1 Data Analysis in Earth Sciences

1.1 Introduction

Earth scientists make observations and gather data about natural processes on Earth. They formulate and test hypotheses on the forces that have operated in a certain region to create its structure. They also make predictions about future changes of the planet. All these steps in exploring the system Earth include the acquisition and analysis of numerical data. An earth scientist needs a solid knowledge in statistical and numerical methods to analyze these data, as well as the ability to use suitable software packages on a computer.

This book introduces some of the most important methods of data analysis in earth sciences and illustrates the use of these methods by MATLAB examples. The examples can be used as recipes for the analysis of the reader's real data after learning their application on synthetic data. The introductory Chapter 1 deals with data acquisition (Chapter 1.2), the expected types of data (Chapter 1.3) and the suitable methods for analyzing data in earth sciences (Chapter 1.4). Therefore, we first explore the characteristics of a typical data set. Subsequently, we investigate the various ways of analyzing data with MATLAB.

1.2 Collecting Data

Data sets in earth sciences have a very limited sample size. They also contain a significant amount of uncertainties. Such data sets are typically used to describe rather large natural phenomena, such as a granite body, a large landslide and a widespread sedimentary unit. The methods described in this book help finding a way of predicting the characteristics of a larger *population* from the collected *samples* (Fig. 1.1). A proper sampling strategy is the first step to obtain a good data set. The development of a successful strategy for field sampling includes decisions on

M. H. Trauth, *MATLAB® Recipes for Earth Sciences*

- the *sample size* – This parameter includes the sample volume, the sample weight and the number of samples collected in the field. The rock weight or volume can be a critical factor if the samples are later analyzed in the laboratory. Most statistical methods also have a minimum required sample size. The sample size also restricts the number of subsamples that can be collected from the single sample. If the population is heterogeneous, then the sample needs to be large enough to represent the population's variability. On the other hand, a sample should be as small as possible to save time and effort to analyze it. It is recommended to collect a smaller pilot sample before defining a suitable sample size.

- the *spatial sampling scheme* – In most areas, samples are taken as the availability of outcrops permits. Sampling in quarries typically leads to

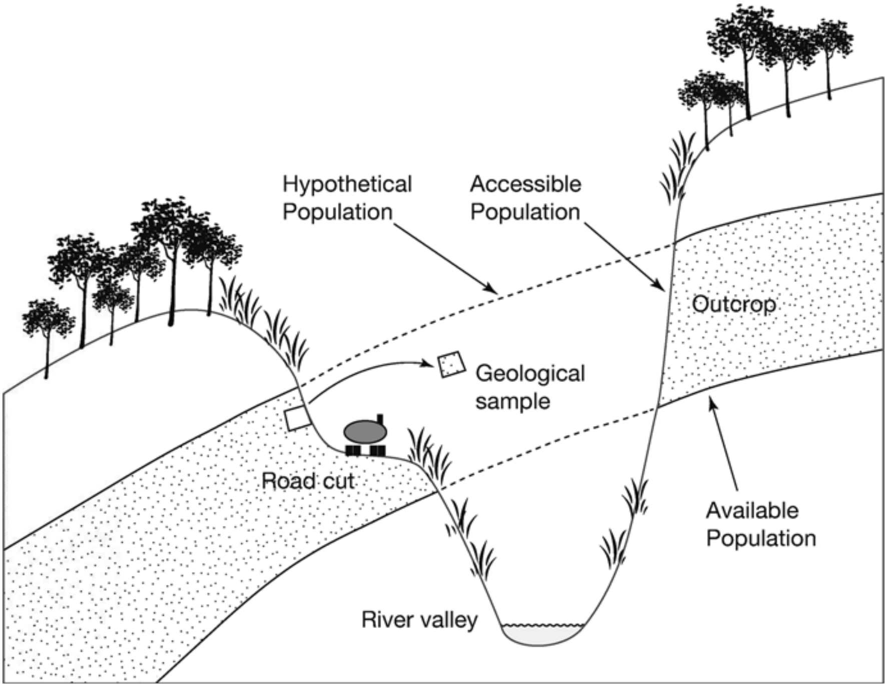

Fig. 1.1 Samples and population. Deep valley incision has eroded parts of a sandstone unit (*hypothetical population*). The remnants of the sandstone (*available population*) can only be sampled from outcrops, i.e., road cuts and quarries (*accessible population*). Note the difference between a statistical sample as a representative of a population and a geological sample as a piece of rock.

