6

Further Graphics for Exploratory Data Analysis

6.1 Scatterplots (xy-plots)

Once there is more than one variable, it is important to study the relationships between the variables. The most frequently used diagram for this purpose is a two-dimensional scatterplot, the xy-plot, where two variables are selected and plotted against one another, one along the x-axis, the other one along the y-axis. Scatterplots can be used in two different ways. Firstly, in a truly "exploratory" data analysis approach, variables can be plotted against one another to identify any unusual structures in the data and to then try to determine the process(es) that cause(s) these features. Secondly, the more "scientific" (and probably more frequently used) way of using scatterplots is to have a hypothesis that a certain variable will influence the behaviour of another variable and to demonstrate this in a plot of variable x versus variable y. When working with compositional data the problem of data closure needs to be considered. It will often be a serious issue whether simple scatterplots show the true relationships between the variables (consult Section 10.5).

Both methods have their merits, the exploratory approach facilitates the identification of unexpected data behaviour. In pre-computer times this approach was not widely used as it was time-consuming with the risk of time and resources expended with no reward. Even with computers many data analysis procedures require considerable effort to draw such displays. The process must be simple and fast so that it becomes possible "to play with the data" and to follow up ideas and hypotheses immediately without any tedious editing of data files before the next plots are displayed.

Figure 6.1 shows a scatterplot for Cu versus Ni in the moss samples from the Kola Project. These two elements are of special interest because approximately 2000 tons of Ni and 1000 tons of Cu are emitted to the atmosphere annually by the Russian nickel industry, situated in Nikel, Zapoljarnij, and Monchegorsk (see Figure 1.1 for locations).

Figure 6.1 reveals that a positive, sympathetic relationship exists between Cu and Ni. However, there are a substantial number of samples with high analytical results for both Cu and Ni that mask the behaviour of the main body of data. Plotting the same diagram with logarithmic axes (Figure 6.1, right), decreases the influence of the high values and increases the visibility of the main body of data. It can now also be seen that the spread (variance) of

Statistical Data Analysis Explained: Applied Environmental Statistics with R. C. Reimann, P. Filzmoser, R. G. Garrett, R. Dutter © 2008 John Wiley & Sons, Ltd. ISBN: 978-0-470-98581-6

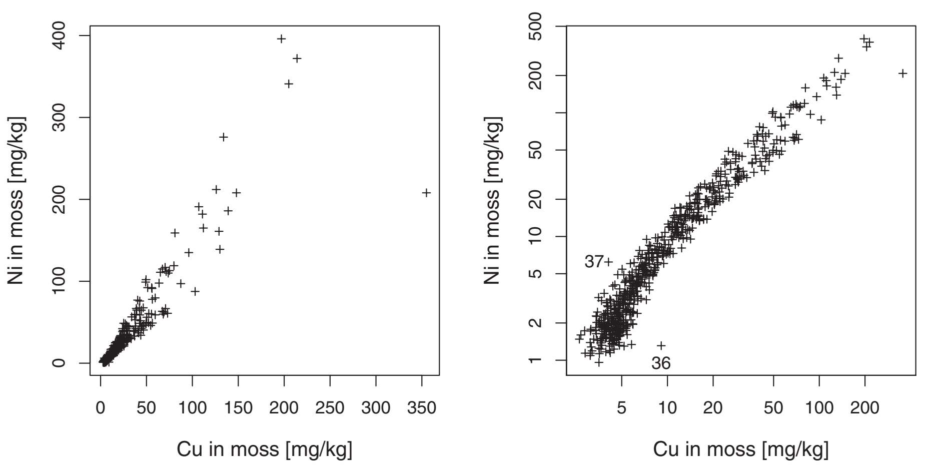

Figure 6.1 Scatterplot of copper (Cu) versus nickel (Ni) in moss samples from the Kola Project. Left plot: original data scale; right plot: log-scale. In the right plot two unusual samples were identified

the data is relatively homogeneous over the whole range of the data. It is also apparent that Cu and Ni are strongly related over the whole range of data. In addition, attention is drawn to a number of unusual samples in both diagrams that deviate from the general trend displayed by the majority of samples. It should be possible to identify these individuals by just "clicking" on them on the screen. Samples like numbers 36 and 37 (Figure 6.1, right) should undergo an additional "quality check" to ensure that they are truly different and that the discordant data are not due to a sample mix-up or poor analytical quality.

6.1.1 Scatterplots with user-defined lines or fields

It is often desirable to include a line or one or several fields in a scatterplot. The simplest case is probably a line indicating a certain proportion between the data plotted along the two axes. One example would be two variables where the same element was analysed with two different analytical techniques. Here the results might be expected to follow a strict one to one relationship. Plotting a 1:1 line onto the diagram allows for the easy detection of any deviations from the expected relationship, and to see at a glance whether the results obtained with one of the methods are higher than expected.

Figure 6.2 shows two such diagrams. In the left plot La was determined with two different analytical techniques. It takes but a glance to notice that although the analyses are similar, by and large, the analytical results obtained with instrumental neutron activation analysis (INAA) are clearly higher than those from an acid extraction. Only INAA provides truly total La-concentrations in the samples and thus this result is no surprise. In the right hand plot, results for Ag in two different soil horizons are depicted (Figure 6.2). The analytical techniques are comparable (both strong acid extractions), but the plot indicates that Ag is highly enriched in the O-horizon and does not show any relation to the C-horizon results. This is surprising; many applied geochemists or environmental scientists would argue that the C-horizon, the mineral or parent soil material, will provide the "geochemical background values" (see Chapter 7) for

