CHAPTER 2

Data preparation

Preparing data before analysis

Before ecological data can be analysed, they need to be prepared and put into the right format. Data that are entered in the wrong format cannot be analysed or will yield wrong results.

Different statistical programs require data in different formats. You should consult the manual of the statistical software to find out how data need to be prepared. Alternatively, you could check example datasets. An example of data preparation for the R package is presented at the end of this session.

Before you embark on the data analysis, it is essential to check for mistakes in data entry. If you detect mistakes later in the analysis, you would need to start the analysis again and could have lost considerable time. Mistakes in data entry can often be detected as exceptional values. The best procedure of analysing your results is therefore to start with checking the data.

An example of species survey data

Imagine that you are interested in investigating the hypothesis that soil depth influences tree species diversity. The data that will allow you to test this hypothesis are data on soil depth and data on diversity collected for a series of sample plots. We will see in a later chapter that diversity can be estimated from information on the species identity of every tree. Figure 2.1 shows species and soil depth data for the first four sample plots that were inventoried (to test the hypothesis, we need several sample plots that span the range from shallow to deep soils). For site A, three species were recorded (S1, S2 and S3) and a soil depth of 1 m. For site B, only two species were recorded (S1 with four trees and S3 with one tree) and a soil depth of 2 m.

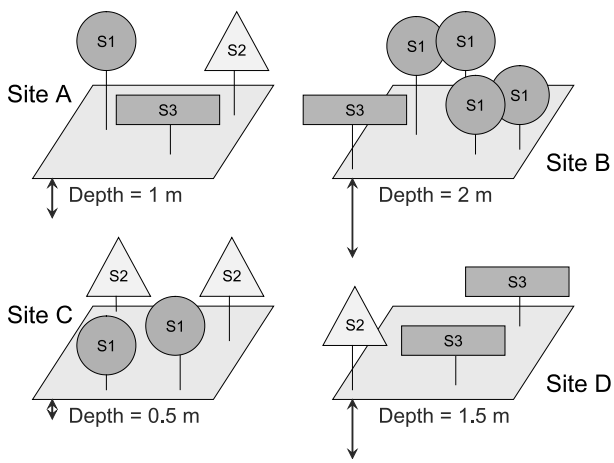

Figure 2.1 A simplified example of information recorded on species and environmental data.

The species information from Figure 2.1 can be recorded as follows:

Site	Species S1 (count)	Species S2 (count)	Species S3 (count)
A	1	1	1
B	4	0	1
C	2	2	0
D	0	1	2

The environmental information from Figure 2.1 can be recorded in a similar fashion:

Site	**Soil depth (m)**
A	1.0
B	2.0
C	0.5
D	1.5

This chapter deals with the preparation of data matrices as the two matrices given above. **Note that the example of Figure 2.1 is simplified**: typical species matrices have more than 100 rows and more than 100 columns. These matrices can be used as input for the analyses shown in the following chapters. They can be generated by a decent data management system. These matrices are usually not the ideal method of capturing, entering and storing data. Recording species data in the field is typically done with data collection forms that are filled for each site separately and that contain tables with a single column for the species name and a single column for the abundance. This is also the ideal method of storing species data.

A general format for species survey data

As seen above, all information can be recorded in the form of **data matrices**. All the types of data that are described in this manual can be prepared as two matrices: the **species matrix** and the **environmental matrix**. Table 2.1 shows a part of the species matrix for a well-studied dataset in community ecology, the dune meadow dataset. This dataset contains 30 species of which only 13 are presented. The data were collected on the vegetation of meadows on the Dutch island of Terschelling (Jongman et al. 1995). Table 2.2 shows the environmental data for this dataset.

You can notice that the rows of both matrices have the same names – they reflect the data that were collected for each **site** or **sample unit**. Sites could be sample plots, sample sites, farms, biogeographical provinces, or other identities. Sites are defined as the areas from which data were collected during a specific time period. We will use the term "**site**" further on in this manual. Sites will always refer to the **rows** of the datasets.

Some studies involve more than one type of sampling unit, often arranged hierarchically. For example, villages, farms in the village and plots within a farm. Sites of different types (such as plots, villages and districts) should not be mixed within the same data matrix. Each site of the matrix should be of the same type of sampling unit.

The columns of the matrices indicate the variables that were measured for each site. The cells of the matrices contain **observations** – bits of data recorded for a specific site and a specific variable.

