1
Introduction

Statistical data analysis is about studying data – graphically or via more formal methods. Exploratory Data Analysis (EDA) techniques (Tukey, 1977) provide many tools that transfer large and cumbersome data tabulations into easy to grasp graphical displays which are widely independent of assumptions about the data. They are used to "visualise" the data. Graphical data analysis is often criticised as non-scientific because of its apparent ease. This critique probably stems from many scientists trained in formal statistics not being aware of the power of graphical data analysis.

Occasionally, even in graphical data analysis mathematical data transformations are useful to improve the visibility of certain parts of the data. A logarithmic transformation would be a typical example of a transformation that is used to reduce the influence of unusually high values that are far removed from the main body of data.

Graphical data analysis is a creative process, it is far from simple to produce informative graphics. Among others, choice of graphic symbols, and data subsets are crucial ingredients for gaining an understanding of the data. It is about iterative learning, from one graphic to the next until an informative presentation is found, or as Tukey (1977) said "It is important to understand what you can do before you learn to measure how well you seem to have done it".

However, for a number of purposes graphics are not sufficient to describe a given data set. Here the realms of descriptive statistics are entered. Descriptive statistics are based on model assumptions about the data and thus more restrictive than EDA. A typical model assumption used in descriptive statistics would be that the data follow a normal distribution. The normal distribution is characterised by a typical bell shape (see Figure 4.1 upper left) and depends on two parameters, mean and variance (Gauss, 1809). Many natural phenomena are described by a normal distribution. Thus this distribution is often used as the basic assumption for statistical methods and estimators. Statisticians commonly assume that the data under investigation are a random selection of many more possible observations that altogether follow a normal distribution. Many formulae for statistical calculations, e.g., for mean, standard deviation and correlation are based on a model. It is always possible to use the empirical data at hand and the given statistical formula to calculate "values", but only if the data follow the model will the values be representative, even if another random sample is taken. If the distribution of the samples deviates from the shape of the model distribution, e.g., the bell shape of the normal distribution, statisticians will often try to use transformations that force the data to approach a normal

Statistical Data Analysis Explained Clemens Reimann, Peter Filzmoser, Robert G. Garrett, Rudolf Dutter

distribution. For environmental data a simple log-transformation of the data will often suffice to approach a normal distribution. In such a case it is said that the data come from a lognormal distribution.

Environmental data are frequently characterised by exceptionally high values that deviate widely from the main body of data. In such a case even a data transformation will not help to approach a normal distribution. Here other statistical methods are needed, that will still provide reliable results. Robust statistical procedures have been developed for such data and are often used throughout this book.

Inductive statistics is used to test hypotheses that are formulated by the investigator. Most methods rely heavily on the normal distribution model. Other methods exist that are not based on these model assumptions (non-parametric statistical tests) and these are often preferable for environmental data.

Most data sets in applied earth sciences differ from data collected by other scientists (e.g., physicists) because they have a spatial component. They present data for individual specimens, termed as samples by earth scientists, that were taken somewhere on Earth. Thus, in addition to the measured values, for example geochemical analyses, the samples have spatial coordinates. During data analysis of environmental and earth science research results this spatial component is often neglected, to the detriment of the investigation. At present there exist many computer program systems either for data analysis, often based on classical statistics that were developed for physical measurements, or for "mapping" of spatial data, e.g., geographical information systems (GIS). For applied earth sciences a data analysis package that takes into account the special properties of spatial data and permits the inclusion of "space" in data analysis is needed. Due to their spatial component, earth science and environmental data have special properties that need to be identified and understood prior to statistical data analysis in order to select the "right" data analysis techniques. These properties include:

- The data are spatially dependent (the closer two sample sites are the higher the probability that the samples show comparable analytical results) – all classical statistical tests assume independence of individuals.
- At each sample site a multitude of different processes can have had an influence on the measured analytical value (e.g., for soil samples these include: parent material, topography, vegetation, climate, Fe/Mn-oxyhydroxides, content of organic material, grain size distribution, pH, mineralogy, presence of mineralisation or contamination). For most statistical tests, however, it is necessary that the samples come from the same distribution – this is not possible if different processes influence different samples in different proportions. A mixture of results caused by many different underlying processes may mimic a lognormal data distribution – but the underlying "truth" is that the data originate from multiple distributions and should not be treated as if they were drawn from a single normal distribution.
- Like much scientific data (consider, for example, data from psychological investigations), applied earth sciences data are imprecise. They contain uncertainty. Uncertainty is unavoidably introduced at the time of sampling, sample preparation and analysis (in psychological investigations some people may simply lie). Classical statistical methods call for "precise" data. They will often fail, or provide wrong answers and "certainties", when applied to imprecise data. In applied earth sciences the task is commonly to optimally "visualise" the results. This book is all about "visualisation" of data behaviour.
- Last but not least environmental data are most often compositional data. The individual variables are not independent of each other but are related by, for example, being expressed as a

