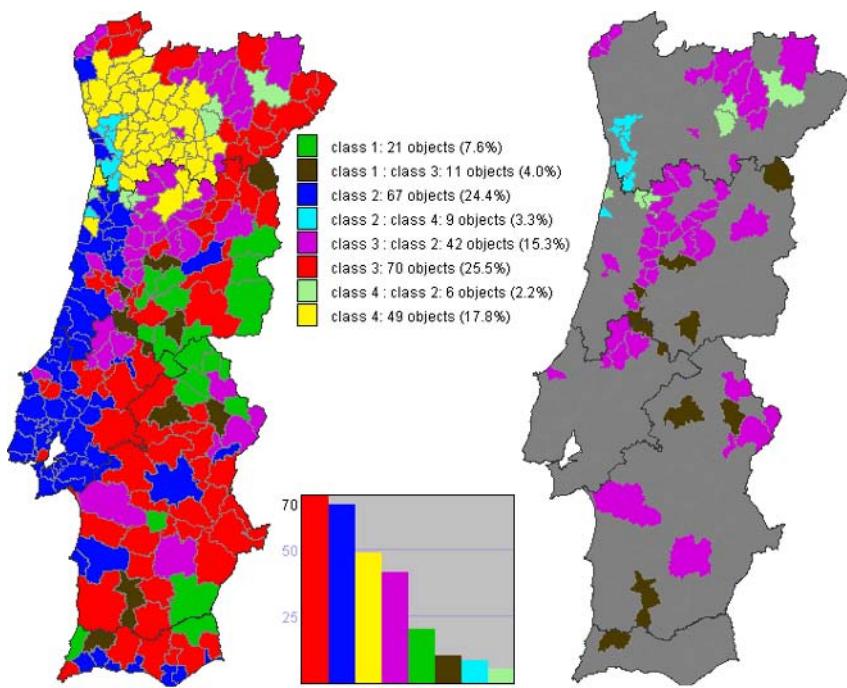
Fig. 4.121C. The result of dividing the districts into four clusters by the method of “simple k-means” (Sect. 4.7.4)



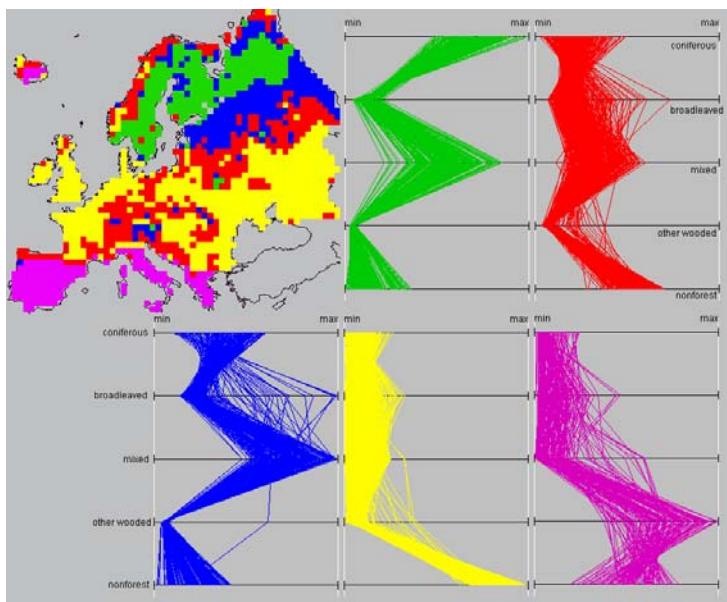
**Fig. 4.122C.** The result of dividing the districts into three clusters by the method of EM (expectation maximisation) (Sect. 4.7.4)



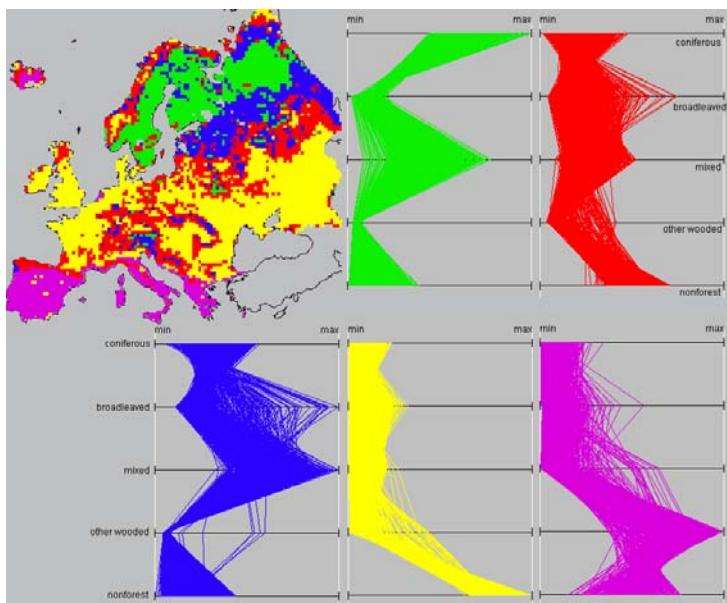
**Fig. 4.123C.** The result of dividing the districts into four clusters by the method of EM (Sect. 4.7.4)



**Fig. 4.125C.** The differences between the results of the methods of simple  $k$ -means and expectation maximisation can be seen better when the results are overlaid on the same map (left). On the right, only the districts that have moved to different clusters are shown in colour, while the districts that have preserved their membership are shown in grey (Sect. 4.7.4)



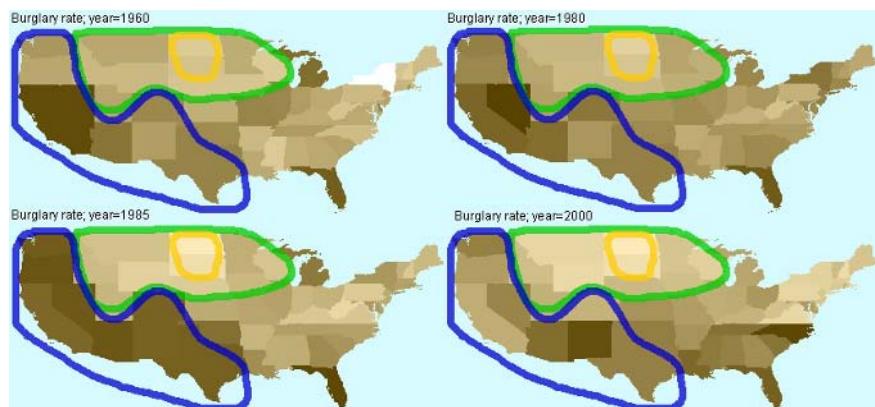
**Fig. 4.132C.** Results of applying a clustering tool to transformed raster data about the forest structure in Europe (Sect. 4.7.6)