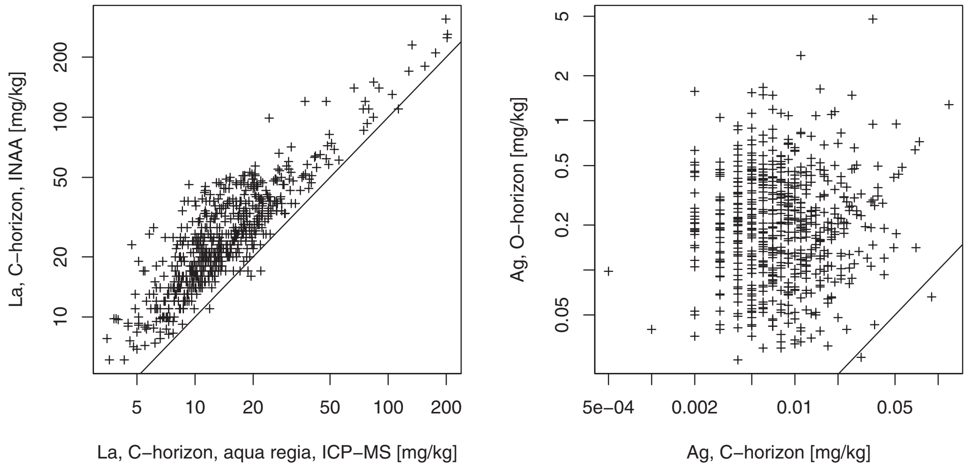

Figure 6.2 Scatterplots with 1:1 line added. Left plot: La determined with two different analytical techniques; right plot: Ag determined with comparable techniques in two different sample materials. Log-scales used for both graphics

the upper soil horizon. The diagram indicates that the relationship is not that simple. One could be tempted to argue that then all the Ag in the O-horizon must be of anthropogenic origin. There is, however, no likely major Ag-emission source in either the survey area or anywhere near. Thus the conclusion must be that there exist one or more processes that can lead to a high enrichment of Ag in the organic layer – independent of the observed concentrations in the soil parent material.

Petrologists frequently use xy-plots where they mark predefined fields to discriminate between different rock types. To include such "auxiliary" background information in a scatterplot is desirable and feasible, but the actual definition of the field limits in the diagrams is a highly specialised task. One software package, based on R, that can plot these lines and fields for many pre-defined plots used by petrologists is freely available at `http://www.gla.ac.uk/gcdkit/` (Janousek *et al.*, 2006).

6.2 Linear regression lines

The discussion of Figure 6.2 demonstrated that it is informative to study dependencies between two or more variables. In Figure 6.2 a 1:1 relation was a reasonable starting hypothesis and thus the line could be drawn without any further considerations. However, many variables may be closely related but do not follow a 1:1 relation (e.g., Cu and Ni as displayed in Figure 6.1). In such cases it might aid interpretation if it were possible to visualise this relationship. In most cases the first step will be to look for linear rather than non-linear relationships. To study linear relationships between two or more variables, linear regression is a widely used statistical method (see also Chapter 16). The best-known approach for fitting a line to two variables is the least squares procedure. To fit a line, it is important to take a conscious decision which variable is the y-variable, i.e. the variable to be predicted. The least squares procedure determines the line where the sum of squared vertical distances from each y-value to this line is minimised.

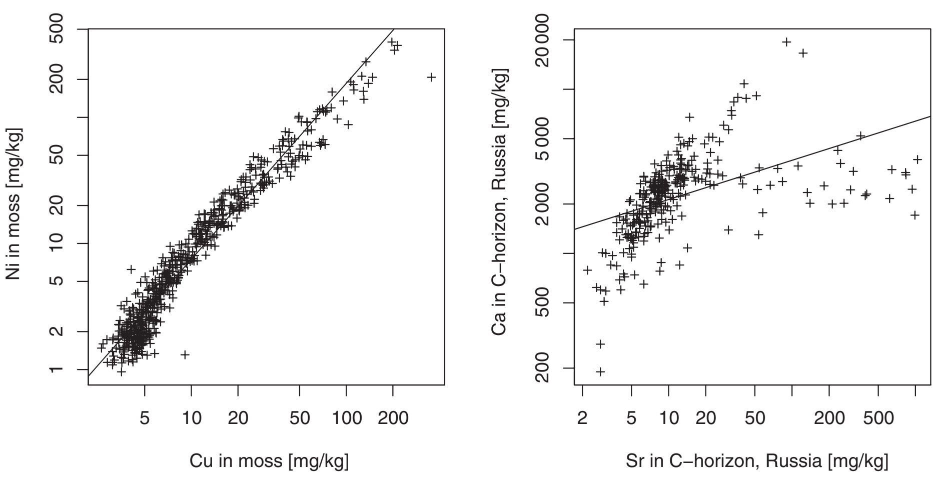

Figure 6.3 Two scatterplots with least squares linear regression lines. Left plot: Cu versus Ni, Kola Project Moss data set with log-scale as used in Figures 6.1 and 6.2; right plot: Sr versus Ca (log-scale), Kola Project C-horizon, Russian samples

This situation can be avoided by using less common procedures for estimating the linear trend, e.g., total least squares or orthogonal least squares regression (Van Huffel and Vandewalle, 1991) which minimise a function of the orthogonal distances of each data point to the line.

Figure 6.3 (left) shows the relationship, regression line, between Cu and Ni in moss (Figure 6.1). When plotting a regression line onto the diagram using the log-scale it is visible that the relation is not really linear but slightly curvilinear. However, a straight line is quite a good approximation of the relationship between the two variables. The other diagram in Figure 6.3 shows the relation between Ca and Sr for the C-horizon samples collected in Russia (log-scale). A number of samples with unusual high Sr-concentrations disturb the linear trend visible for the main body of samples. These samples exert a strong influence (leverage) on the regression line, such that it is drawn towards the relatively few high Sr values. This example demonstrates that the least squares regression is sensitive to unusual data points – in the example the line follows neither the main body of data nor the trend indicated for the outliers (Figure 6.3, right).

It is thus necessary to be able to fit a regression line that is robust against a certain number of outliers to the data. This can be achieved by down-weighting data points that are far away from the linear trend as indicated by the main body of data. The resulting line is called a robust regression line.

Figure 6.4 shows the two plots from Figure 6.3 with both the least squares and a robust regression line. For the left hand plot, Cu versus Ni in moss, both methods deliver almost the same regression line, the difference in the right hand plot (Ca versus Sr in the C-horizon of the Russian samples) is striking. The robust regression line follows the main body of data and is practically undisturbed by the outlying values. For environmental (geochemical) data it is advisable to always use a robust regression line, just like MEDIAN and MAD are better measures of central value and spread than MEAN and SD.

Regression analysis can be used not only to visualise relationships between two variables but also to predict one variable based on the values of one or more other variables (see Chapter 16).