We prefer using rows to represent samples and columns to represent variables to the alternative form where rows represent variables. Our preference is simply based on the fact that some general statistical packages use this format. Data can be presented by swapping rows and columns, since the contents of the data will remain the same.

Table 2.1 An example of a species matrix, where rows correspond to sites, columns correspond to species and cell entries are the abundance of the species at a particular site

Site	Achmil	Agrsto	Airpra	Alogen	Antodo	Belper	Brarut	Brohor	Calcus	Chealb	Cirarv	Elepal	Elyrep	…
X1	1	0	0	0	0	0	0	0	0	0	0	0	4	…
X2	3	0	0	2	0	3	0	4	0	0	0	0	4	…
X3	0	4	0	7	0	2	2	0	0	0	0	0	4	…
X4	0	8	0	2	0	2	2	3	0	0	2	0	4	…
X5	2	0	0	0	4	2	2	2	0	0	0	0	4	…
X6	2	0	0	0	3	0	6	0	0	0	0	0	0	…
X7	2	0	0	0	2	0	2	2	0	0	0	0	0	…
X8	0	4	0	5	0	0	2	0	0	0	0	4	0	…
X9	0	3	0	3	0	0	2	0	0	0	0	0	6	…
X10	4	0	0	0	4	2	2	4	0	0	0	0	0	…
X11	0	0	0	0	0	0	4	0	0	0	0	0	0	…
X12	0	4	0	8	0	0	4	0	0	0	0	0	0	…
X13	0	5	0	5	0	0	0	0	0	1	0	0	0	…
X14	0	4	0	0	0	0	0	0	4	0	0	4	0	…
X15	0	4	0	0	0	0	4	0	0	0	0	5	0	…
X16	0	7	0	4	0	0	4	0	3	0	0	8	0	…
X17	2	0	2	0	4	0	0	0	0	0	0	0	0	…
X18	0	0	0	0	0	2	6	0	0	0	0	0	0	…
X19	0	0	3	0	4	0	3	0	0	0	0	0	0	…
X20	0	5	0	0	0	0	4	0	3	0	0	4	0	…

Table 2.2 An example of an environmental matrix, where rows correspond to sites and columns correspond to variables

Site	A1	Moisture	Management	Use	Manure
X1	2.8	1	SF	Haypastu	4
X2	3.5	1	BF	Haypastu	2
X3	4.3	2	SF	Haypastu	4
X4	4.2	2	SF	Haypastu	4
X5	6.3	1	HF	Hayfield	2
X6	4.3	1	HF	Haypastu	2
X7	2.8	1	HF	Pasture	3
X8	4.2	5	HF	Pasture	3
X9	3.7	4	HF	Hayfield	1
X10	3.3	2	BF	Hayfield	1
X11	3.5	1	BF	Pasture	1
X12	5.8	4	SF	Haypastu	2
X13	6	5	SF	Haypastu	3
X14	9.3	5	NM	Pasture	0
X15	11.5	5	NM	Haypastu	0
X16	5.7	5	SF	Pasture	3
X17	4	2	NM	Hayfield	0
X18	4.6	1	NM	Hayfield	0
X19	3.7	5	NM	Hayfield	0
X20	3.5	5	NM	Hayfield	0

The species matrix

The species data are included in the species matrix. This matrix shows the values for each species and for each site (see data collection for various types of samples). For example, the value of 5 was recorded for species *Agrostis stolonifera* (coded as Agrsto) and for site 13. Another name for this matrix is the **community matrix**.

The species matrix often contains **abundance** values – the number of individuals that were counted for each species. Sometimes species data reflect the biomass recorded for each species. Biomass can be approximated by percentage **cover** (typical for surveys of grasslands) or by **cross-sectional area** (the surface area of the stem, typical for forest surveys). Some survey methods do not collect precise values but collect values that indicate a **range** of possible values, so that data collection can proceed faster. For instance, the value of 5 recorded for species *Agrostis stolonifera* and for site 13 indicates a range of 5-12.5% in cover percentage. The species matrix should not contain a range of values in a single cell, but a single number (the database can contain the range that is used to calculate the coding for the range). An extreme method of collecting data that only reflect a range of values is the **presence-absence** scale, where a value of 0 indicates that the species was not observed and a value of 1 shows that the species was observed.