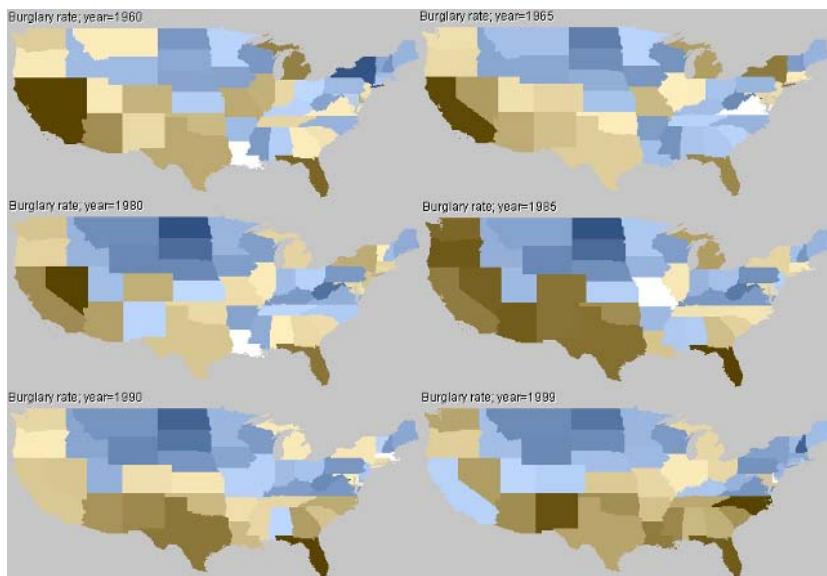
**Fig. 4.133C.** The result of reapplying the clustering tool after the resolution of the aggregating grid was increased. The profiles of the clusters are consistent with those of the previously obtained clusters shown in Fig. 4.132C (Sect. 4.7.6)



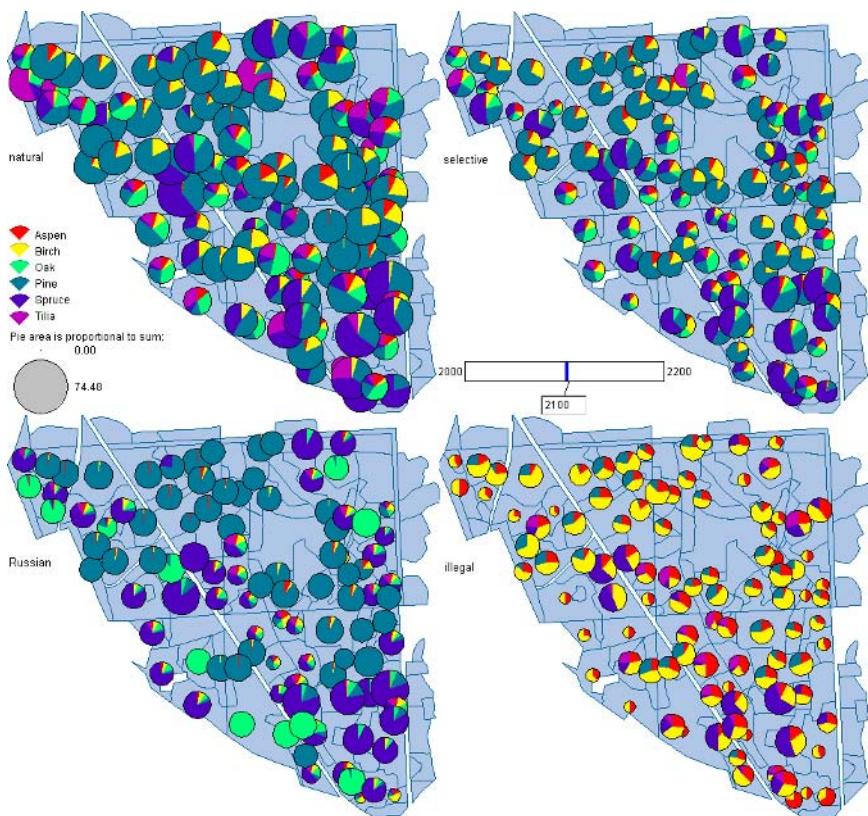
**Fig. 5.4C.** The most prominent clusters of similar local behaviours of the burglary rates in the states of the USA (Sect. 5.3)



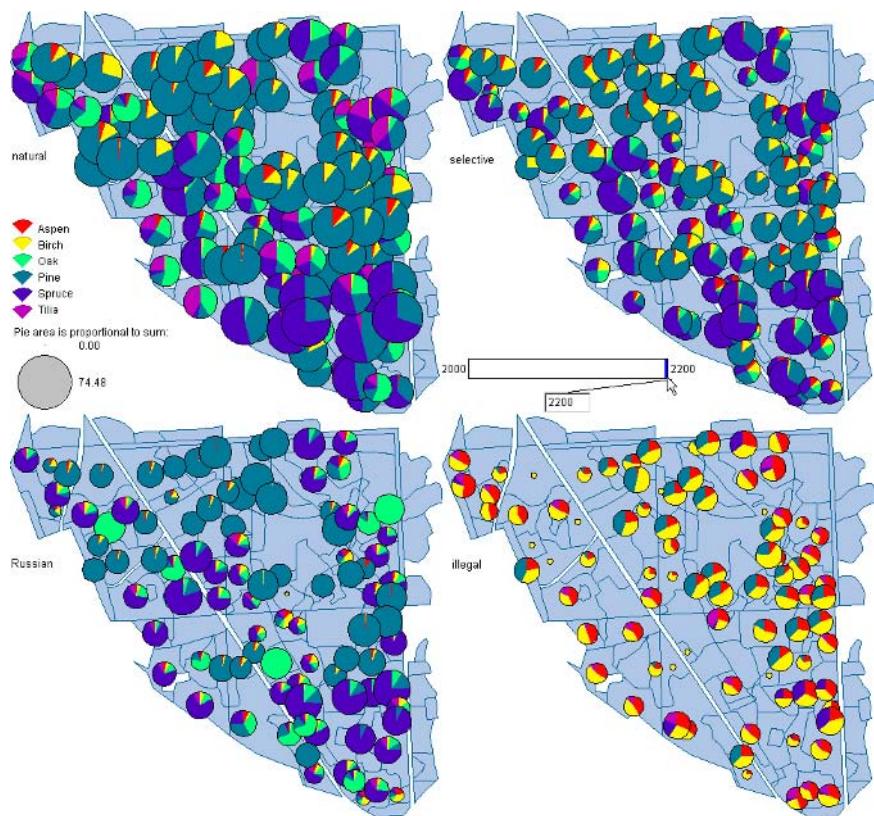
**Fig. 5.5C.** Comparison of the clusters of states with similar local behaviours with the evolution of the spatial distribution over time (Sect. 5.3)