clustered data, whereas road cuts, shoreline cliffs or steep gorges cause traverse sampling schemes. If there are no financial limitations or the area allows hundred percent access to the rock body, a more uniform sampling pattern can be designed. A regular sampling scheme results in a gridded distribution of sample locations, whereas a uniform sampling strategy includes the random location of a sampling point within a grid square. You might expect that these sampling schemes represent the superior method to collect the samples. However, evenly-spaced sampling locations tend to miss small-scale variations in the area, such as thin mafic dykes in a granite body or the spatially-restricted occurrence of a fossil (Fig. 1.2).

The proper sampling strategy depends on the type of object to be analyzed, the aims of the investigation and the required level of confidence of the result. Having chosen a suitable sampling strategy, the quality of the set of samples can be influenced by a number of disturbances. The samples might not be representative of the larger population. Chemical or physical alteration, contamination by other material or dislocation by natural and anthropogenic processes may result in erroneous results and interpretations. Therefore, it is recommended to test the quality of the sample, the method of data analysis employed and the validity of the conclusions based on the analysis in all stages of the investigation.

1.3 Types of Data

Most data in geosciences consist of numerical measurements, although some information can also be represented by a list of names such as fossils and minerals (Fig. 1.3). The available methods for data analysis may require certain types of data in earth sciences. These are

- *nominal data* – Information in earth sciences is sometimes presented as a list of names, e.g., the various fossil species collected from a limestone bed or the minerals identified in a thin section. In some studies, these data are converted into a binary representation, i.e., *one* for present and *zero* for absent. Special statistical methods are available for the analysis of such data sets.

- *ordinal data* – These are numerical data representing observations that can be ranked, but the intervals along the scale are not constant. Mohs' hardness scale is one example for an ordinal scale. The hardness value