percentage (or parts per million – ppm (mg/kg)). They sum up to a constant, e.g., 100 percent or 1. To understand the problem of closed data, it just has to be remembered how percentages are calculated. They are ratios that contain all variables that are investigated in their denominator. Thus, single variables of percentage data are not free to vary independently. This has serious consequences for data analysis. Possibilities of how to deal with compositional data and the effect of data closure are discussed in Chapter 10 (for an in depth discussion see Aitchison, 1986, 2003; Buccianti *et al.*, 2006).

The properties described above do not agree well with the assumptions of "classical" (Gaussian) statistics. These are, however, the statistical methods taught in all basic statistics courses at universities because they are the most fundamental statistical methods. As a consequence, up to this day techniques that are far from ideal for the data at hand are widely applied by earth scientists in data analysis. Rather, applied earth science data call for the use of robust and non-parametric statistical methods. Instead of "precise" statistics so called "simple" exploratory data analysis methods, as introduced by Tukey in 1977, should always be the first choice. To overcome the problem of "closed data" the data array may have to be opened (see Section 10.5) prior to any data analysis (note that even graphics can be misleading when working with closed data). However, working with the resulting ratios has other severe shortcomings and at present there is no ideal solution to the problems posed by compositional data. All results based on correlations should routinely be counterchecked with opened data. Closure cannot be overcome by not analysing the major components in a sample or by not using some elements during data analysis (e.g. by focussing on trace elements rather than using the major elements). Even plotting a scatterplot of only two variables can be severely misleading with compositional data (compare Figures 10.7 and 10.8).

Graphical data analyses were largely manual and labour intensive 30 or 40 years ago, another reason why classical statistical methods were widely used. Interactive graphical data analysis has become widely available in these days of the personal computer – provided the software exists to easily prepare, modify and finally store the graphics. To make full use of exploratory data analysis, the software should be so powerful and convenient that it becomes fun to "play" with the data, to look at the data in all kinds of different graphics until one intuitively starts to understand their message. This leads to a better understanding about what kind of advanced statistical methods can and should (or rather should not) be applied to the data at hand. This books aims at providing such a package, where it becomes easy and fun to "play" with spatial data and look at them in many different ways in a truly exploratory data analysis sense before attempting more formal statistical analyses of the data. Throughout the book it is demonstrated how more and more information is extracted from spatial data using "simple" graphical techniques instead of advanced statistical calculations and how the "spatial" nature of the data can and should be included in data analysis.

Such a data analysis and mapping package for spatial data was developed 20 years ago under the name "DAS" (Data Analysis System – Dutter *et al.*, 1990). A quite similar system, called IDEAS, was used at the Geological Survey of Canada (GSC) (Garrett, 1988). DAS was only available for the DOS environment. This made it more and more difficult to run for people used to the Microsoft Windows operating system. During recent years, R has developed into a powerful and much used open source tool (see: `http://www.r-project.org/`) for advanced statistical data analysis in the statistical community. R could actually be directly used to produce all the tables and graphics shown in this book. However, R is a command-line language, and as such it requires more training and experience than the average non-statistician will usually have, or be willing to invest in gaining. The R-scripts for all the graphics and tables

shown in this book are provided and can be used to learn R and to produce these outputs with the reader's own data. The program package accompanying this book provides the link between DAS and R and is called DAS+R. It uses the power of R and the experience of the authors in practical data analysis of applied geoscience data. DAS+R allows the easy and fast production of tables, graphics and maps, and the creation and storage of data subsets, through a graphical user interface that most scientists will find intuitive and be able to use with very little training. R provides the tools for producing and changing tables, graphics and maps and, if needed, the link to some of the most modern developments in advanced statistical data analysis techniques.