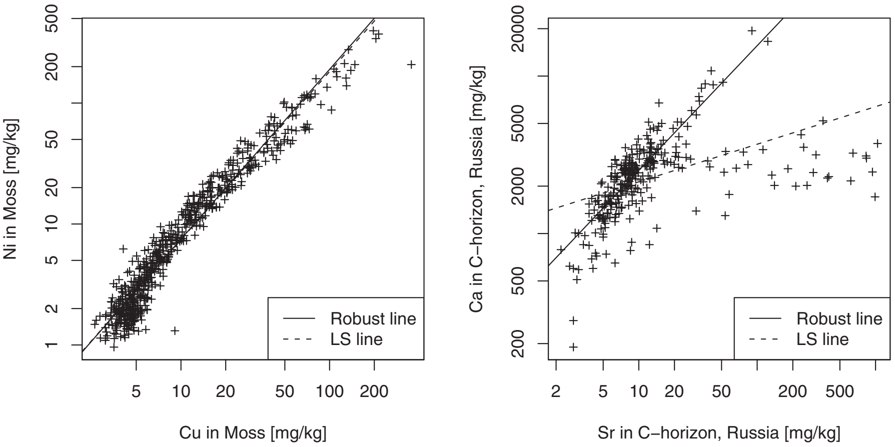

Figure 6.4 Two scatterplots with least squares (LS) and robust regression lines. Left plot: Cu versus Ni, Kola Project Moss data set with log-scale as used in Figure 6.3 (left); right plot: Sr versus Ca (log-scale), Kola Project C-horizon, Russian samples

Relationships between two variables do not need to follow a linear trend. In fact, any nonlinear function could be fitted to the data, or the data points could be simply connected by a line. However, to be able to discern the trend, the function will often need to be smoothed. Many different data-smoothing algorithms do exist. Usually a window of specified width is moved along the x-axis. In the simplest case one central value for the points within the window is calculated (e.g., the MEAN or MEDIAN). When connected, these averages provide a smoothed line revealing the overall data behaviour. Depending on the window width, the resulting line will be more or less smoothed. More refined methods use locally weighted regression within the window.

In environmental sciences the most usual application of smoothed lines will be in studying time or spatial trends as described below (Sections 6.3, 6.4).

6.3 Time trends

In environmental sciences “monitoring” is often an important task; it is necessary to determine whether analytical results change with time. The large test data set used here comes from the regional part of the Kola Project. It permits the study of the distribution of the measured elements in space via mapping (Chapter 5) but not to study changes in element concentration over time. It would become possible if the regional mapping exercise was repeated at certain time intervals. This would be a very expensive undertaking, and it would be more effective to collect the required data in a monitoring exercise at selected locations more frequently than to repeatedly re-map the whole area.

For interpreting the regional Kola data set, more detailed information on local variability versus regional variability and on changes of element concentrations with the season of the year was needed. These data were obtained in a catchment study, which took place one year before the regional mapping program (see, e.g., de Caritat *et al.*, 1996a,b; Boyd *et al.*,1997; Räisänen *et al.*, 1997; Reimann *et al.*, 1997c). For the catchment study eight small catchments were sampled at a density of one site per km^2, and most sample materials were collected at different times through all seasons of one year. Stream water was one of the most intensely studied sample materials, and at some streams water samples were taken once a week throughout a hydrological year (de Caritat *et al.*, 1996a,b).

These data are used here to demonstrate the investigation of time trends in a special form of the scatterplot where the x-axis is the time scale and the measured concentration of diverse variables is plotted along the y-axis. This kind of work requires careful planning as to how the time of sampling is recorded so that the computer software can process it. A simple trick to overcome such problems could be to give all samples a consecutive number according to the date of collection and select the number as the variable for plotting the time axis. If the field sampling has not occurred at regular intervals, this procedure will lead to distortions along the x-axis. Another procedure is to record the date as the "Julian day number", which numbers days consecutively from January 1, 4713 BC and is an internationally agreed-upon system employed by astronomers. Fortunately, R is also able to handle "usual" dates in a variety of formats. To easily discern the trends, a line should be used to link results for a particular variable. Often it is advantageous to be able to estimate a smoothed line running through the results by a variety of different algorithms (see Section 6.4). In the simple case of about 50 water measurements per catchment, spread almost evenly over the year, a simple line directly connecting the results will suffice. However, where multi-year data are acquired, some systematic periodicity, seasonal effect, may be expected and smoothing may help identify such effects.

In Figure 6.5 it is demonstrated how pH and EC change in the course of one year in catchment two, which is in close proximity to the Monchegorsk smelter. A drop in pH and conductivity during May is obvious (Figure 6.5). This indicates the influence of the snow melt period on the stream water samples.

If an abundance of data exists that exhibits high variability, it can also be informative to summarise days, weeks, or months (maybe even years) in the form of boxplot comparisons (see Section 8.4). It is then possible to study differences between the MEDIANS and differences in

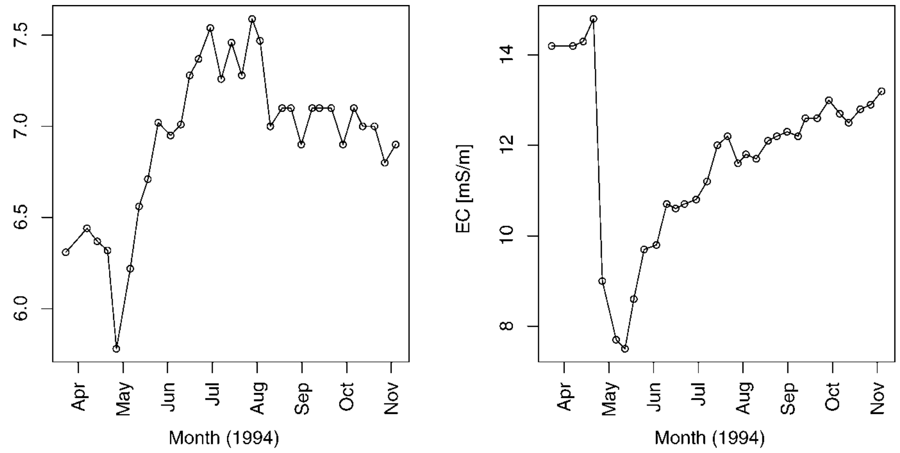

Figure 6.5 Time trends of acidity (pH) and electrical conductivity (EC) in stream water collected on a weekly base in catchment two, Monchegorsk