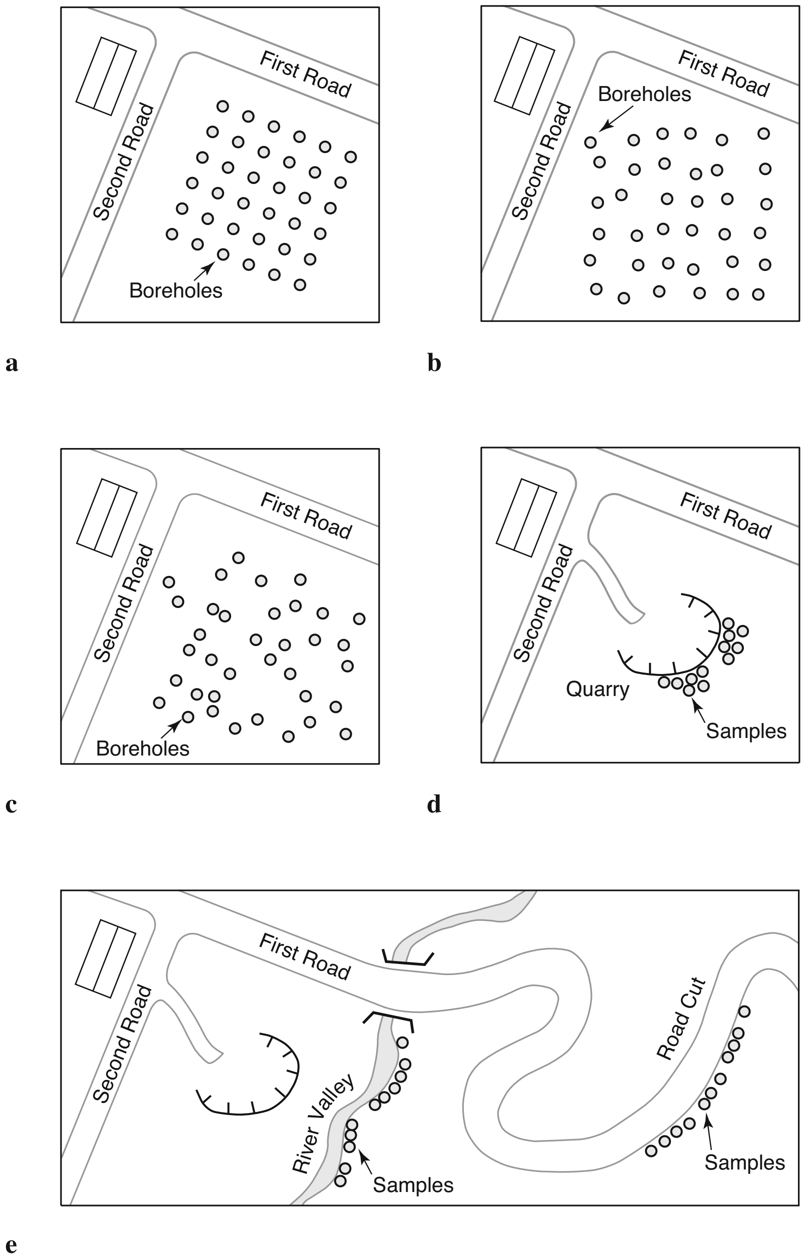

Fig. 1.2 Sampling schemes. **a** *Regular sampling* on an evenly-spaced rectangular grid, **b** *uniform sampling* by obtaining samples randomly-located within regular grid squares, **c** *random sampling* using uniform-distributed *xy* coordinates, **d** *clustered sampling* constrained by limited access in a quary, and **e** *traverse sampling* along road cuts and river valleys.