To demonstrate the methods, and to teach a user not trained to think "graphically" in data analysis, an existing multidimensional data set from a large environmental geochemical mapping project in the European Arctic, the Kola Ecogeochemistry Project (Reimann *et al.*, 1998a), is used as an example. The Kola Project data include many different sample materials, more than 60 chemical elements were determined, often by several different analytical techniques. The book provides access to the data set and to the computer scripts used to produce the example statistical tables, graphics and maps. It is assumed that the reader has a spreadsheet package like Microsoft Excel™ or Quattro Pro™ installed and has the basic knowledge to effectively use a spreadsheet program.

The Kola Project data set is used to demonstrate step by step how to use exploratory data analysis to extract more and more information from the data. Advantages and disadvantages of certain graphics and techniques are discussed. It is demonstrated which techniques may be used to good effect, and which should better be avoided, when dealing with spatial data. The book and the software can be used as a textbook for teaching exploratory data analysis (and many aspects of applied geochemistry) or as a reference guide to certain techniques. The program system can be used with the reader's own data, and the book can then be used as a "handbook" for graphical data analysis. Because the original R scripts are provided for all tables, graphics, maps, statistical tests, and more advanced statistical procedures, the book can also be used to become familiar with R programming procedures.

Because many readers will likely use the provided software to look at their own data rather than to study the Kola Project data, the book starts out with a description of the file structure that is needed for entering new data into R and DAS+R (Chapter 2). Some common problems encountered when editing a spreadsheet file as received from a laboratory to the DAS+R (or R) format are discussed in the same chapter. A selection of graphics for displaying data distributions are introduced next (Chapter 3), before the more classical distribution measures are introduced and discussed (Chapter 4). The spatial structure of applied earth science data should be an integral part of data analysis and thus spatial display, mapping, of the data is discussed next (Chapter 5) and before further graphics used in exploratory data analysis are introduced (Chapter 6). A "classical" task in the analysis of applied geochemical data is the definition of background and threshold, coupled with the identification of "outliers" and of element sources – techniques used up to this time are discussed in their own chapter (7). A key component of exploratory data analysis lies in comparing data. The use of data subsets is an especially powerful tool. The definition of subsets of data and using a variety of graphics for comparing them are introduced in Chapter 8, while Chapter 9 covers the more formal statistical tests. Statistical tests often require that the data are drawn from a normal distribution. Many problems can arise when using formal statistics with applied earth science data that are not normally distributed or drawn from multiple statistical populations. Chapter 10 covers techniques that may be used to improve the data behaviour for statistical analysis as a preparation for entering the realms of multivariate data analysis. The following chapters cover some of

the widely used multivariate techniques such as correlation analysis (Chapter 11), multivariate graphics (Chapter 12), multivariate outlier detection (Chapter 13), principal component and factor analysis (Chapter 14), cluster analysis (Chapter 15), regression analysis (Chapter 16) and discriminant analysis (Chapter 17). In all chapters the advantages and disadvantages of the methods as well as the data requirements are discussed in depth. Chapter 18 covers different aspects of an integral part of collecting data in applied earth sciences: quality control. One could argue that Chapter 18 should be at the front of this book, due to its importance. However, quality control is based on graphics and statistics that needed to be introduced first and thus it is treated in Chapter 18, notwithstanding the fact that it should be a very early consideration when designing a new project. Chapter 19 provides an introduction to R and the R-scripts used to produce all diagrams and tables in this book. The program system and graphical user interface under development to make R easily accessible to the non-statistician is also explained.

The following books can be suggested for further reading. Some cover "graphical statistics", some the interpretation of geoscience and environmental data, some give an introduction to the computer language S (the base of R) and the commercial software package S-Plus (comparable to R), and several provide the mathematical formulae that were consciously avoided in this book. Davis (1973, 2002) is still one of the "classic" textbooks about computerised data analysis in the geosciences. Tukey (1977) coined the term "Exploratory Data Analysis" (EDA) and introduced a number of powerful graphics to visualise data (e.g., the boxplot). Velleman and Hoaglin (1981) provide an introduction to EDA including the computer codes for standard EDA methods (Fortran programs). Chambers *et al.* (1983) contains a comprehensive survey of graphical methods for data analysis. Rollinson (1993) is a classical textbook on data analysis in geochemistry, the focus is on interpretation rather than on statistics. Rock (1988), and Helsel and Hirsch (1992) provide an excellent compact overview of many statistical techniques used in the earth sciences with an early focus on robust statistical methods. Cleveland's papers (1993, 1994) are general references for visualising and graphing data. Millard and Neerchal (2001) provide an extensive introduction to environmental statistics using S-Plus. Venables and Ripley (2002) explain a multitude of statistical methods and their application using S. Murrell (2006) provides an excellent and easy to read description of the graphics system in R.