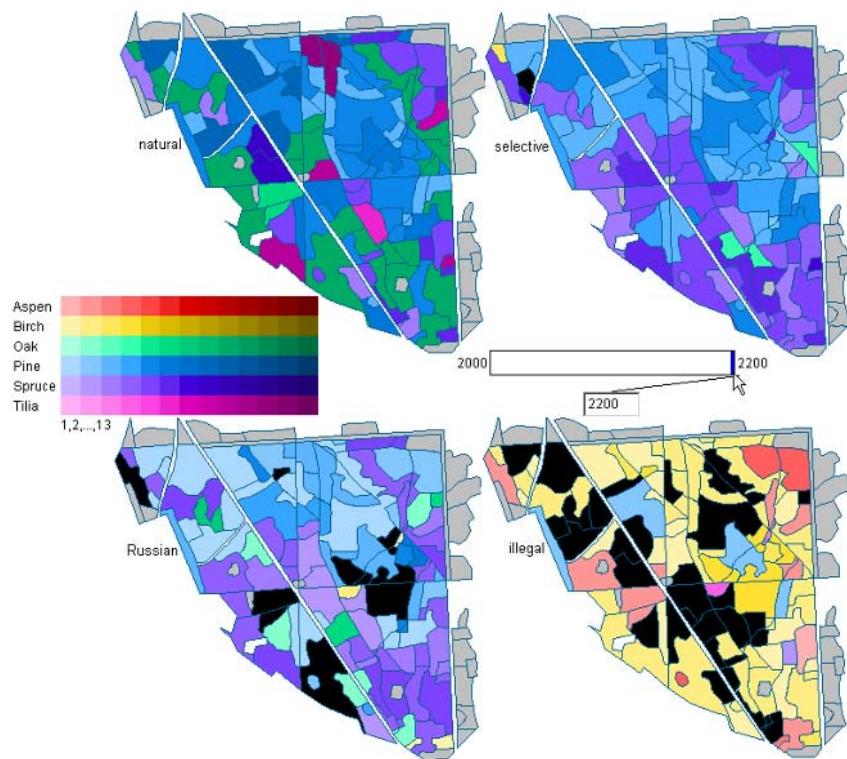
**Fig. 5.6C.** Changes of the spatial behaviour of the burglary rate over time. To increase expressiveness, visual comparison with the yearly country median has been applied in each map (Sect. 5.3)



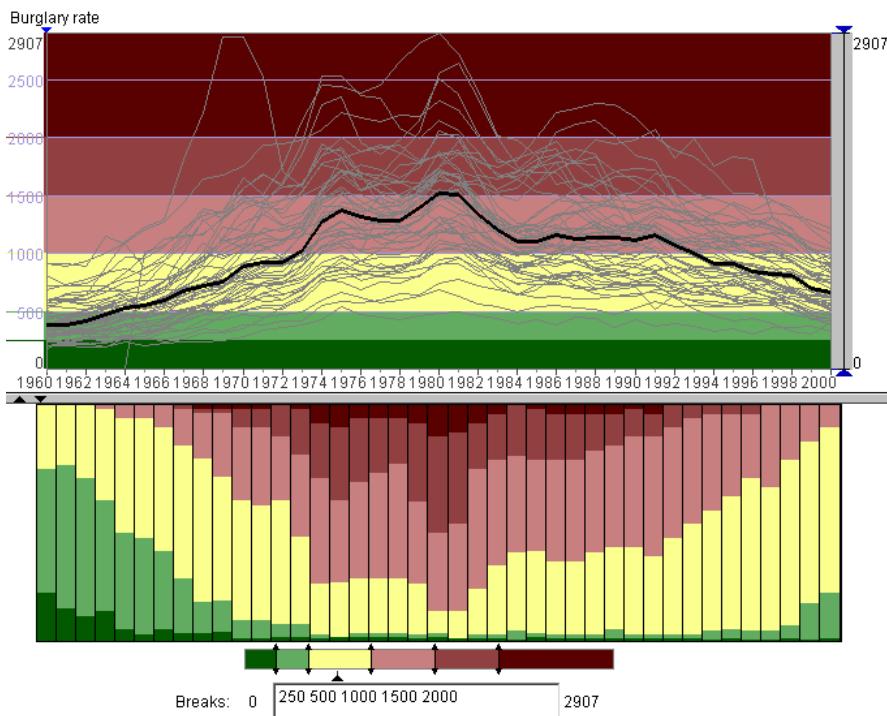
**Fig. 5.8C.** Visualisation of the forest management data. In order to reduce the dimensionality, the data have been aggregated: the areas occupied by different age groups of the same species have been summed. The result of the transformation has been visualised by means of a collection of four animated maps, each map corresponding to one forest management scenario: natural (upper left), selective cutting (upper right), Russian legal system (lower left), and illegal cutting (lower right). The screenshot shown here corresponds to the 100th simulation year of the whole 200-year long simulation period. The pie charts represent the areas in each forest compartment occupied by different tree species (all age groups being combined). The sizes of the pie charts are proportional to the total areas occupied by all the species together (Sect. 5.4.1.1)