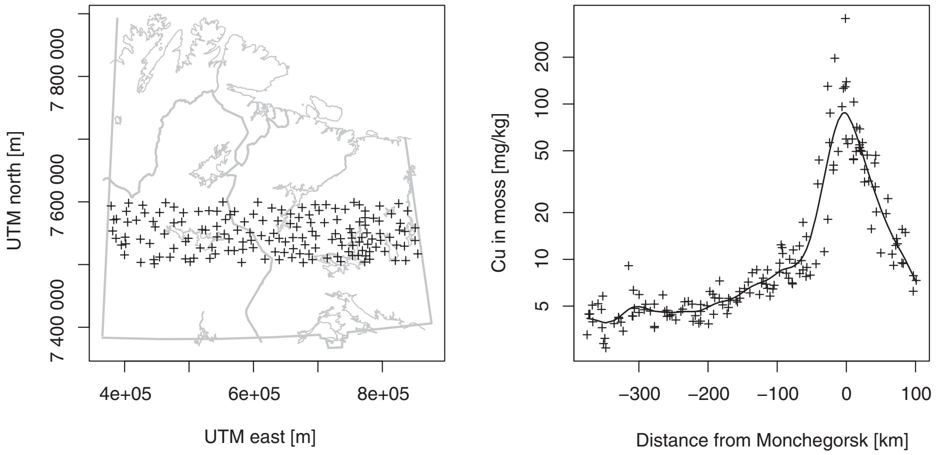

Figure 6.6 Right plot: east–west transect through Monchegorsk, showing the decrease of the concentration of Cu in moss (y-axis, log-scale) with distance from Monchegorsk (x-axis). The map (left plot) shows the location of the samples used to plot the transect

variation (and skewness) at the same time, and trends may be easier to discern when looking at such a graphical summary instead of looking at all the data points and a trend line.

6.4 Spatial trends

Instead of looking at a geochemical map (see Chapter 5), it may also be informative to study geochemical transects, another special case of the xy-plot, where the x-variable is linked to the coordinates of the sample site locations. With the Kola Project data it is, for example, informative to look at an east–west-transect running through Monchegorsk from the eastern project border to the western project border. This facilitates a study of the impact of contamination and the decrease of element concentrations with distance from Monchegorsk. Another interesting study would be a north–south-transect along the western project boundary, in the "background" area, to investigate the impact of the input of marine aerosols at the coast and of the influence of the change of vegetation zones from north to south on element concentrations in moss or soils.

For constructing such a transect, it is necessary to have a tool to define the samples belonging to the transect subset. It is of great advantage if this can be done interactively directly in the map. Such samples could then, for example, be identified as belonging to the subset "EW_Monchegorsk" and "NS_Background". Once the subset is selected, it is easy to study the distribution of all the measured variables along such a transect via selecting the x- (for the east–west-transect) or y-coordinate (for the north–south-transect) for the x-axis of the plot and any other variable for the y-axis and plotting these as xy-plots. In instances where the coordinates of the coastline or of Monchegorsk are known, new variables for the x-axis can be defined "distance from coast" and "distance from Monchegorsk" via a simple subtraction. If the transects are not parallel to the coordinate system, distances can be estimated by plane geometry.

A trend line is then fitted to the data along the *x*-axis. Different smoothing techniques exist for constructing such a line. The basic principle is to choose a window of a certain bandwidth, which is moved along the *x*-axis and relevant statistics computed sequentially. Within the window the points are either averaged or a weighted regression is performed to obtain an average value. All average values are connected by a line. The smoothness of the line is mainly controlled by the chosen bandwidth. A very simple method for smoothing is Tukey's moving median (Tukey, 1977). A more modern and prominent technique, which has been used here, is based on local polynomial regression fitting and known as the "loess" method (Cleveland *et al.*, 1992).

In Figure 6.6 it can easily be seen that Cu values in moss increase by two orders of magnitude when approaching the location of the Monchegorsk smelter. At the same time it is also apparent that the values decrease rapidly with distance from the smelter. At about 200 km from Monchegorsk the contamination signal "disappears" into the natural background variation. At this distance anthropogenic contamination and natural background can no longer be separated using direct chemical analysis. Although the example data display a clear trend even without a smoothed line, the line provides a more informative picture of the overall trend without distortion by local variation.

A north–south transect at the western project border in a pristine area without influence from industry shows other processes (Figure 6.7). The steady input of marine aerosols from the coast of the Barents Sea has an important influence on moss and O-horizon chemistry. Elements like Na and Mg, which are enriched in sea water, show a steady decrease from the coast to a distance of about 200 to 300 km inland (Figure 6.7). This indicates the important influence of a natural process, the steady input of marine aerosols along the coast, on the element concentrations observed in moss and O-horizon samples on a regional scale. An "exploratory" surprise was the opposite trend as detected for Bi (and several other elements – see Reimann *et al.*, 2000b) (Figure 6.7). This trend follows the distribution of the major vegetation zones in the survey area (tundra–subarctic birch forest–boreal forest) and is interpreted as a "bio-productivity" signal: the higher the bio-productivity, the higher the enrichment of certain elements in the O-horizon.

In the examples shown so far, the trends were visible when studying the distribution of the sample points in the figures (Figure 6.6 and 6.7) by eye. Trend lines will be even more interesting when the data are far noisier and trends are not visible at first glance. The distribution of the pH values in the O-horizon samples along the transects provides such an example (Figure 6.8). Along the north–south transect, pH is clearly influenced by distance from the coast. The steady input of marine aerosols along the coast (see Figure 6.8) leads to a replacement of protons in the O-horizon and results in a steadily increasing pH towards the coast. The replaced protons end up in the lakes, resulting in lake water acidification along the coast (Reimann *et al.*, 1999a, 2000a). Along the east–west transect one would expect clear signs of acidification (lower pH) when approaching Monchegorsk with its enormous SO_2 emissions. Figure 6.8 (right) suggests the opposite: pH increases slightly when approaching Monchegorsk, the anthropogenic effect of one of the largest SO_2-emission sources on Earth on the pH in the O-horizon is clearly lower than the natural effect of input of marine aerosols along the coastline. The fact that pH actually increases towards Monchegorsk is explained by co-emissions of alkaline elements like Ca and Mg which more than counteract the acidic effects of the SO_2 emissions. The diagram suggests that it would be a severe mistake to reduce the dust emissions of the smelter before the SO_2 emissions are reduced because then the positive environmental effects of these co-emissions would be lost.