1.1 The Kola Ecogeochemistry Project

The Kola Ecogeochemistry Project (web site `http://www.ngu.no/Kola`) gathered chemical data for up to more than fifty chemical elements from four different primary sample materials (terrestrial moss, and the O-, B-, and C-horizon of podzolic soils) in parts of northern Finland, Norway and Russia. Two additional materials were collected for special purposes (Topsoil: 0–5 cm, and lake water – the latter in the Russian survey area only). The size of the survey area in the European Arctic (Figure 1.1) was 188 000 km^2. The four primary materials were collected because in combination they can reflect atmospheric input (moss), interactions of the biosphere with element cycles (moss and O-horizon), the atmosphere–biosphere–lithosphere interplay (O-horizon), the influence of soil-forming processes (B-horizon), and the regional geogenic background distribution (the lithosphere) (C-horizon) for the elements investigated. Topsoil was primarily collected for the determination of radionuclides, but later a number of additional parameters were determined. Lake water reflects the hydrosphere, samples were

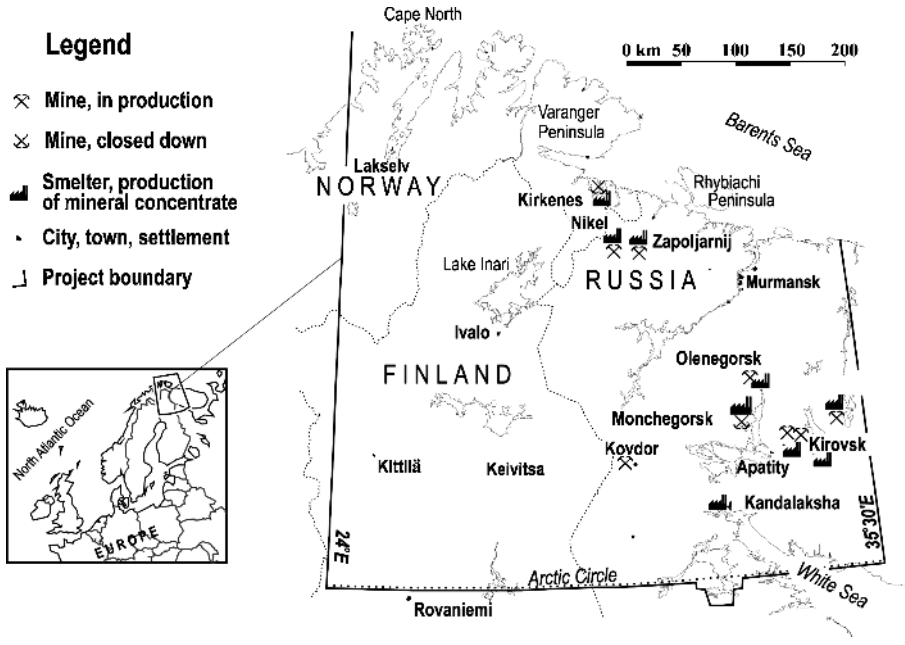

Figure 1.1 Location of the Kola Ecogeochemistry Project survey area

collected in Russia only because the 1000 lakes project (Henriksen *et al.*, 1996, 1997) collected lake water samples over all of Scandinavia at the same time (1995). All results for the four primary sample materials and topsoil are documented in the form of a geochemical atlas (Reimann *et al.*, 1998a). Lake water geochemistry is documented in a number of publications in international journals (see, e.g., Reimann *et al.*, 1999a, 2000a).

The main aim of the project was the documentation of the impact of the Russian nickel industry on the vulnerable Arctic environment. The result was a database of the concentration of more than 50 chemical elements in the above sample materials, reflecting different compartments of the ecosystem, in the year 1995. Each material can be studied for itself, the main power of the project design, however, lies in the possibility to directly compare results from all the different sample materials at the same sites.

This book provides access to all the regional Kola data, including topsoil and lake waters. Throughout the book examples were prepared using these data and the reader is thus able to reproduce all these diagrams (and many others) by using DAS+R and the Kola data. There are certainly many features hidden in the data sets that have not yet been covered by publications. Feel free to use the data for your own publications, but if a fellow scientist wants to use these data for publications, due reference should be given to the original source of the data (Reimann *et al.*, 1998a).