A site will often only contain a small subset of all the species that were observed in the whole survey. Species distribution is often patchy. Species data will thus typically contain many zeros. Some statistical packages require that you are explicit that a value of zero was collected – otherwise the software could interpret an empty cell in a species matrix as a **missing value**. Such a missing value will not be used for the analysis, so you could obtain erroneous results if the data were recorded as zero but treated as missing.

The environmental matrix

The environmental dataset is more typical of the type of dataset that a statistical package normally handles. The columns in the environmental dataset contain the various environmental variables. The rows indicate the sites for which the values were recorded. The environmental variables can be referred to as **explanatory variables** for the types of analysis that we describe in this manual. Some people prefer to call these variables **independent variables**, and others prefer the term **x variables**. For instance, the information on the thickness of the A1 horizon of the dune meadow dataset shown in Table 2.2 can be used as an explanatory variable in a model that explains where species *Agrostis stolonifera* occurs. The research hypotheses will have indicated which explanatory variables were recorded, since an infinite number of environmental variables could be recorded at each site.

The environmental dataset will often contain two types of variables: **quantitative variables** and **categorical variables**.

Quantitative variables such as the thickness of the A1 horizon of Table 2.2 contain observations that are measured quantities. The observation for the A1 horizon of site 1 was for example recorded by the number 2.8. Various statistics can be calculated for quantitative variables that cannot be calculated for categorical variables. These include:

- The mean or average value
- The standard deviation (this value indicates how close the values are to the mean)
- The median value (the middle value when values are sorted from low to high) (synomyms for this value are the 50% quantile or 2nd quartile)
- The 25% and 75% quantiles = 1st and 3rd quartiles (the values for which 25% or 75% of values are smaller when values are sorted from low to high)
- The minimum value
- The maximum value