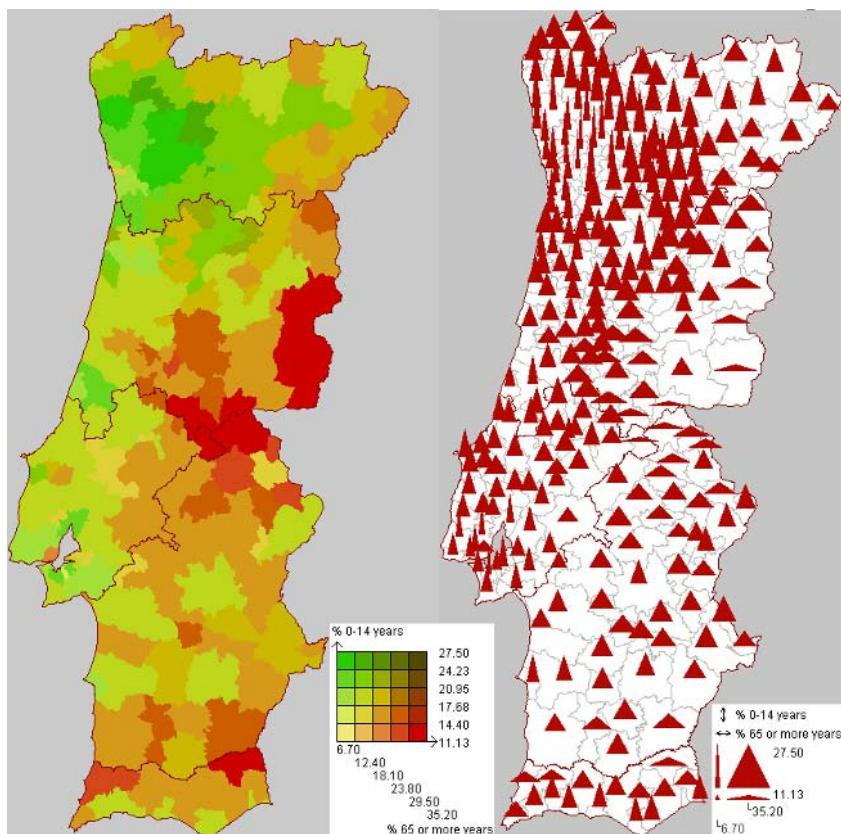
**Fig. 5.9C.** Another screenshot of the same visualisation as in Fig. 5.8C, showing the situation in the 200th simulation year achievable under each forest management strategy (Sect. 5.4.1.1)



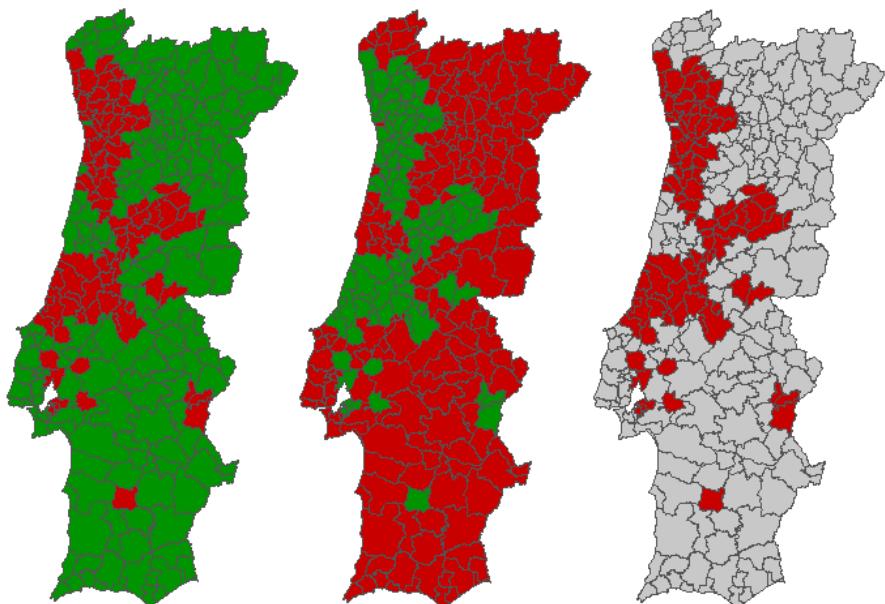
**Fig. 5.10C.** The maps here portray the dominant species and age groups by forest compartment in the 200th simulation year under the four different forest management strategies. Colour hues are used to represent the species, and degrees of darkness represent the age groups, with light shades corresponding to young ages and dark shades to older ages. Black signifies compartments that have no or very few trees because of cutting (Sect. 5.4.1.1)



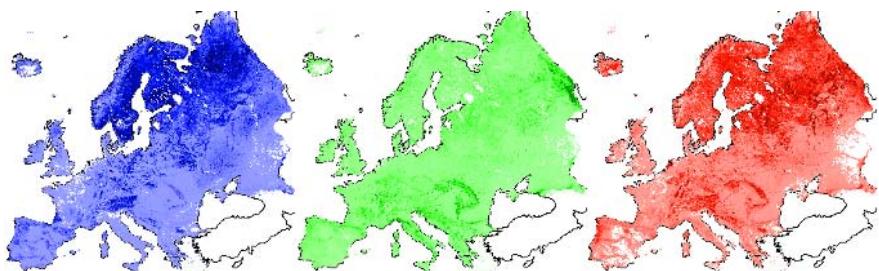
**Fig. 5.11C.** An aggregated representation of multiple time series based on dividing the value range of the attribute into intervals. The lower display represents the sizes of the aggregates at each moment in time by proportional heights of the coloured bar segments (Sect. 5.4.1.1)



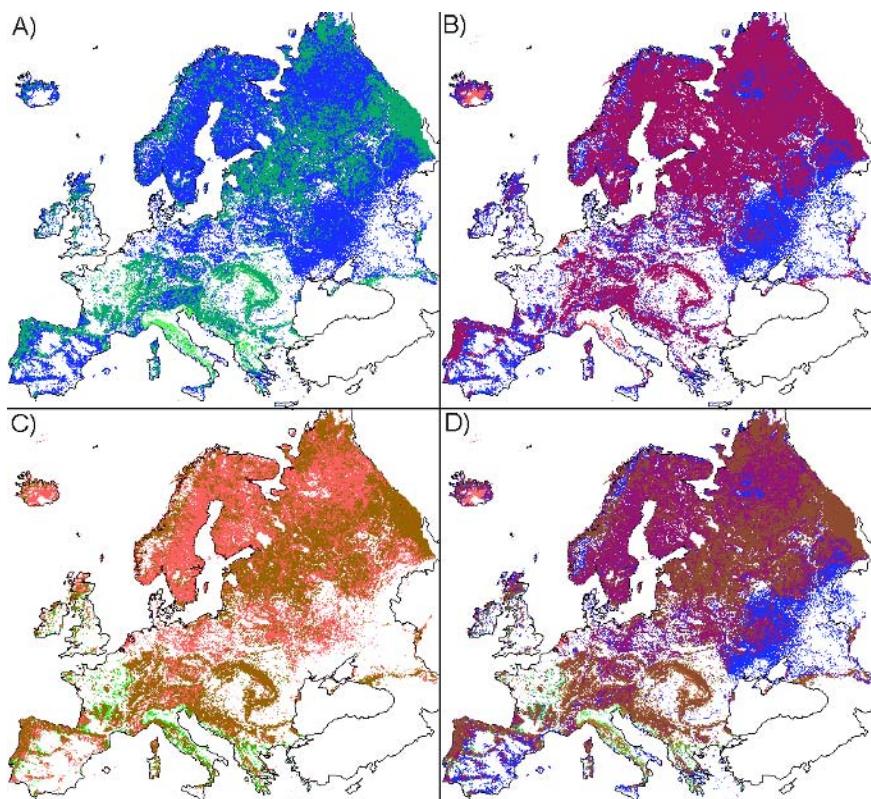
**Fig. 5.12C.** Two age structure attributes, “% 0–14 years” and “% 65 or more years”, are jointly represented on each map. On the left, the colouring of the districts corresponds to the values of the two attributes. The degree of greenness corresponds to the proportion of children in the population (the more children, the greener the colour), and the degree of redness corresponds to the proportion of elderly people (the more elderly people, the redder the colour). Low values of both attributes are reflected in yellow shades. On the right, the values of the attributes are “packed” into the dimensions of the triangular marks: the widths represent the proportion of elderly people, and the heights represent the proportion of children (Sect. 5.4.1.2)