1.1.1 Short description of the Kola Project survey area

The survey area is described in detail in the geochemical atlas (Reimann *et al.*, 1998a). This book also provides a multitude of maps, which can be helpful when interpreting the data.

The project covered the entire area north of the Arctic Circle between longitudes 24° and 35.5° east and thence north to the Barents Sea (Figure 1.1). Relative to most of Europe, the Finnish and Norwegian parts of the area are still almost pristine. Human activities are mostly limited to fishery, reindeer-herding and forestry (in the southern part of the project area). Exceptions are a large iron ore mine and mill at Kirkenes, N-Norway; a small, brown coal-fired power station near Rovaniemi at the southern project border in Finland and some small mines. Population density increases gradually from north to south. In contrast, the Russian part of the project area is heavily industrialised with the nickel refinery at Monchegorsk, the nickel smelter at Nikel and the Cu/Ni-ore roasting plant at Zapoljarnij, which are three of the world's largest point-source emitters of SO_2 and Cu, Ni and Co and a number of other metals. These three sources together accounted for emissions of 300 000 t SO_2, 1900 t Ni and 1100 t Cu in 1994 (Reimann *et al.*, 1997c). Apatite ore is mined and processed near Apatity, iron ore at Olenegorsk and Kovdor, Cu-Ni-ore near Zapoljarnij. An aluminium smelter is located near Kandalaksha. The major towns of Murmansk and Apatity have large oil- and coal-fired thermal heating and power plants.

Topographically, large parts of the area can be characterised as highlands. In Norway, the general landscape in the coastal areas is quite rugged, and the mountains reach elevations of 700 m above sea level (a.s.l.). In Russia, in the south-western part of the Kola Peninsula, there are mountains reaching 200–500 m a.s.l. Near Monchegorsk and Apatity and near the coast of the White Sea there are some higher mountains (over 1000 m a.s.l.).

The geology of the area is complex and includes a multitude of different bedrock types (Figure 1.2). Some of the rock types occurring in the area are rare on a global scale and have unusual geochemical compositions. The alkaline intrusions that host the famous apatite deposits near Apatity are an example. The main rock types in the area that are occasionally mentioned in later chapters due to their special geochemical signatures, are:

- Sedimentary rocks of Caledonian (c. 1600–400 million years (Ma) old) and Neoproterozoic (1000–542 Ma) age that occur along the Norwegian coast and on the Rhybachi Peninsula in Russia (lithologies 9 and 10 in the data files).
- The rocks of the granulite belt that runs from Norway through northern Finland into Russia. These rocks are of Archean age (2300–1900 Ma) and "foreign" to the area (see Reimann and Melezhik, 2001). They are subdivided into "felsic" and "mafic" granulites (lithologies 31 and 32 in the data files).
- Diverse greenstone belts, which occur throughout the area. These are Palaeoproterozoic (2400–1950 Ma) rocks of volcanic origin (lithologies 51 and 52 in the data files). These rocks host many of the ore occurrences in the area, e.g., the famous Cu-Ni-deposits near Zapoljarnij in Russia.
- Alkaline and ultramafic alkaline intrusions of Palaeoproterozoic to Palaeozoic age (1900–470 Ma) (see above). These host the important phosphate deposits near Apatity (lithologies 81, 82 and 83 in the data files).
- Granitic intrusions of Palaeoproterozoic age (1960–1650 Ma), occurring in the south-western corner of the survey area in Finland and as small bodies throughout the survey area (lithology 7 in the data files).
- Large gneiss masses of Archean (3200–2500 Ma) and uncertain age (lithologies 1, 4 and 20 in the data files) that do not show any geochemical peculiarities.

