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):

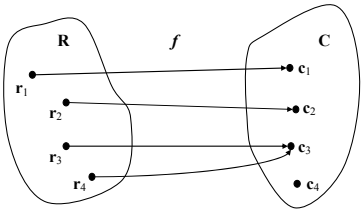

Fig. 1.1. The functional view of a dataset

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:

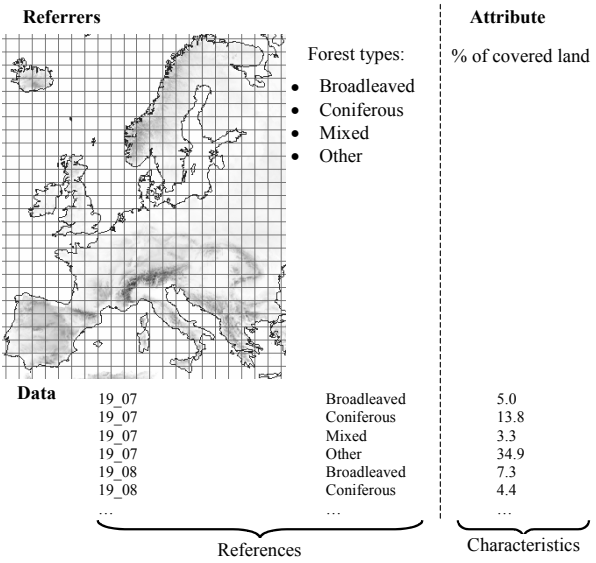

Fig. 1.2. A visual representation of the structure of a dataset

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?"

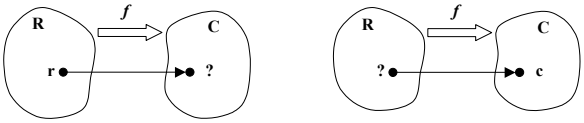

Fig. 1.3. Two types of tasks are represented schematically on the basis of the functional view of data

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:

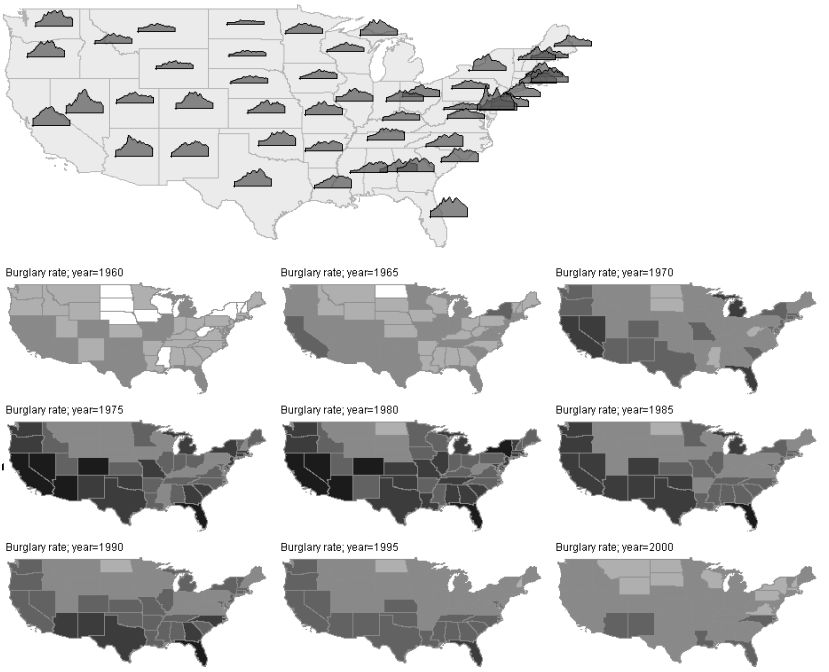

Fig. 1.4. Two possible views of the same space- and time-referenced data

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

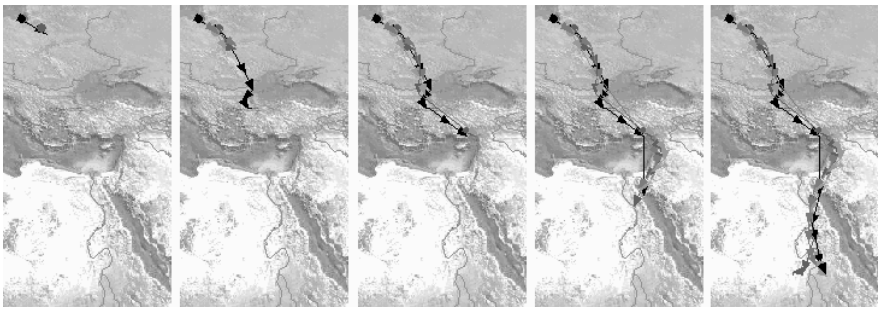

Fig. 1.5. A visualisation of the movement of white storks.