For the thickness of A1 horizon of Table 2.2, we obtain following summary statistics.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.800   3.500   4.200   4.850   5.725  11.500
```

These statistics summarize the values that were obtained for the quantitative variable. Another method by which the values for a quantitative variable can be summarized is a **boxplot graph** as shown in Figure 2.2. The whiskers show the minimum and maximum of the dataset, except if some values are farther than 1.5 × the interquartile range (the difference between the 1st and 3rd quartile) from the median value. Note that various software packages or options within such package will result in different statistics to be portrayed in boxplot graphs – you may want to check the documentation of your particular software package. An important feature of Figure 2.2 is that it shows that there are some **outliers** in the dataset. If your data are normally distributed, then you would only rarely (less than 1% of the time) expect to observe an outlier. If the boxplot indicates outliers, check whether you entered the data correctly (see next page).

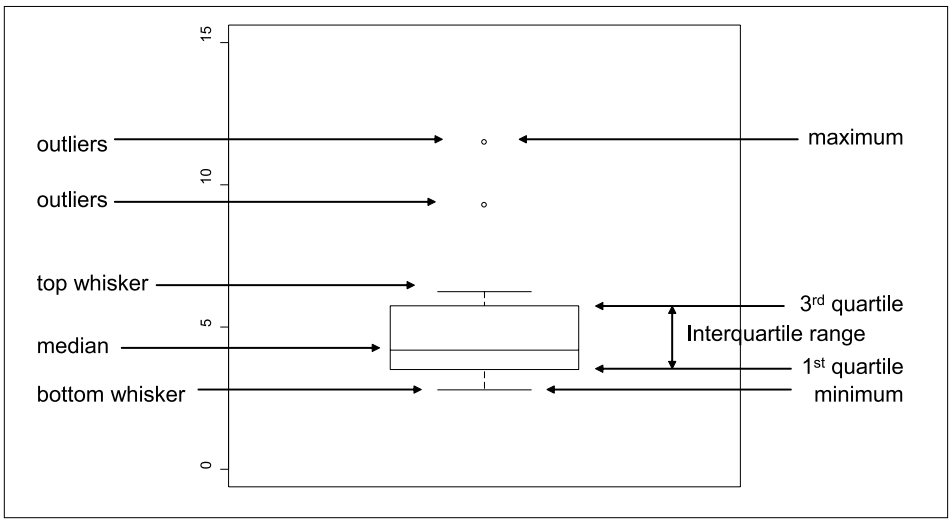

Figure 2.2 Summary of a quantitative variable as a boxplot. The variable that is summarized is the thickness of the A1 horizon of Table 2.2.

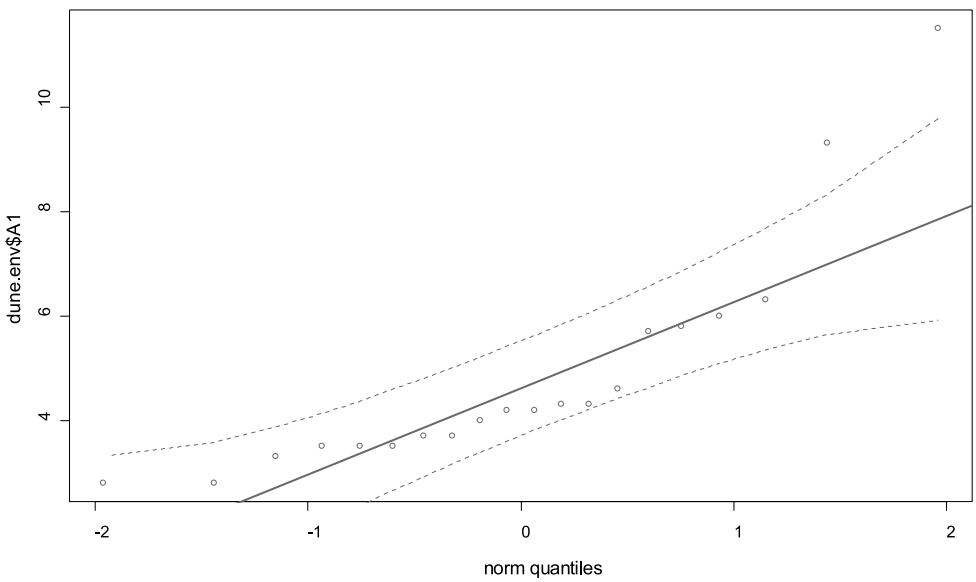

Figure 2.3 Summary of a quantitative variable as a Q-Q plot. The variable that is summarized is the thickness of the A1 horizon of Table 2.2. The two outliers (upper right-hand side) correspond to the outliers of Figure 2.2.

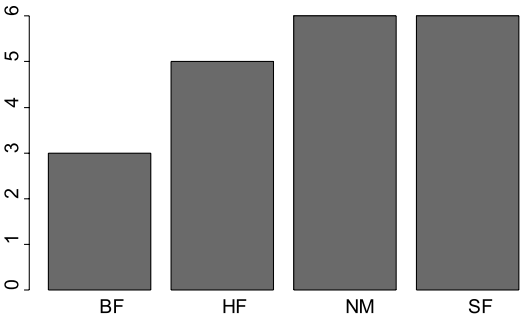

Figure 2.4 Summary of a categorical variable by a bar plot. The management of Table 2.2 is summarized.

There are other graphical methods for checking for outliers for quantitative variables. One of these methods is the **Q-Q plot**. When data are normally distributed, all observations should be plotted roughly along a straight line. Outliers will be plotted further away from the line. Figure 2.3 gives an example. Another method to check for outliers is to plot a histogram. The key point is to check for the exceptional observations.

Categorical variables (or **qualitative variables**) are variables that contain information on data categories. The observations for the type of management for the dune meadow dataset (presented in Table 2.2) have four values: "standard farming", "biological farming", "hobby farming" and "nature conservation management". The observation for the type of management is thus not a number. In statistical textbooks, categorical variables are also referred to as **factors**. Factors can only contain a limited number of **factor levels**.

The only way by which categorical variables can be summarized is by listing the number of observations or frequency of each category. For instance, the summary for the management variable of Table 2.2 could be presented as:

	Category			
	BF	HF	NM	SF
observations	3	5	6	6

Graphically, the summary can be represented as a **barplot**. Figure 2.4 shows an example for the management of Table 2.2.

Some researchers record observations of categorical variables as a number, where the number represents the code for a specific type of value – for instance code "1" could indicate "standard farming". We do not encourage the usage of numbers to code for factor levels since statistical software and analysts can confuse the variable with a quantitative variable. The statistical software could report erroneously that the average management type is 2.55, which does not make sense. It would definitely be wrong to conclude that the average management type would be 3 (the integer value closest to 2.55) and thus be hobby-farming. A better way of recording categorical variables is to include characters. You are then specific that the value is a factor level – you could for instance use the format of "c1", "c2", "c3" and "c4" to code for the four management regimes. Even better techniques are to use meaningful abbreviations for the factor levels – or to just use the entire description of the factor level, since most software will not have any problems with long descriptions and you will avoid confusion of collaborators or even yourself at later stages.