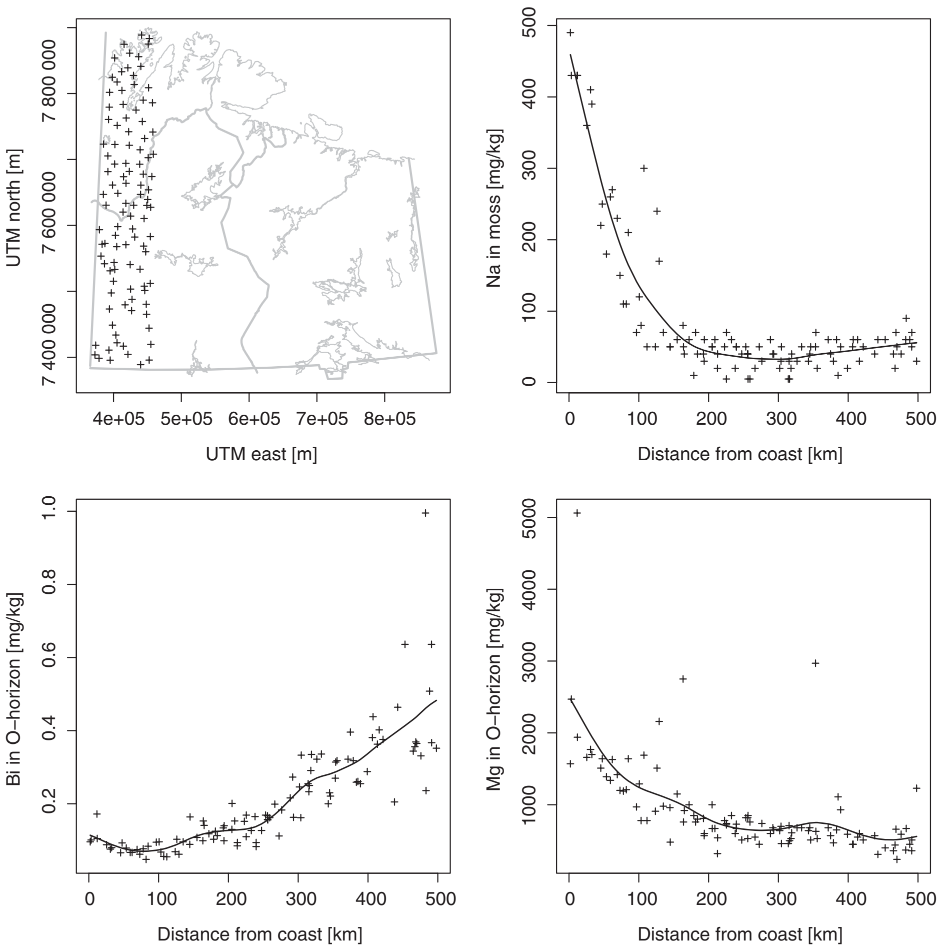

Figure 6.7 North-south transects along the western project border, the most pristine part of the Kola Project survey area. The map (upper left) indicates the samples selected for plotting as transects. Upper right: Na in moss; lower left: Bi in O-horizon soil; and lower right: Mg in O-horizon soil

6.5 Spatial distance plot

There may be cases where it is difficult to define a sensible transect. In such instances it could still be interesting to study systematic changes of a variable with distance from a defined point. This can be done in the spatial distance plot where a subarea (of course the whole survey area can also be used) and a reference point are defined.

The calculated distance of each sample site from the reference point is plotted along the x-axis. Along the y-axis the corresponding concentrations of the variable at the sample sites are plotted. To better recognise a possible trend a smoothed line is fitted to the points.

Figure 6.9 shows an example of such a spatial distance plot using the variable Cu in moss and the location of Nikel/Zapoljarnij as the reference point. The plot clearly demonstrates the

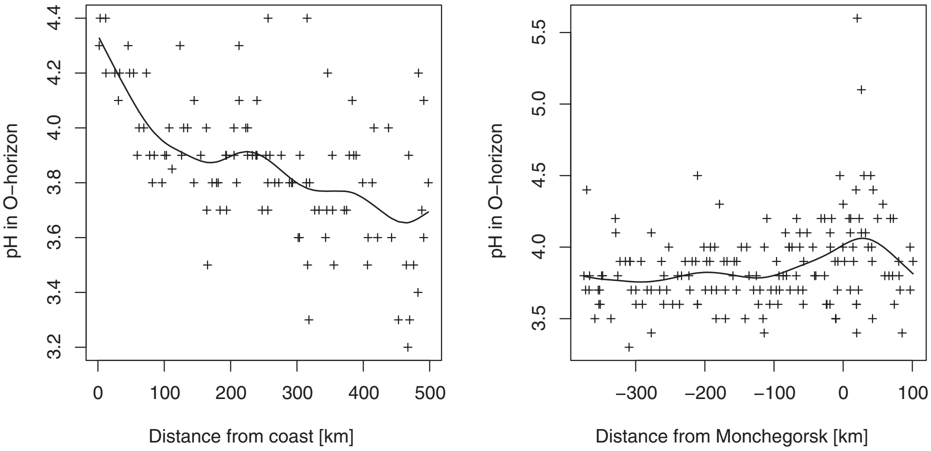

Figure 6.8 pH in the O-horizon. Left plot: north–south transect. Right plot: east–west transect through Monchegorsk

decreasing Cu concentrations with distance from the industrial centre, with background levels being reached at a distance of about 200 km from Nikel/Zapoljarnij.

The spatial distance plot can accommodate any form and any direction of the chosen subarea. It is, for example, easily possible to study the effects of Nikel/Zapoljarnij and Monchegorsk at the same time. When choosing a north–west to south–east running subarea that includes both sites, the reference point is set in the north-western part of the survey area; this results in the plot shown in Figure 6.10. Figure 6.10 shows two peaks of Cu concentrations, the first (left) peak is related to Nikel/Zapoljarnij, the second (right) peak to Monchegorsk. In this plot

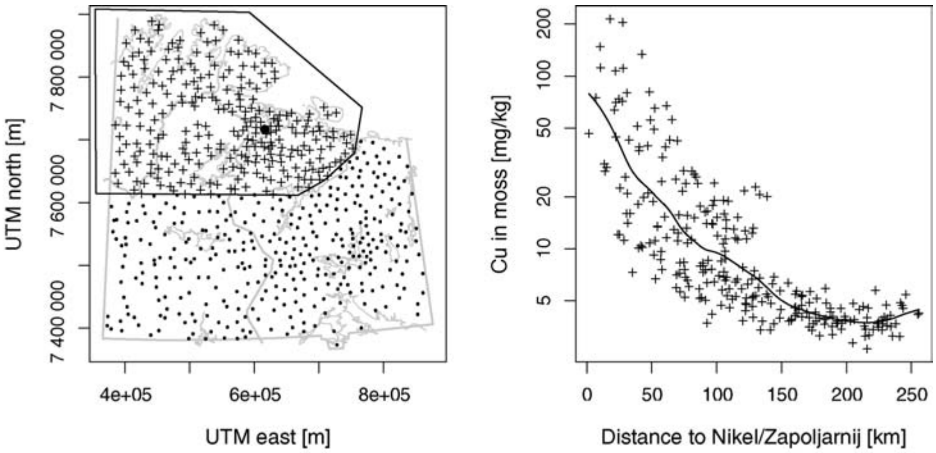

Figure 6.9 Right plot: spatial distance plot for Cu in moss (log-scale). The location of Nikel/Zapoljarnij (dot on left map) is chosen as reference point, the outline indicates the sample sites included in the spatial distance plot (right)