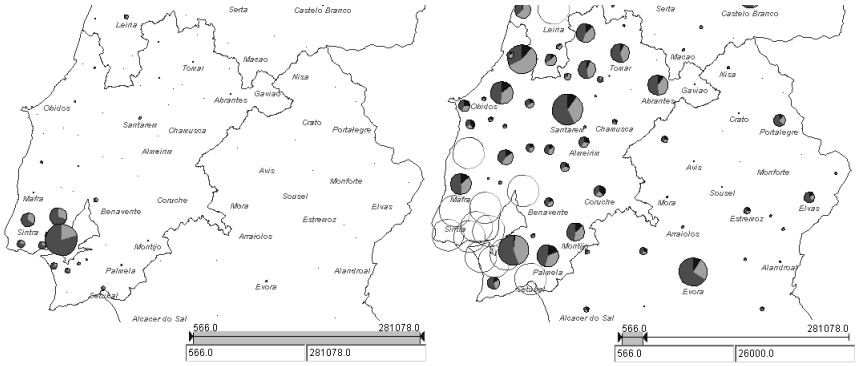

Fig. 1.6. An example of a display manipulation technique: focusing

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).

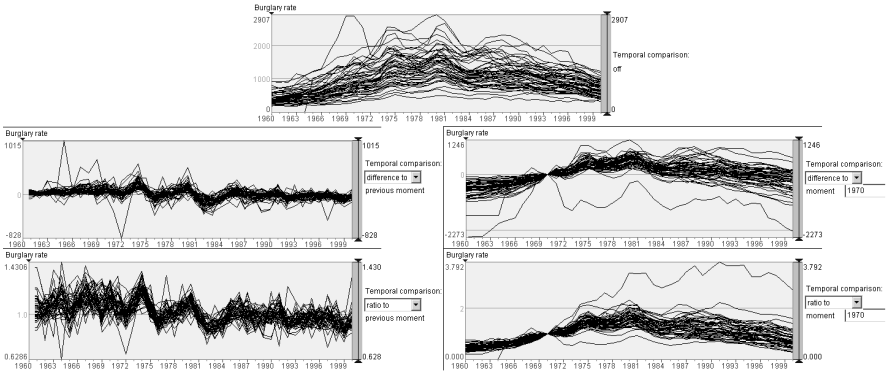

Fig. 1.7. Examples of various transformations of time-series data.

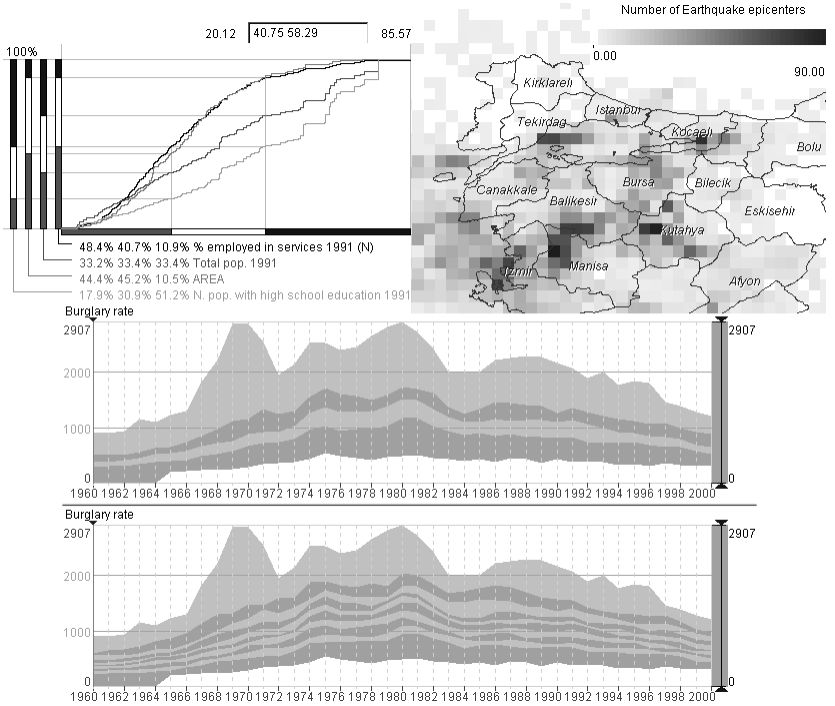

Fig. 1.8. A few examples of data aggregation

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.