Ordinal variables are somewhere between quantitative and categorical variables. The manure variable of the dune meadow dataset is an ordinal variable. Ordinal variables are not measured on a quantitative scale but the order of the values is informative. This means for manure that progressively more manure is used from manure class 0 until 4. However, since the scale is not quantitative, a value of 4 does not mean that four times more manure is used than for value 1 (if it was, then we would have a quantitative variable). For the same reason manure class 3 is not the average of manure class 2 and 4.

You can actually choose whether you treat ordinal variables as quantitative or categorical

variables in the statistical analysis. In many statistical packages, when the observations of a variable only contain numbers, the package will assume that the variable is a quantitative variable. If you want the variable to be treated as a categorical variable, you will need to inform the statistical package about this (for example by using a non-numerical coding system). If you are comfortable to assume for the analysis that the ordinal variables were measured on a quantitative scale, then it is better to treat them as quantitative variables. Some special methods for ordinal data are also available.

Checking for exceptional observations that could be mistakes

The methods of summarizing quantitative and categorical data that were described in the previous section can be used to check for exceptional data. Maximum or minimum values that do not correspond to the expectations will easily be spotted. Figure 2.5 for instance shows a boxplot for the A1 horizon that contained a data entry error for site 3 as the value 43 was entered instead of 4.3. Compare with Figure 2.2. You should be aware of the likely ranges of all quantitative variables.

Some mistakes for categorical data can easily be spotted by calculating the frequencies of observations for each factor level. If you had entered "NN" instead of "NM" for one management observation in the dune meadow dataset, then a table with the number of observations for each management type would easily reveal that mistake. This method is especially useful when the number of observations is fixed for each level. If you designed your survey so that each type of management should have 5 observations, then spotting one type of management with 4 observations and one type with 1 observation would reveal a data entry error.

Some exceptional observations will only be spotted when you plot variables against each other as part of exploratory analysis, or even later when you started conducting some statistical analysis. Figure 2.6 shows a plot of all possible pairs of the environmental variables of the dune meadow dataset. You can notice the two outliers for the thickness of the A1 horizon, which occur at moisture category 4 and manure category 1, for instance.

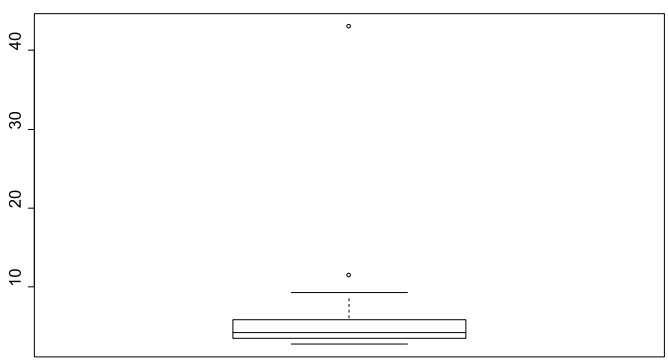

Figure 2.5 Checking for exceptional observations.

After having spotted a potential mistake, you need to record immediately where the potential mistake occurred, especially if you do not have time to directly check the raw data. You can include a text file where you record potential mistakes in the folder where you keep your data. Alternatively, you could give the cell in the spreadsheet where you keep a copy of the data a bright colour. Yet another method is to add an extra variable in your dataset where comments on potential mistakes are listed. However the best method is to directly check and change your raw data (if a mistake is found). Always record the changes that you have made and the reasons for them. Note that an observation that looks odd but which can not be traced to a mistake should not be changed or assumed to be missing. If it is clearly a nonsense value, but no explanation can be found, then it should be omitted. If it is just a strange value then various courses are open to you. You can try analysing the data with and without the observation to check if it makes a big difference to results. You might have to go back to the field and take the measurement again, finding a field explanation if the odd value is repeated.