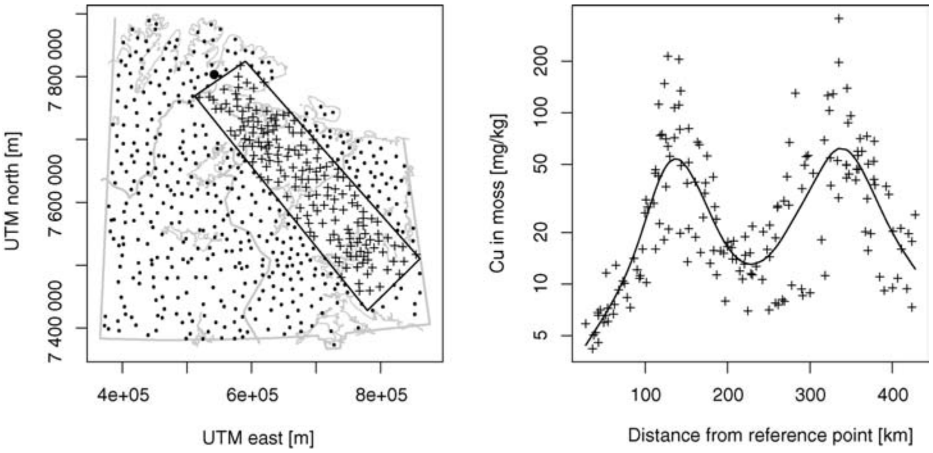

Figure 6.10 Right plot: spatial distance plot for Cu in moss (log-scale). The subarea (left map) includes Nikel/Zapoljarnij and Monchegorsk, and the reference point (dot) is placed in the north-western part of the survey area; the outline indicates the sample sites included in the spatial distance plot (right)

it is clearly visible that the Monchegorsk Cu emissions travel further than the emissions from Nikel/Zapoljarnij, as indicated by the width of the peaks. One likely explanation could be that the emissions at Nikel/Zapoljarnij have a higher proportion of particulates.

6.6 Spiderplots (normalised multi-element diagrams)

Spiderplots, where a selection of elements are plotted along the *x*-axis against the ratio of the element to the concentration of the same element in some reference value (e.g., "chondrite", "mantle", "shale", "upper continental crust") on the *y*-axis, are popular in petrology (see, e.g., Rollinson, 1993). The technique was originally developed to compare the distribution of the Rare Earth Elements (REEs) between different rock types. REEs with an even atomic number show in general higher concentrations in rocks than REEs with uneven atomic numbers (Taylor and McClennan, 1985). Through normalising REEs to a reference value, the resulting saw-blade-pattern can be brought to a more or less straight line by plotting the ratio against the elements. Deviations from the line are then immediately visible. Usually the elements are sorted according to atomic number, but other sorting criteria (e.g., ionic potential or bulk partition coefficient) are also used.

Only a few samples can be plotted before the graphic becomes increasingly unreadable; it is thus not really suited for large data sets. The appearance of the plot will strongly depend on the sequence of elements along the *x*-axis and, of course, on the selection of the set of reference values used for normalisation. Except for the REEs, there are no standardised display orders, a fact that limits the usefulness of these diagrams.

To display the Kola Project results in such a spider plot, the MEDIAN concentration of selected variables per layer (moss, O-, B-, or C-horizon) could be shown. It could be argued at length as to what reference value the results should be compared against. "Upper crust" or "world soil" might be the most appropriate candidates at first glance. However, the average

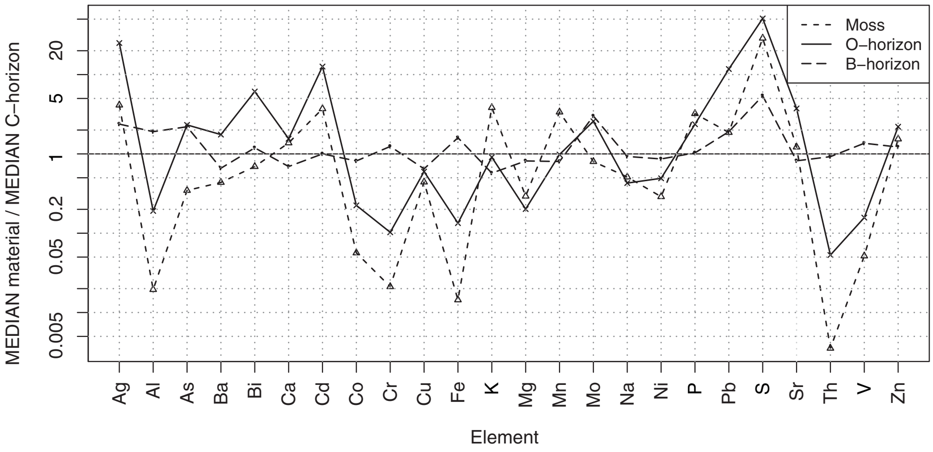

Figure 6.11 Spiderplot showing element concentrations in the B-horizon, O-horizon and moss sample materials relative to the element concentration in the C-horizon from the Kola Project plotted in chemical abbreviation order (from Ag to Zn)

values quoted for "upper crust" or "world soil" are true total concentrations, which are not really comparable to the aqua regia and concentrated HNO_3-extraction data of the Kola Project. Thus there are problems in selecting the most appropriate reference values for the display. The upper soil horizons could be referenced to the C-horizon results from the Kola Project as the best estimate of parent material geochemistry for the survey area (bearing the "cost" of being "only" able to directly compare B-horizon with O-horizon and moss). It would also be possible to use another large-scale data set where comparable analytical techniques were used (e.g., the Baltic Soil Survey data – Reimann *et al.*, 2003). For the example plot the MEDIAN of each variable in moss, O- and B-horizon was divided by the corresponding MEDIAN in the C-horizon. The line at "1" in the plot is the reference line of the C-horizon against which the other sample materials are compared. In the first example (Figure 6.11) the elements have been ordered alphabetically (according to chemical abbreviation from Ag to Zn).