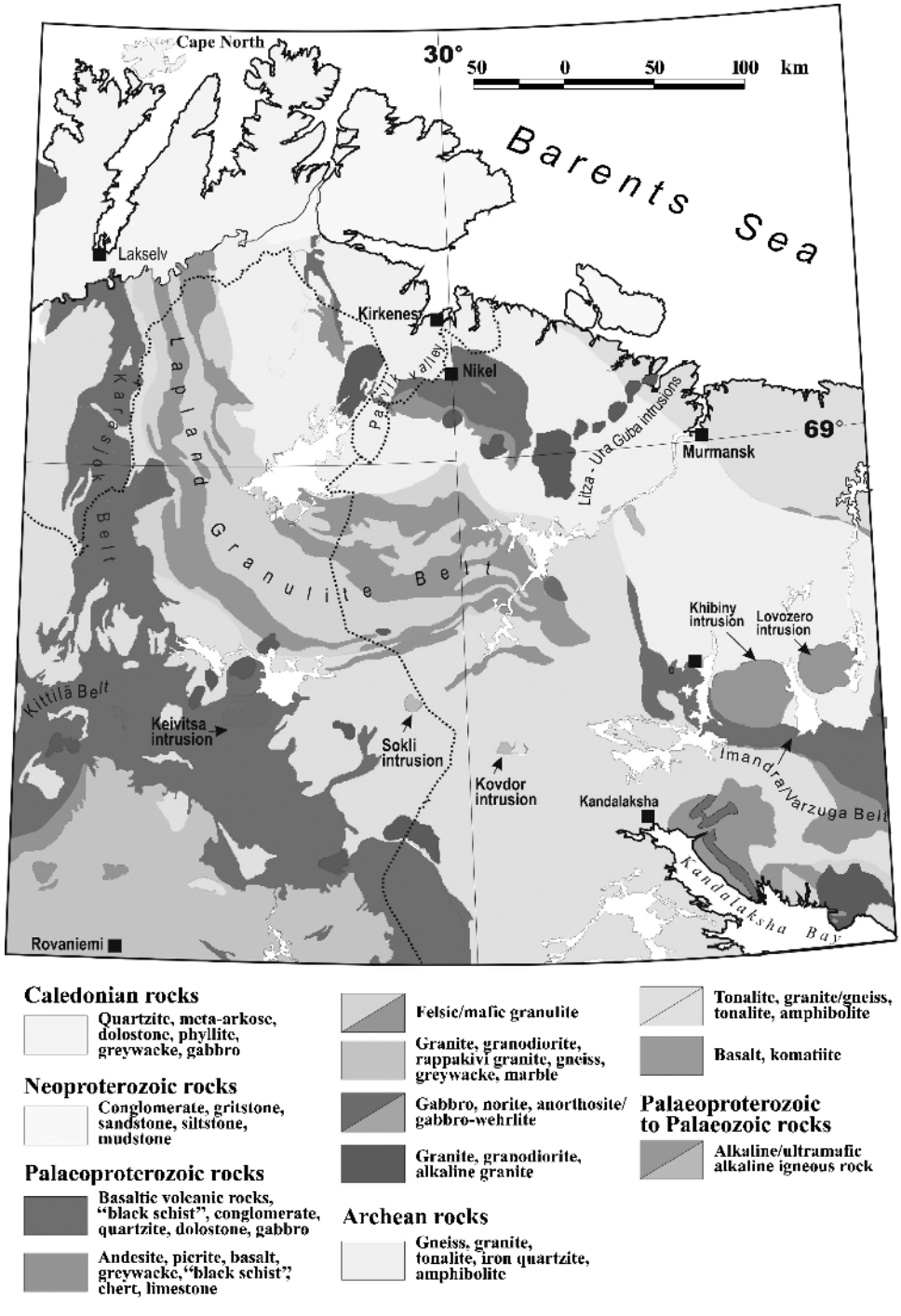

Figure 1.2 Geological map of the Kola Project survey area (modified from Reimann et al., 1998a). A colour reproduction of this figure can be seen in the colour section, positioned towards the centre of the book

The study area is part of the glaciated terrain of Northern Europe. The main Quaternary deposits are till and peat. There are also large areas without any surficial cover, dominated by outcrops and boulder fields (Niemelä *et al.*, 1993).

The north–south extent of the survey area is about 500 km. Within this distance, three vegetation zones gradually replace each other (for a vegetation map in colour see Kashulina *et al.*, 1997; Reimann *et al.*, 1998a). The southern and central parts of the area fall into the northern boreal coniferous zone. Towards the north, this zone gradually gives way to subarctic birch forest, followed by the subarctic tundra zone close to the coast of the Barents Sea. These changes in vegetation zones can also occur with altitude. Major characteristics of the forest ecosystems in this area are the sparseness of the tree layer, a slow growth rate, and the large proportion of ground vegetation in the total biomass production.