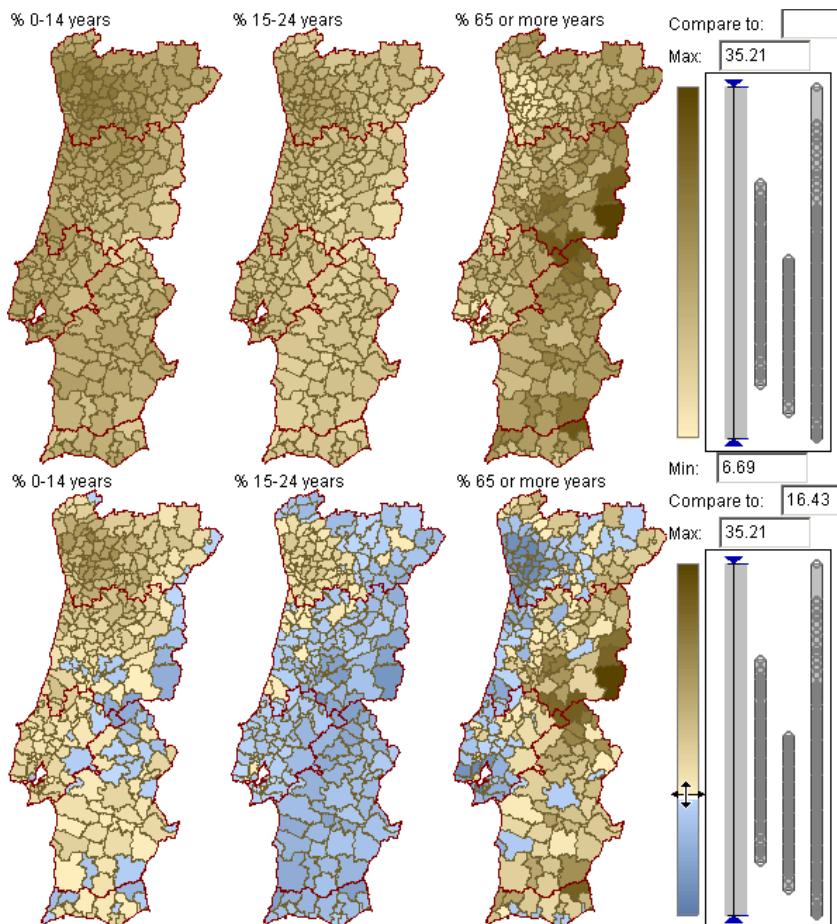
**Fig. 5.17C.** When equally bright, saturated colours are used to represent classes, this may impede the differentiation into figure and background, and hence complicate the visual grouping and perception of the overall pattern. Thus, spatial clusters of districts with close characteristics are better perceived from the map on the right (one can see red figures against a grey background) than from either of the two other maps, where the same two classes are represented using red and green colours (Sect. 5.4.3)



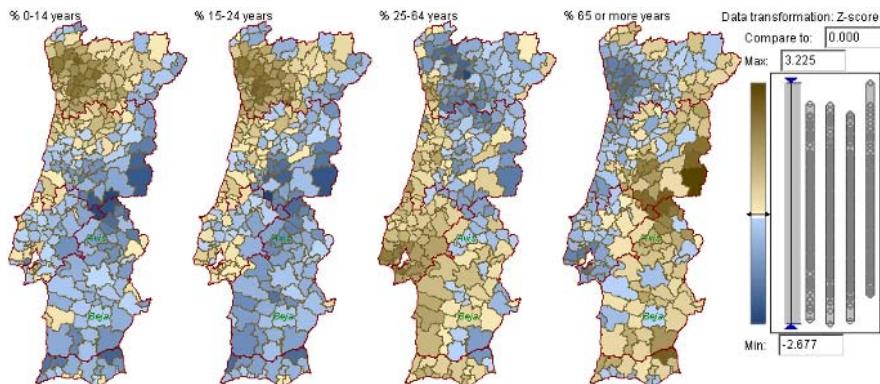
**Fig. 5.18C.** Three concurrent map displays representing three attributes characterising the forest structure in Europe: the percentage of coniferous forest (blue, on the left), the percentage of broadleaved forest (green, in the centre), and the percentage of mixed forest (red, on the right) (Sect. 5.4.4)



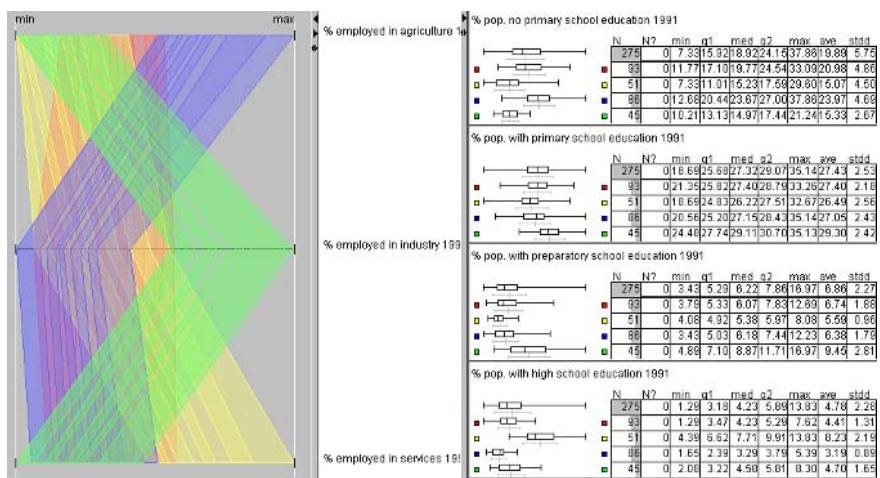
**Fig. 5.19C.** The same three forest structure attributes as in Fig. 5.18C are represented here as different map layers overlaid in a single map display. The maps A–D correspond to different layer combinations: A, coniferous and broadleaved; B, coniferous and mixed; C, broadleaved and mixed; D, all three layers. The layers drawn on top of others are shown in a semi-transparent mode. In all the layers, small attribute values have been filtered out by means of a dynamic query tool. The query constraints were selected so as to make the characteristic features of the spatial behaviours well exposed (Sect. 5.4.4)