The second example (Figure 6.12) shows the same plot where the elements are ordered according to a steadily decreasing ratio in moss. Although the same data are plotted, the resulting diagram looks very different. Here it is easier to compare the differences of the other two sample media (B- and O-horizon) from the two extremes, C-horizon and moss.

It is immediately apparent that the average composition of the B-horizon – with the exception of very few elements – closely follows that of the C-horizon. Furthermore, it is clear that a number of elements show unusually high concentrations in the moss, while others have much lower values in the moss than in the C-horizon. It is also quite easy to identify those elements that show a different behaviour in the O-horizon than they do for the moss or B-horizon.

6.7 Scatterplot matrix

The scatterplot matrix is a collection of scatterplots of all selected variables against each other (Figures 6.13 and 6.14). The scatterplot matrix is one of the most powerful tools in graphical

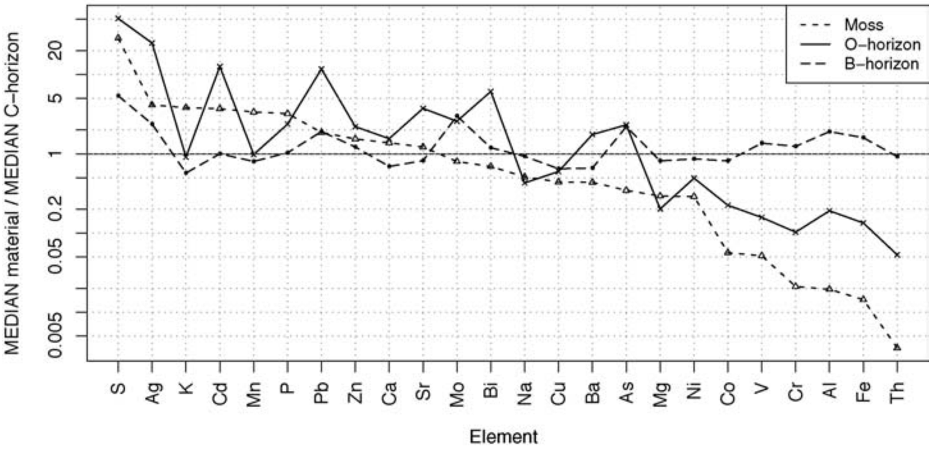

Figure 6.12 The same spiderplot as Figure 6.11, but the elements are sorted according to the continuously decreasing ratio for the moss samples relative to the underlying C-horizon

data analysis, and is an entrance point to multivariate data analysis. It can, for example, be used to identify those pairs of elements where a more detailed study in xy-plots appears to be necessary.

When using the scatterplot matrix with a multi-element data set, one of the major problems is screen and paper size. The individual plots become very small on the screen if more than 10 to 12 variables are displayed. If an A0 (or 36 inch) printer or plotter is available, up to a 60 × 60 matrix may be plotted. An important consideration before plotting a scatterplot matrix is the handling of outliers. Extreme outliers in several variables will greatly disturb the plot (Figure 6.13). Thus transformations need to be considered prior to plotting in order to down-weight the influence of outliers (Figure 6.14). Alternatively samples that exhibit extreme outliers can be removed (trimmed) prior to plotting a scatterplot matrix. Although the scatterplot matrix is a graphic and not a "formal" correlation analysis, the special problems of the closure of geochemical data still remain (see Section 10.5), especially for major and minor elements. The scatterplot matrix may help, however, to visualise some of these problems right away. Furthermore, by additive or centred logratio transformation of the major and minor element data (see Sections 10.5.1 and 10.5.2), the "opened" data may be displayed.

6.8 Ternary plots

Ternary plots show the relative proportions of three variables in one diagram (see, e.g., Cannings and Edwards, 1968). They have been used in petrology since the early 1900s to study the evolution of magmas or to differentiate between certain rock types. When three variables for a single observation sum to 100 per cent, they can be plotted as a point in a ternary plot. To construct an informative ternary plot, the three variables should have analytical results within the same order of magnitude, or all points will plot into the corner of the variable with the highest values. If there are large differences between the variables, it is necessary to multiply one or two of the variables by a factor to bring the values into the same data range as that of the other variables. The values for all three variables are then re-calculated to 100 per cent and

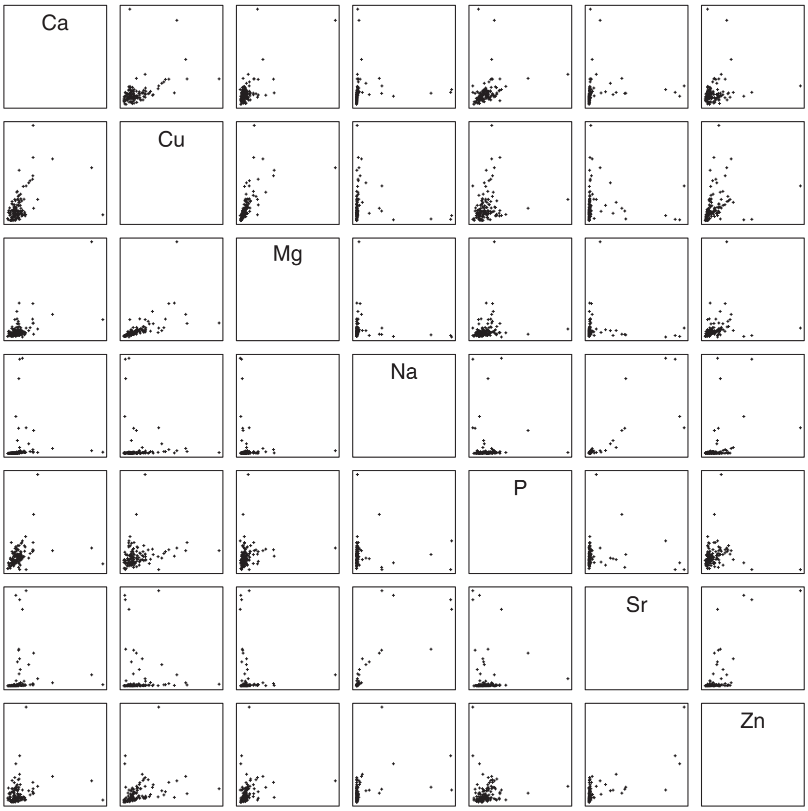

Figure 6.13 Scatterplot matrix of some selected elements, Kola Project C-horizon data. Axes could be labelled, but that will result in even smaller scatterplots, which contain the important information of this graphic. Some extreme values dominate the plot

plotted into the ternary space. In a ternary plot only two variables are independent, the value of the third variable is automatically known; the data are closed (Aitchison, 1986), and thus these diagrams should not be used to infer "correlations".