The dominant soil-forming process in forested and treeless regions of Northern Europe is podzolisation of mineral soils (Borggaard, 1997). Podzols are thus the most important soil type present throughout the survey area. Soils in the area are young. Their age ranges between 5000 and 8000 years, having formed since the retreat of the continental ice sheet in Northern Europe. A typical podzol profile consists of five main layers, the O (organic), E (eluvial), B (illuvial), BC (transitional) and C (weathered parent)-horizon. Colour photographs of typical podzol profiles from the area can be found in Reimann *et al.* (1998a) or on the Kola Project web site: `http://www.ngu.no/Kola/podzol.html`. The O-horizon of podzol is characterised by a low pH-value (MEDIAN in the survey area 3.85 – Reimann *et al.*, 1998a). pH increases systematically with depth. In the C-horizon it reaches a MEDIAN value of 5.8. The O-horizon varies in thickness from less than 0.5 cm to more than 20 cm, the MEDIAN for the survey area being 2.5 cm. The thickness of the soil profiles can vary considerably – the depth to the top of the C-horizon varies between 10 and 123 cm, the MEDIAN being 35 cm (maps for all these parameters are presented in Reimann *et al.*, 1998a).

Major climatic differences occur within the survey area. In the northwest, summers are cool and winters mild, with most precipitation between September and January (coastal climate). In the central part, warm summers and relatively cold winters are typical; the main precipitation takes the form of rain showers in July and August (continental climate). The easternmost part is similar to the central part, but with colder winters. Precipitation is in general low. The average for the whole area is <500 mm/year. The yearly average temperature is -1°C. Maps showing average rainfall and temperature for summer and winter can be found in Reimann *et al.* (1998a).

1.1.2 Sampling and characteristics of the different sample materials

Sampling of the 188 000 km^2 area took place from July to September 1995. The average sample density was 1 site per 300 km^2. Samples were collected from 617 sites. Depending on availability the total number of samples, n, ranges for any one sample material from 594 (moss) to 617 (O-horizon). Sample site selection for the project was such that only locations where podzol had developed on glacial drift were visited, giving genetically comparable samples over the whole area. At each site the following samples were collected (for more detailed information see Äyräs and Reimann, 1995; Reimann *et al.*, 1998a):

Moss ($n = 594$): Terrestrial moss, preferably the species *Hylocomium splendens*, and if it was absent *Pleurozium schreberi*, was collected. Only shoots representing the previous three years' growth were taken.

Terrestrial moss (*Hylocomium splendens* and *Pleurozium schreberi*) receives most of its nutrients from the atmosphere. In Scandinavia it has been used to monitor the atmospheric deposition of chemical elements for more than 30 years (see, e.g., Rühling and Tyler, 1968, 1973; Tyler, 1970). It was thus included in the Kola Project to reflect the input of elements from the atmosphere (wet and dry deposition, including both geogenic and anthropogenic dust) over the previous 2–3 years. At the same time the moss reflects element concentrations in an important component of the arctic ecosystem in terms of total biomass production and ecological function. It is an important supplier of litter for the formation of the organic horizon of the soils (Kashulina *et al.*, 1997). Results of an interspecies comparison between *Hylocomium splendens* and *Pleurozium schreberi* are presented in Halleraker *et al.* (1998). A review of the moss technique and a discussion of possible problems related to it as well as a comparison with the chemical composition of lichen and crowberry from the Kola Project area are given in Reimann *et al.* (1999b, 2001c). Results of regional mapping are also discussed in Äyräs *et al.* (1997a); de Caritat *et al.* (2001); Reimann *et al.* (1997b).

O-horizon ($n = 617$): This was collected with a custom-built tool to facilitate and systematise sampling, and to allow easy measurement of the volume of each sample (Äyräs and Reimann, 1995). For the O-horizon samples only the top 3 cm of the organic layer were taken. If the total thickness of the O-horizon was less than 3 cm, only the organic layer was sampled, and the thickness was recorded on the field notes. From seven to ten sub-samples were collected at each site to give a composite sample with a minimum volume of one litre.

The O-horizon consists mostly of plant residues in differing stages of decay and humus, almost inevitably mixed with some minerogenic particles. Due to its location and genesis, the organic horizon reflects the complex interplay between the lithosphere, the biosphere and the atmosphere. The O-horizon is a major sink for plant nutrients in northern ecosystems. It can accumulate and enrich many chemical elements, e.g., via organic complexing and the formation of metallo-organic compounds. For many elements, both from natural and anthropogenic sources, it thus acts as a very effective "geochemical barrier" as defined first by Goldschmidt (1937). A separate interpretation of the regional distribution patterns found in the O-horizon from the survey area is given in Äyräs and Kashulina (2000). Other publications discussing special properties of the O-horizon include Reimann *et al.* (1998c, 2000b, 2001a).