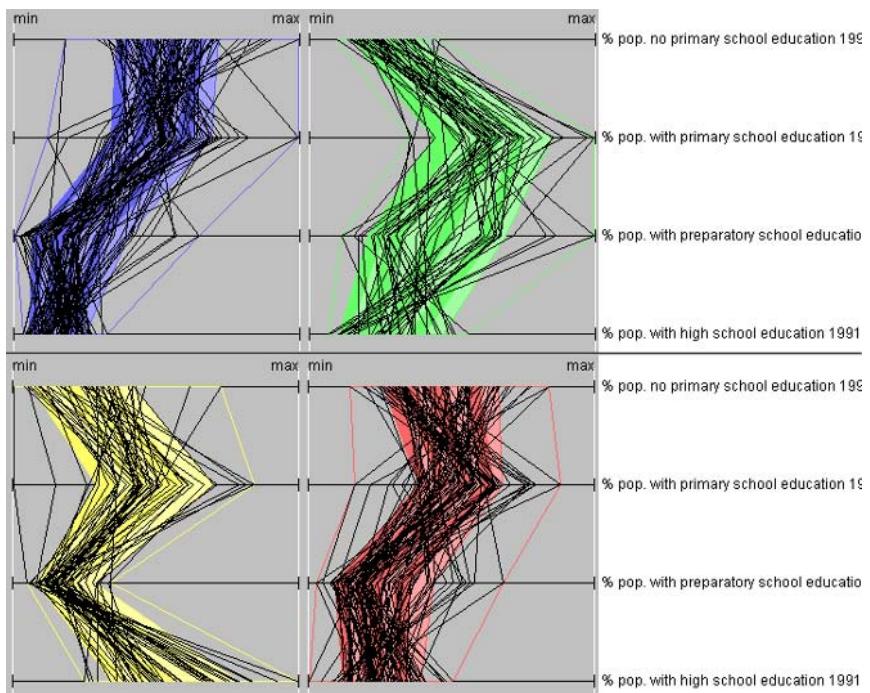
**Fig. 5.20C.** Behaviours of several numeric attributes with close value ranges may be compared using multiple displays with a common visual encoding function and common display manipulation tools. Here, three attributes are represented in unclassified choropleth maps with a common function for encoding the values by colour shades. In the lower row, the operation of visual comparison has been simultaneously applied to all three maps. The reference value in the visual comparison is the same in all the maps (Sect. 5.4.4)



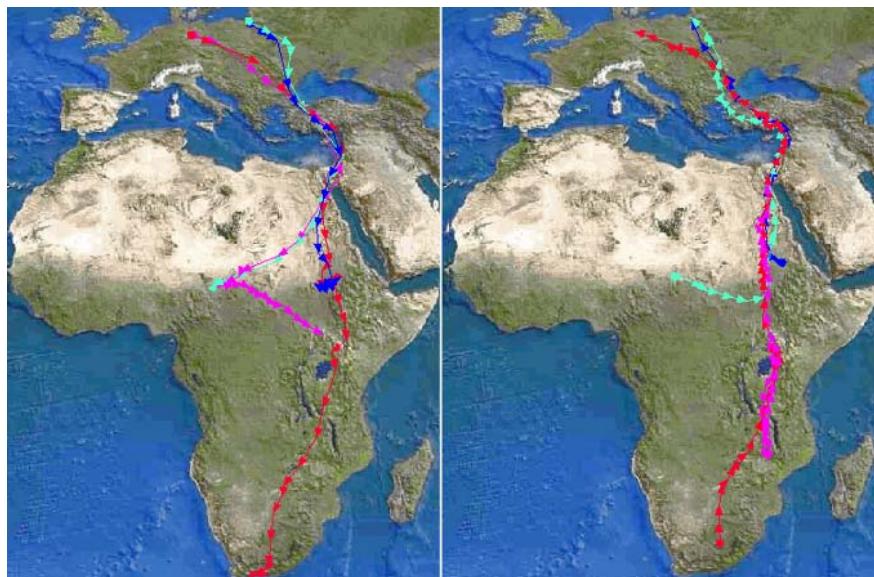
**Fig. 5.21C.** Transformation from the original attribute values to *z*-scores makes the behaviours of different attributes more comparable. In particular, the similarities between the behaviours of “% 0–14 years” and “% 15–24 years” can be seen more clearly than in Fig. 5.20C (Sect. 5.4.4)



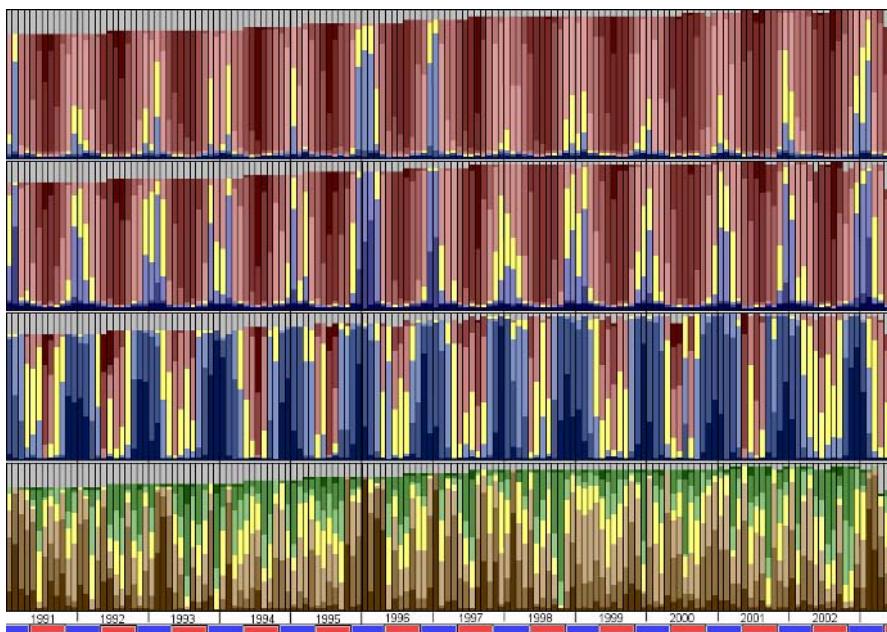
**Fig. 5.34C.** By use of a clustering tool, the districts of Portugal have been divided into four classes according to the employment of the population in different sectors of the economy, namely agriculture, industry, and services. The characteristics of the classes are represented in an aggregated form in the parallel-coordinates display on the left. On the right, the statistics of the values of four attributes reflecting the education level of the population are shown for the entire country and for the four classes of districts (Sect. 5.4.8)