However, they can be used in environmental and other applied geochemical studies to aid "pattern recognition" and comparison with other studies.

Figure 6.15 shows two ternary plots constructed with the Kola Moss data set. The scale of the three axes varies between 0 and 100 per cent for each element (100 per cent in the corner of the element). The grey dashed lines in Figure 6.15 show the percentages. Just as in the scatterplot, it should be possible to identify interesting samples (usually outliers) via a simple click of the mouse. The Cu-Ni-Pb plot demonstrates that Pb is not an important component of the smelter emissions, which are dominated by Cu and Ni. Sample 907 was taken close to the Monchegorsk smelter. The Mn-Al-Fe ternary plot was constructed because both Al and Fe concentrations will be strongly influenced by the input of dust to the moss, while Mn is a plant nutrient. It is apparent that there are samples where Al dominates the dust component and others where Fe is more important. In addition, it appears that the majority of samples tend to the Mn corner of the diagram.

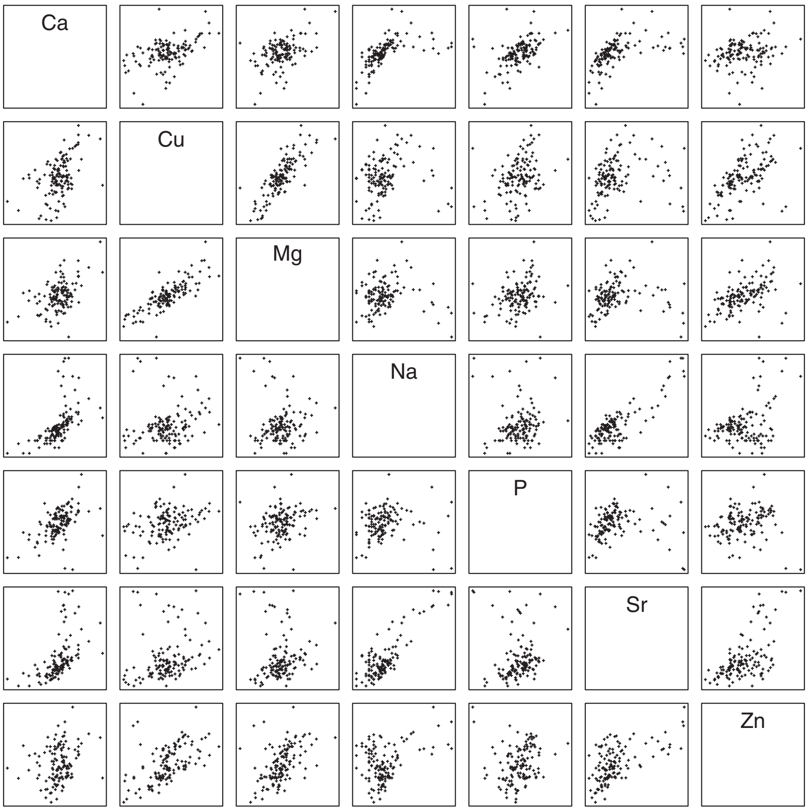

Figure 6.14 The same scatterplot matrix as above (Figure 6.13) but with log-transformed data to decrease the influence of extreme values

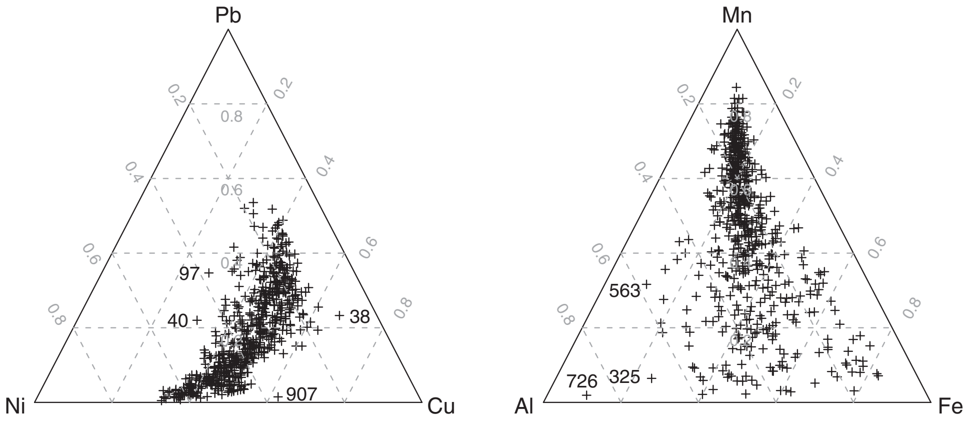

Figure 6.15 Two ternary plots for selected elements of the Kola Project Moss data set

For petrological applications it should again be possible to display pre-defined fields (or lines) in these diagrams that outline where rocks with particular names plot. Just as for xy-plots, one software package, based on R, which can plot these fields and lines for many pre-defined ternary plots (Janousek *et al*., 2006), is freely available at `http://www.gla.ac.uk/gcdkit/`.

6.9 Summary

Once the statistical and spatial data structure is documented, the relationships between different variables are an important avenue of study. While up to this point it was no major problem to study and document variable by variable, now almost countless possibilities are encountered. The number of possible plots showing bivariate data relationships increases quadratically with the number of variables. It may thus be tempting to commence right away with multivariate data analysis. However, all multivariate methods are built on many statistical assumptions about the data. It is thus advisable to use scatterplots first as the basis for studying bivariate data behaviour. Such bivariate plots used in an exploratory manner can be very informative, especially when adding user-defined lines, regression lines and trend lines. Plotting spatial and time trends are special cases of bivariate data analysis. The scatterplot matrix can be used to gain a graphical impression of the relationships between all possible pairs of variables. Three variables can be displayed in ternary plots. All of these simple plots can be used to great advantage to gain an improved knowledge of the data and its structure, interrelationships, before starting any formal statistical data analysis.