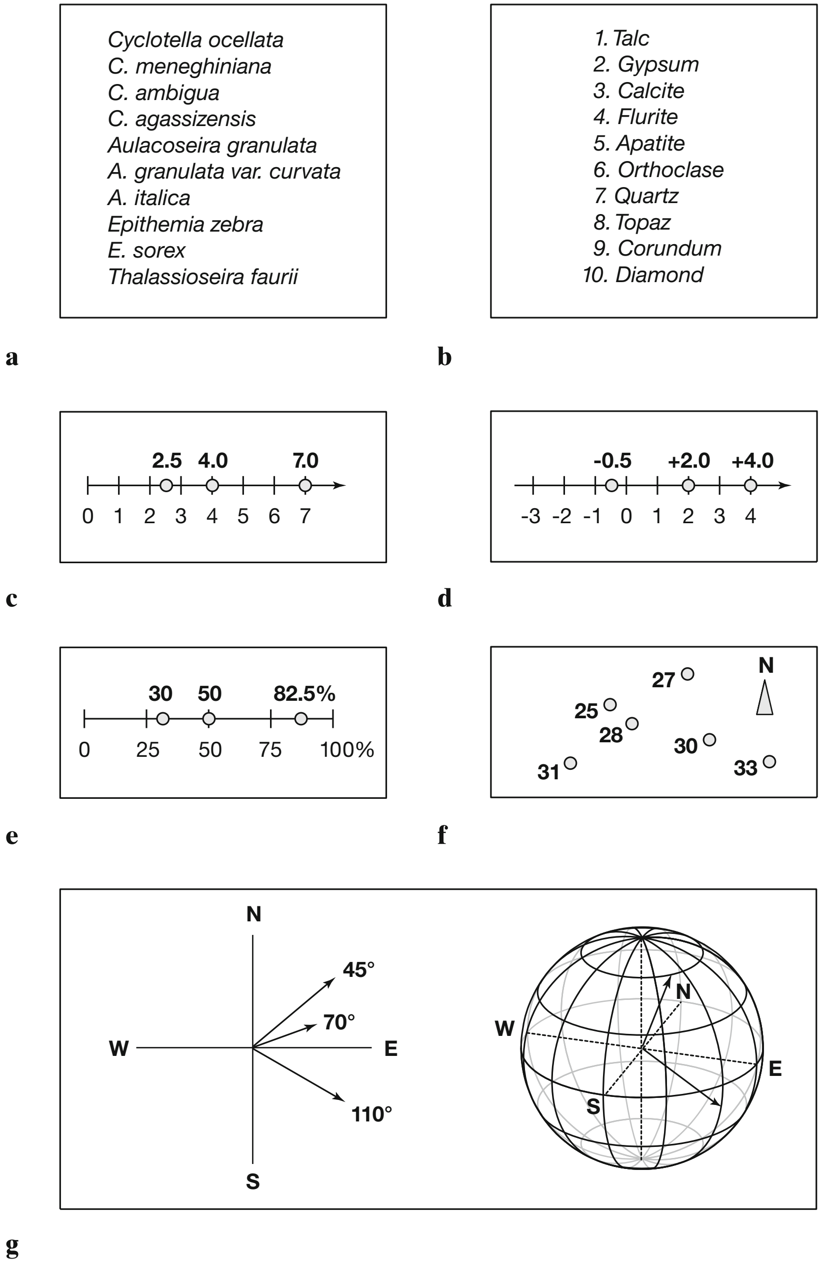

Fig. 1.3 Types of data in earth sciences. **a** *Nominal data,* **b** *ordinal data,* **c** *ratio data,* **d** *interval data,* **e** *closed data,* **f** *spatial data* and **g** *directional data*. All data types are described in the book.

indicates the materials resistance to scratching. Diamond has a hardness of 10, whereas this value for talc is 1. In terms of absolute hardness, diamond (hardness 10) is four times harder than corundum (hardness 9) and six times harder than topaz (hardness 8). The Modified Mercalli Scale to categorize the size of earthquakes is another example for an ordinal scale. It ranks earthquakes from intensity I (barely felt) to XII (total destruction).

- *ratio data* – These data are characterized by a constant length of successive intervals. Therefore, ratio data offer a great advantage in comparison to ordinal data. However, the zero point is the natural termination of the data scale. Examples of such data sets include length or weight data. This type of data allows either a discrete or continuous data sampling.

- *interval data* – These are ordered data that have a constant length of successive intervals. The data scale is not terminated by zero. Temperatures C and F represent an example of this data type although zero points exist for both scales. This types of data may be sampled continuously or in discrete intervals.

Besides these standard types of data, earth scientists frequently encounter special kinds of data, such as

- *closed data* – These data are expressed as proportions and added to a fixed total such as 100 percent. Compositional data represent the majority of closed data, such as element compositions of rock samples.

- *spatial data* – These are collected in a 2D or 3D study area. The spatial distribution of a certain fossil species, the spatial variation of the sandstone bed thickness and the 3D tracer concentration in groundwater are examples for this type of data. This is likely to be the most important data type in earth sciences.

- *directional data* – These data are expressed in angles. Examples include the strike and dip of a bedding, the orientation of elongated fossils or the flow direction of lava. This is a very common type of data in earth sciences.

Most of these data require special methods to be analyzed, that are outlined in the next chapter.

1.4 Methods of Data Analysis

Data analysis methods are used to describe the sample characteristics as precisely as possible. Having defined the sample characteristics we hypothesize about the general phenomenon of interest. The particular method that is used for describing the data depends on the data type and the project requirements.

- *Univariate methods* – Each variable is explored separately assuming the variables are independent of each other. The data are presented as a list of numbers representing a series of points on a scaled line. Univariate statistical methods include the collection of information about the variable, such as the minimum and maximum value, the average and the dispersion about the average. Examples are the sodium content of volcanic glass shards that were affected by chemical weathering or the size of snail shells in a sediment layer.