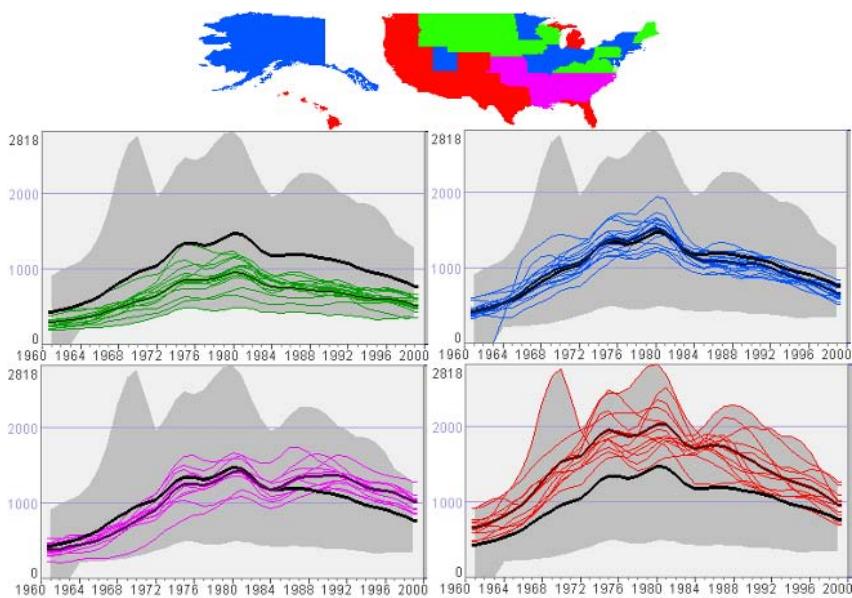
**Fig. 5.35C.** For the four classes of districts of Portugal defined according to the employment structure of the population, the profiles in terms of the four education-related attributes are shown here in four different parallel-coordinates displays (Sect. 5.4.8)



**Fig. 5.36C.** A satellite image with a superimposed representation of the movement of the storks demonstrates links between the movement and the characteristics of the underlying ground surface. On the left, the movement during the period from 20 August 1998 to 31 January 1999 is shown, and on the right, the movement from 1 February 1999 to 1 May 1999 (Sect. 5.4.8)

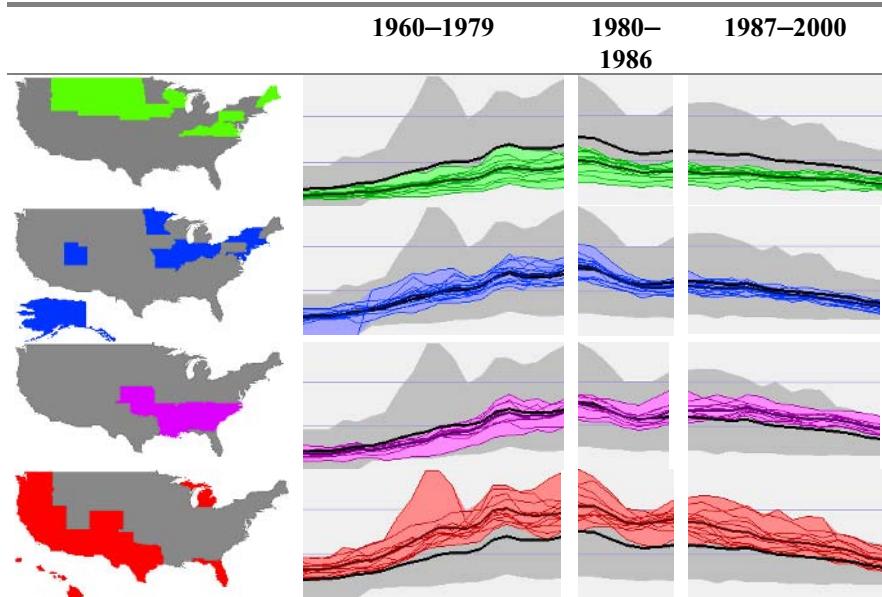


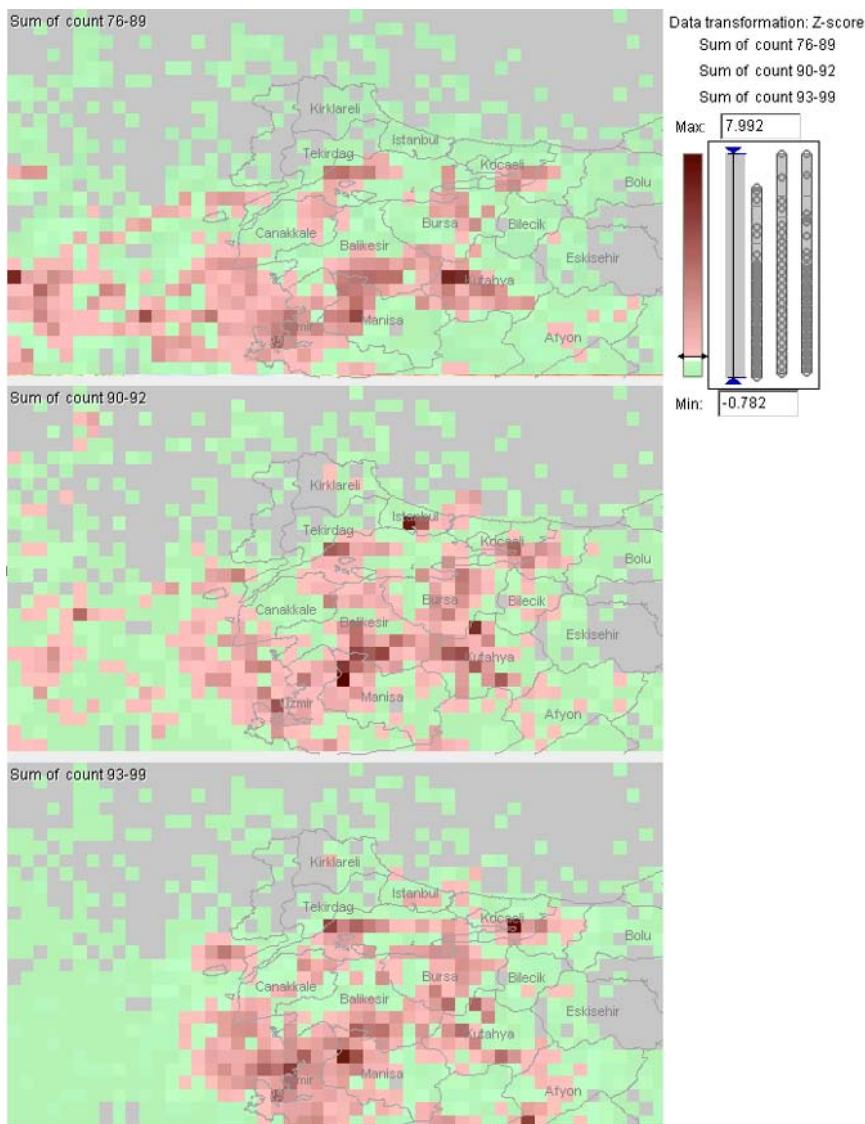
**Fig. 5.37C.** This display represents the temporal variation of the values of four climate attributes aggregated over Germany by months. From top to bottom: the monthly mean of the daily mean temperature, the monthly mean of the daily minimum temperature, the total monthly sunshine duration, and the total monthly precipitation (Sect. 5.4.8)