Do not get confused when you have various datasets in various stages of correction. Commonly scientists end up with several versions of each data file and loose track of which is which. The best method is to have only one dataset, of which you make regular backups.

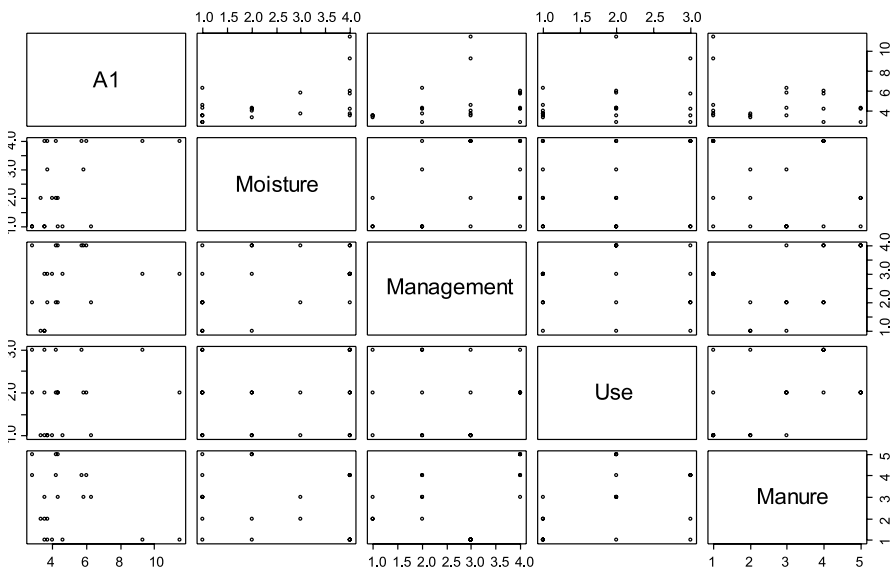

Figure 2.6 Checking for exceptional data by pairwise comparisons of the variables of Table 2.2.

Methods of transforming the values in the matrices

There are many ways in which the values of the species and environmental matrices can be transformed. Some methods were developed to make data more conform to the normal distribution. What transformation you use will depend on your objectives and what you want to assume about the data. For several types of analysis described in later chapters you do not need to transform the species matrix, and most analyses do not actually require the explanatory variables to be normally distributed. It is therefore not good practice to always transform explanatory variables to be normally distributed. Moreover, in many cases it will not be possible to find a transformation that will result in normally distributed data.

We recommend only transforming variables if you have a good reason to investigate a particular pattern that will be revealed by the transformation. For example, an extreme way of transforming the species matrix is to change the values to 1 if the species is present and 0 if the species is absent. The subsequent analysis will thus not be influenced by differences in species' abundances. By comparing the results of the analysis of the original data with the results from the transformed data, you can get an idea of the influence of differences in abundance on the results. If one species dominates and the ordination results are only influenced by that one species, then you could use a logarithmic or square-root transformation to diminish the influence of the dominant species – again this means that there is a good reason for the transformation and such should not be a standard approach. The fact that the results are influenced by the dominant species is actually a clear demonstration of an important pattern in your dataset.

Examples of the analysis with the menu options of Biodiversity.R

See in chapter 3 how data can be loaded from an external file:

Data > Import data > from text file…

→Enter name for dataset: data (choose any name)

→Click "OK"

→Browse for the file and click on it

To save data to an external file:

Data > Active Dataset > export active dataset…

→File name: export.txt (choose any name)

Select the species and environmental matrices:

Biodiversity > Environmental Matrix > Select environmental matrix

→Select the dune.env dataset

Biodiversity > Community matrix > Select community matrix

→Select the dune dataset

To summarize the data and check for exceptional cases:

Biodiversity > Environmental Matrix > Summary…

→Select variable: A1

→Click "OK"

→Click "Plot"

Examples of the analysis with the command options of Biodiversity.R

To load data from an external file:

```
data <- read.table(file="D://my files/data.txt")
data <- read.table(file.choose())
```

To save data to an external file:

```
write.table(data, file="D://my files/data.txt")
write.table(data, file.choose())
```

To summarize the data and check for exceptional cases:

```
summary(dune.env)
boxplot(dune.env$A1)
points(mean(dune.env$A1),cex=1.5)
table(dune.env$Management)
plot(dune.env$Management)
pairs(dune.env)
```

To transform the data:

```
dune.ln.transformed <- log(dune+1)
dune.squareroot.transformed <- dune^0.5
dune.speciesprofile <- decostand(dune,"total")
dune.env$A1.standard <- scale(dune.env$A1)
```

Checking whether data is normally distributed:

```
qq.plot(dune.env$A1)
shapiro.test(dune.env$A1)
ks.test(dune.env$A1,pnorm)
```