- *Bivariate methods* – Two variables are investigated together to detect relationships between these two parameters. For example, the correlation coefficient may be calculated to investigate whether there is a linear relationship between two variables. Alternatively, the bivariate regression analysis helps to find an equation that describes the relationship between the two variables. An example for a bivariate plot is the *Harker Diagram*, which is one of the oldest method to visualize geochemical data and plots oxides of elements against SiO_2 from igneous rocks.

- *Time-series analysis* – These methods investigate data sequences as a function of time. The time series is decomposed into a long-term trend, a systematic (periodic, cyclic, rhythmic) and an irregular (random, stochastic) component. A widely used technique is spectral analysis, to describe cyclic components of the time series. Examples for the application of these techniques are the investigation of cyclic climate variations in sedimentary rocks or the analysis of seismic data.

- *Signal processing* – This includes all techniques for manipulating a signal to minimize the effects of noise, to correct all kinds of unwanted distortions or to separate various components of interest. It includes the design, realization and application of filters to the data. These methods are widely used in combination with time-series analysis, e.g., to increase the signal-to-noise ratio in climate time series, digital images or geophysical data.

- *Spatial analysis* – The analysis of parameters in 2D or 3D space. Therefore, two or three of the required parameters are coordinate numbers. These methods include descriptive tools to investigate the spatial pattern of geographically distributed data. Other techniques involve spatial regression analysis to detect spatial trends. Finally, 2D and 3D interpolation techniques help to estimate surfaces representing the predicted continuous distribution of the variable throughout the area. Examples are drainage-system analysis, the identification of old landscape forms and lineament analysis in tectonically-active regions.

- *Image processing* – The processing and analysis of images has become increasingly important in earth sciences. These methods include manipulating images to increase the signal-to-noise ratio and to extract certain components of the image. Examples are the analysis of satellite images, the identification of objects in thin sections and counting annual layers in laminated sediments.

- *Multivariate analysis* – These methods involve the observation and analysis of more than one statistical variable at a time. Since the graphical representation of multidimensional data sets is difficult, most methods include dimension reduction. Multivariate methods are widely used on geochemical data, for instance in tephrochronology, where volcanic ash layers are correlated by geochemical fingerprinting of glass shards. Another important example is the comparison of species assemblages in ocean sediments to reconstruct paleoenvironments.

- *Analysis of directional data* – Methods to analyze circular and spherical data are widely used in earth sciences. Structural geologists measure and analyze the orientation of slickenlines (or striae) on a fault plane. Circular statistical methods are also common in paleomagnetic studies. Microstructural investigations include the analysis of grain shapes and quartz c-axis orientation in thin sections.

Some of these methods require the application of numerical methods, such as interpolation techniques and some methods of signal processing. The following text is mainly on statistical techniques, but also introduces several numerical methods used in earth sciences.

Recommended Reading

Borradaile G (2003) Statistics of Earth Science Data – Their Distribution in Time, Space and Orientation. Springer, Berlin Heidelberg New York

Carr JR (1995) Numerical Analysis for the Geological Sciences. Prentice Hall, Englewood Cliffs, New Jersey

Davis JC (2002) Statistics and Data Analysis in Geology, Third Edition. John Wiley and Sons, New York

Hanneberg WC (2004) Computational Geosciences with Mathematica. Springer, Berlin Heidelberg New York

Mardia KV (1972) Statistics of Directional Data. Academic Press, London

Middleton GV (1999) Data Analysis in the Earth Sciences Using MATLAB. Prentice Hall, New Jersey

Press WH, Teukolsky SA, Vetterling WT (1992) Numerical Recipes in Fortran 77. Cambridge University Press, Cambridge

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) Numerical Recipes in C++. Cambridge University Press, Cambridge

Swan ARH, Sandilands M (1995) Introduction to Geological Data Analysis. Blackwell Sciences, Oxford

Upton GJ, Fingleton B (1990) Spatial Data Analysis by Example, Categorial and Directional Data. John Wiley & Sons, New York