1 Data Analysis in Earth Sciences

1.1 Introduction

Earth scientists make observations and gather data about the natural processes that operate on planet Earth. They formulate and test hypotheses on the forces that have acted on a particular region to create its structure and also make predictions about future changes to the planet. All of these steps in exploring the Earth involve the acquisition and analysis of numerical data. An earth scientist therefore needs to have a firm understanding of statistical and numerical methods as well as the ability to utilize relevant computer software packages, in order to be able to analyze the acquired data.

This book introduces some of the most important methods of data analysis employed in earth sciences and illustrates their use through examples using the MATLAB® software package. These examples can then be used as recipes for the analysis of the reader’s own data, after having learned their application with synthetic data. This introductory chapter deals with data acquisition (Section 1.2), the various types of data (Section 1.3) and the appropriate methods for analyzing earth science data (Section 1.4). We therefore first explore the characteristics of typical data sets and subsequently investigate the various ways of analyzing data using MATLAB.

1.2 Data Collection

Most data sets in earth sciences have a very limited sample size and also contain a significant number of uncertainties. Such data sets are typically used to describe rather large natural phenomena, such as a granite body, a large landslide or a widespread sedimentary unit. The methods described in this book aim to find a way of predicting the characteristics of a larger population from a much smaller sample (Fig. 1.1). An appropriate sampling strategy is the first step towards obtaining a good data set. The development
Fig. 1.1 Samples and populations. Deep valley incision has eroded parts of a sandstone unit (hypothetical population). The remaining 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.

of a successful strategy for field sampling requires decisions on the sample size and the spatial sampling scheme.

The sample size includes the sample volume, the sample weight and the number of samples collected in the field. The sample weights or volumes can be critical factors if the samples are later analyzed in a laboratory and most statistical methods also have a minimum requirement for the sample size. The sample size also affects the number of subsamples that can be collected from a single sample. If the population is heterogeneous then the sample needs to be large enough to represent the population’s variability, but on the other hand samples should be as small as possible in order to minimize the time and costs involved in their analysis. The collection of smaller pilot samples is recommended prior to defining a suitable sample size.

The design of the spatial sampling scheme is dependent on the availabil-
ity of outcrops or other material suitable for sampling. Sampling in quarries typically leads to clustered data, whereas sampling along road cuts, shoreline cliffs or steep gorges results in one-dimensional traverse sampling schemes. A more uniform sampling pattern can be designed where there is 100% exposure or if there are no financial limitations. 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. Although these sampling schemes might be expected to provide superior methods for sampling collection, evenly-spaced sampling locations tend to miss small-scale variations in the area, such as thin mafic dykes within a granite body or the spatially-restricted occurrence of a fossil (Fig. 1.2).

The correct sampling strategy will depend on the objectives of the investigation, the type of analysis required and the desired level of confidence in the results. Having chosen a suitable sampling strategy, the quality of the sample can be influenced by a number of factors resulting in the samples not being truly representative of the larger population. Chemical or physical alteration, contamination by other material or displacement by natural and anthropogenic processes may all result in erroneous results and interpretations. It is therefore recommended that the quality of the samples, the method of data analysis employed and the validity of the conclusions drawn from the analysis be checked at each stage of the investigation.

1.3 Types of Data

Most earth science data sets 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 in which the intervals along the scale are irregularly spaced. Mohs’ hardness scale is one example of an ordinal scale. The hard-
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 uniformly-distributed $xy$ coordinates, d clustered sampling constrained by limited access in a quarry, and e traverse sampling along road cuts and river valleys.
Fig. 1.3 Types of earth science data. a Nominal data, b ordinal data, c ratio data, d interval data, e closed data, f spatial data, and g directional data. All of these data types are described in this book.
ness value indicates the material’s resistance to scratching. Diamond has a hardness of 10, whereas the value for talc is 1, but 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, which attempts to categorize the effects of earthquakes, is another example of 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 offering a great advantage over ordinal data. The zero point is the natural termination of the data scale, and this type of data allows for either discrete or continuous data sampling. Examples of such data sets include length or weight data.

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

In addition to these standard data types, earth scientists frequently encounter special kinds of data such as

- **closed data** – These data are expressed as proportions and add up 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 in thickness of a sandstone bed and the distribution of tracer concentrations in groundwater are examples of this type of data, which 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 bedding, the orientation of elongated fossils or the flow direction of lava. This is another very common type of data in earth sciences.

Most of these different types of data require specialized methods of analysis, which are outlined in the next section.
1.4 Methods of Data Analysis

Data analysis uses precise characteristics of small samples to hypothesize about the general phenomenon of interest. Which particular method is used to analyze the data depends on the data type and the project requirements. The various methods available include:

- **Univariate methods** – Each variable is assumed to be independent of the others, and is explored individually. 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 values, the average, and the dispersion about the average. Examples are the sodium content of volcanic glass shards that have been affected by chemical weathering, or the sizes of snail shells within 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 may be used to find an equation that describes the relationship between the two variables. An example of a bivariate plot is the Harker Diagram, which is one of the oldest methods of visualizing geochemical data from igneous rocks and simply plots oxides of elements against SiO$_2$.

- **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) component and an irregular (random, stochastic) component. A widely used technique to describe cyclic components of a time series is that of spectral analysis. Examples of the application of these techniques include the investigation of cyclic climatic 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 in order to correct all kinds of unwanted distortions or to separate various components of interest. It includes the design and realization of filters, and their application 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** – This is the analysis of parameters in 2D or 3D space and hence 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. Also included are 2D and 3D interpolation techniques, which 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 involve importing and exporting, compressing and decompressing, and displaying images. Image processing also aims to enhance images for improved intelligibility, and to manipulate images in order to increase the signal-to-noise ratio. Advanced techniques are used to extract specific features, or analyze shapes and textures, such as for counting mineral grains or fossils in microscope images. Another important application of image processing is in the use of satellite remote sensing to map certain types of rocks, soils and vegetation, as well as other parameters such as soil moisture, rock weathering and erosion.

• **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 of these 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 usage is in the comparison of species assemblages in ocean sediments for the reconstruction of 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 slickensides (or striae) on a fault plane, circular statistical methods are common in paleomagnetic studies, and microstructural investigations include the analysis of grain shapes and quartz c-axis orientations in thin sections.

Some of these methods of data analysis require the application of numerical methods such as interpolation techniques. While the following text
deals mainly with statistical techniques it also introduces several numerical methods commonly used in earth sciences.

**Recommended Reading**