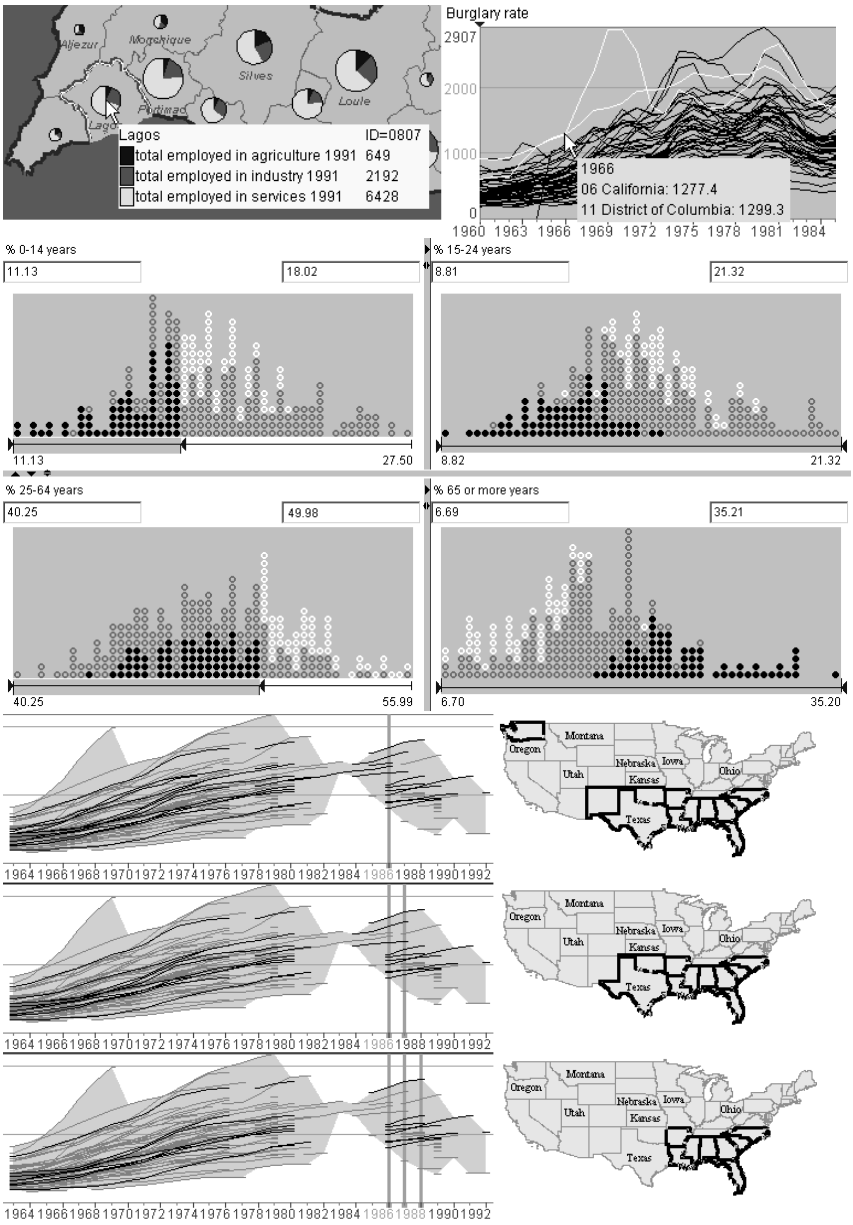

Fig. 1.9. Examples of dynamic querying tools

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.

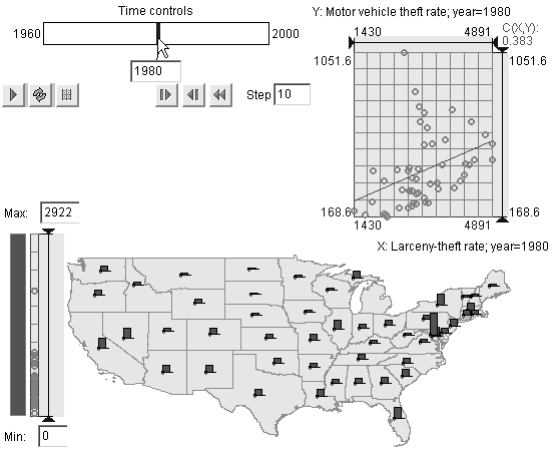

Fig. 1.10. Several tools working in combination

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.

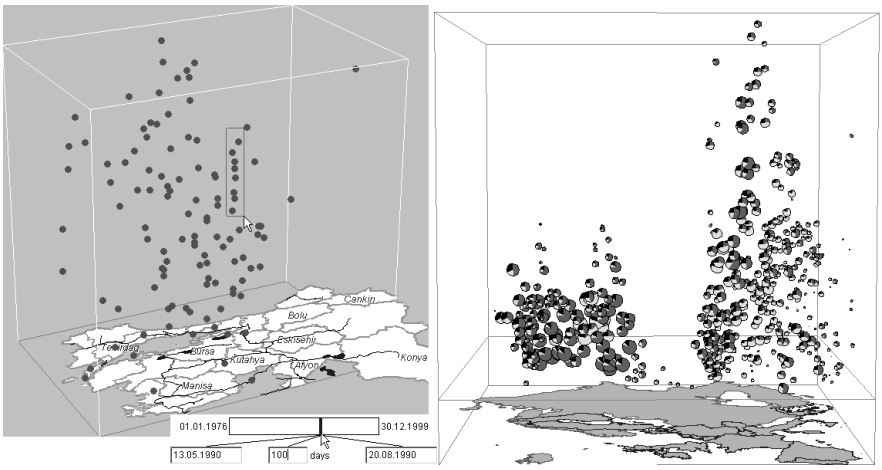

Fig. 1.11. Illustration of some of the principles

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 single referrer, 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)

http://www.springer.com/978-3-540-25994-7

Exploratory Analysis of Spatial and Temporal Data
A Systematic Approach
Andrienko, N.; Andrienko, G.
2006, XV, 703 p., Hardcover
ISBN: 978-3-540-25994-7