Podzol profiles (B-horizon, $n = 609$; C-horizon, $n = 605$): Before sampling a podzol profile, homogeneity of the soil cover was checked over an area of 10×10 m. The exact location of the profile was chosen so that both ground vegetation and micro-topography were representative. The sampling pits were dug by spade to the C-horizon. Samples of the O-, E-, B-, BC- and C-horizons were collected, starting from the bottom to avoid contamination and mixing of the horizons. With the exception of the C-horizon, each layer was sampled over its complete thickness. If there were distinguishable layers within the B-horizon, these were collected in the same ratio as present in the profile. Each sample weighed about 1–1.5 kg, depending on parameters like grain size, mineralogy, and water content. At present only the samples of the B- and C-horizon have been studied, all other samples have been archived in air dried condition for future reference.

The B-horizon is clearly affected by soil-forming processes. Compared to the C-horizon it is relatively enriched in clay minerals, organic matter, and free and organically-bound

amorphous Fe- and Al-oxides and -hydroxides, which have been leached from the upper soil horizons. It is less active than the O-horizon, but can still act as a second "geochemical barrier" for many elements (independent of origin) within the soil profile, e.g., via co-precipitation with the Fe-oxides/-hydroxides. Some soil-forming processes within podzol profiles from the Kola Ecogeochemistry Project area are discussed in Räisänen *et al.* (1997) and Kashulina *et al.* (1998a).

The deepest soil horizon of podzols, the C-horizon, is only slightly influenced by soil-forming processes and sometimes by anthropogenic contamination, and thus mostly reflects the natural, geogenic, element pool and regional variations therein. A geological interpretation of the C-horizon results is given in Reimann and Melezhik (2001b).

Topsoil ($n = 607$): Samples were taken for the analysis of radionuclides. Topsoil samples (0–5 cm) were collected at all sample sites using the same procedure as for the O-horizon. These samples were also analysed by Instrumental Neutron Activation Analysis (INAA) for about 30 elements. These results are not used within this book but provided on the web page `http://www.statistik.tuwien.ac.at/StatDA/R-scripts/data/`.

For reasons of completeness, analytical results of lake water samples collected in the Russian survey area are also included on the web page of the book. These are not used within this book, interested readers can find a description of the sampling procedures and some first interpretations of the results in Reimann *et al.* (1999a).

In addition to the sample media documented in the atlas and provided with this book, snow, rain water, stream water, lake water, ground water, bedrock, organic stream sediments, topsoil 0–5 cm, and overbank sediments were studied during different stages of the project (Äyräs and Reimann, 1995; Äyräs *et al.*, 1997a,b; Boyd *et al.*, 1997; de Caritat *et al.*, 1996a,b, 1997a,b, 1998a,b; Chekushin *et al.*, 1998; Gregurek *et al.*, 1998a,b, 1999a,b; Halleraker *et al.*, 1998; Kashulina *et al.*, 1997, 1998a,b; Niskavaara *et al.*, 1996, 1997; Reimann *et al.*, 1996, 1997a,b,c, 1998c, 1999a,b, 2000a,b; Volden *et al.*, 1997), often on a spatially detailed scale. These data can be used to assist in the interpretation of the observed regional features. They are not provided here but are available from the Norwegian Geological Survey (NGU) upon request. A complete and regularly updated list of publications with data from the Project can be found at the following project website: `http://www.ngu.no/Kola/publist.html`.

1.1.3 Sample preparation and chemical analysis

Detailed descriptions of sample preparation and analysis are provided in Reimann *et al.* (1998a). In general, selection of elements, extractions and detection limits was based on a conscious decision to make all results as directly comparable as possible.

Both the Geological Survey of Finland (GTK) and NGU chemical laboratories and the methods described above for analysing the Kola samples are accredited according to EN 45001 and ISO Guide 25. All methods have been thoroughly validated and trueness, accuracy and precision are monitored continuously. Some of the quality control results are presented in Reimann *et al.* (1998a) – and in the last chapter of this book. The data files include information about the analytical method used and the detection limits reached.