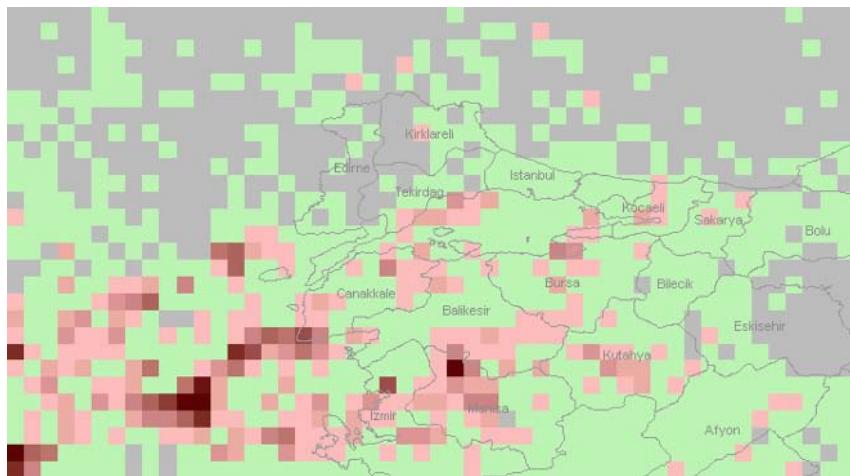
**Fig. 5.38C.** The states of the USA have been divided here into four clusters according to the similarity of their local temporal behaviours (Sect. 5.4.8)

**Table 5.4.** Partial temporal behaviours of the burglary rate by groups of states and by time period (Sect. 5.4.8)

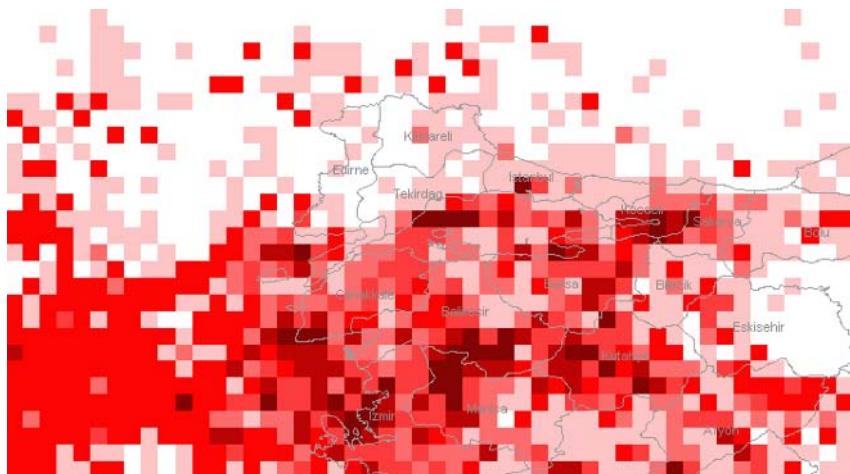




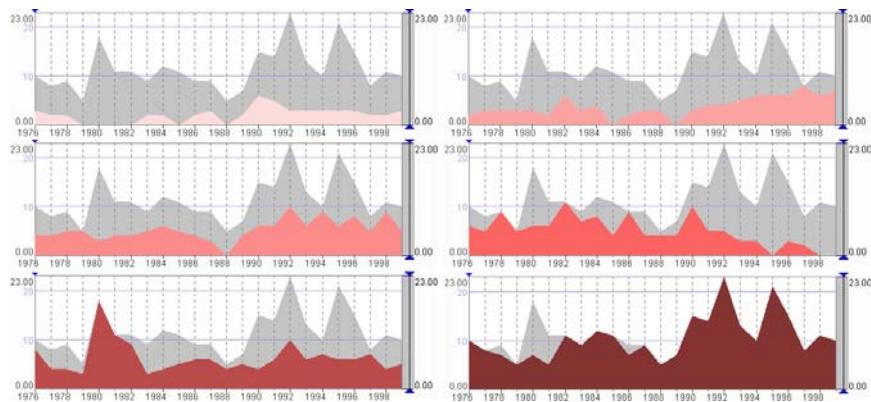
**Fig. 5.45C.** After summing the frequencies over three time periods, namely 1976–1989, 1990–1992, and 1993–1999, we have obtained a sort of generalised portrait of the typical spatial behaviours in these periods. The data for the year 1982 have not been included in the computation, since the behaviour in this year differs from those in the other years of the period 1976–1989 (Sect. 5.6)



**Fig. 5.46C.** The spatial behaviour of the earthquake frequency in 1982 is visualised here in the same way as for the “summarised” behaviours in the three time periods in the previous figure for a more convenient comparison (Sect. 5.6)



**Fig. 5.50C.** Results of clustering according to similarity of the local behaviours (Sect. 5.6)



**Fig. 5.51C.** The outlines of the behaviours united in the clusters shown in Fig. 5.50C (Sect. 5.6)