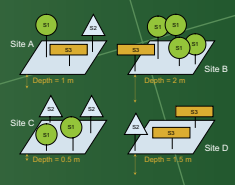
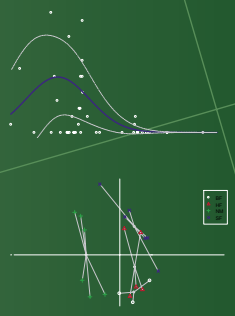
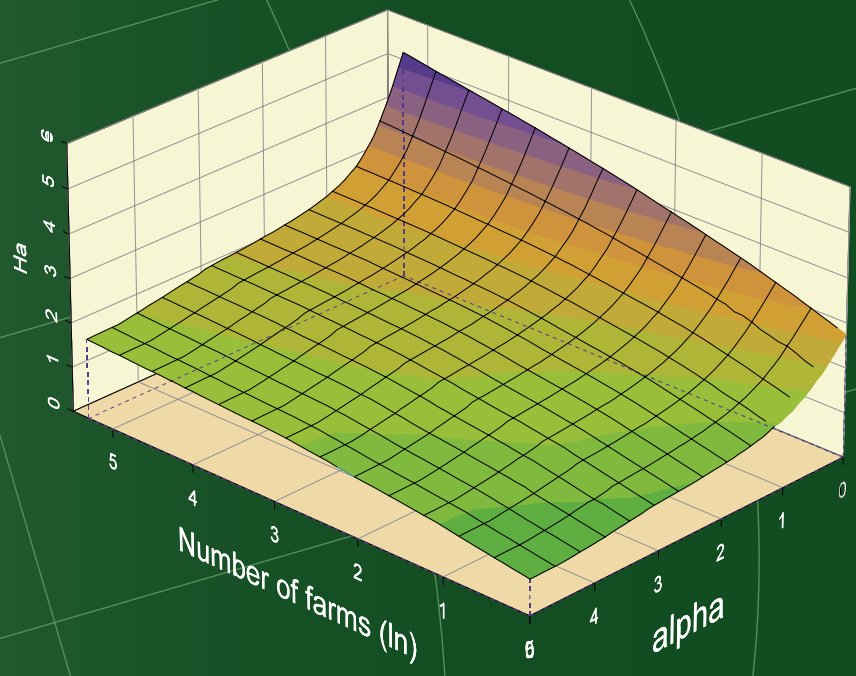
includes CD with software

Tree diversity analysis

A manual and software for common statistical methods for ecological and biodiversity studies







Roeland Kindt and Richard Coe

World Agroforestry Centre
TRANSFORMING LIVES AND LANDSCAPES

Tree diversity analysis

A manual and software for common statistical methods for ecological and biodiversity studies

Roeland Kindt and Richard Coe

World Agroforestry Centre, Nairobi, Kenya

World Agroforestry Centre

TRANSFORMING LIVES AND LANDSCAPES

Suggested citation: Kindt R and Coe R. 2005. *Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies*. Nairobi: World Agroforestry Centre (ICRAF).

Published by the World Agroforestry Centre
United Nations Avenue
PO Box 30677, GPO 00100
Nairobi, Kenya
Tel: +254(0)20 7224000, via USA +1 650 833 6645
Fax: +254(0)20 7224001, via USA +1 650 833 6646
Email: icraf@cgiar.org
Internet:www.worldagroforestry.org

ISBN: 92 9059 179 X

Design and Layout: K. Vanhoutte
Printed in Kenya

Contents

Acknowledgements

We warmly thank all that provided inputs that lead to improvement of this manual. We especially appreciate the comments received during training sessions with draft versions of this manual and the accompanying software in Kenya, Uganda and Mali. We are equally grateful to the thoughtful reviews by Dr Simoneta Negrete-Yankelevich (Instituto de Ecología, Mexico) and Dr Robert Burn (Reading University, UK) of the draft version of this manual, and to Hillary Kipruto for help in editing of this manual.

We highly appreciate the support of the Programme for Cooperation with International Institutes (SII), Education and Development Division of the Netherlands' Ministry of Foreign Affairs, and VVOB (The Flemish Association for Development Cooperation and Technical Assistance, Flanders, Belgium) for funding the development for this manual. We also thank VVOB for seconding Roeland Kindt to the World Agroforestry Centre (ICRAF).

This tree diversity analysis manual was inspired by research, development and extension activities that were initiated by ICRAF on tree and landscape diversification. We want to acknowledge the various donor agencies that have funded these activities, especially VVOB, DFID, USAID and EU.

We are grateful for the developers of the R Software for providing a free and powerful statistical package that allowed development of Biodiversity.R. We also want to give special thanks to Jari Oksanen for developing the vegan package and John Fox for developing the Rcmdr package, which are key packages that are used by Biodiversity.R.

Introduction

This manual was prepared during training events held in East- and West-Africa on the analysis of tree diversity data. These training events targeted data analysis of tree diversity data that were collected by scientists of the World Agroforestry Centre (ICRAF) and collaborating institutions. Typically, data were collected on the tree species composition of quadrats or farms. At the same time, explanatory variables such as land use and household characteristics were collected. Various hypotheses on the influence of explanatory variables on tree diversity can be tested with such datasets. Although the manual was developed during research on tree diversity on farms in Africa, the statistical methods can be used for a wider range of organisms, for different hierarchical levels of biodiversity, and for a wider range of environments.

These materials were compiled as a second-generation development of the Biodiversity Analysis Package, a CD-ROM compiled by Roeland Kindt with resources and guidelines for the analysis of ecological and biodiversity information. Whereas the Biodiversity Analysis Package provided a range of tools for different types of analysis, this manual is accompanied by a new tool (Biodiversity.R) that offers a single software environment for all the analyses that are described in this manual. This does not mean that Biodiversity.R is the only recommended package for a particular type of analysis, but it offers the advantage for training purposes that users only need to be introduced to one software package for statistically sound analysis of biodiversity data.

It is never possible to produce a guide to all the methods that will be needed for analysis of biodiversity data. Data analysis questions are continually advancing, requiring ever changing data collection and analysis methods. This manual focuses on the analysis of species survey data. We describe a number of methods that can be used to analyse hypotheses that are frequently important in biodiversity research. These are not the only methods that can be used to analyse these hypotheses, and other methods will be needed when the focus of the biodiversity research is different.

Effective data analysis requires imagination and creativity. However, it also requires familiarity with basic concepts, and an ability to use a set of standard tools. This manual aims to provide that. It also points the user to other resources that develop ideas further.

Effective data analysis also requires a sound and up to date understanding of the science behind the investigation. Data analysis requires clear objectives and hypotheses to investigate. These have to be based on, and push forward, current understanding. We have not attempted to link the methods described here to the rapidly changing science of biodiversity and community ecology. Data analysis does not end with production of statistical results. Those results have to be interpreted in the light of other information about the problem. We can not, therefore, discuss fully the interpretation of the statistical results, or the further statistical analyses they may lead to.

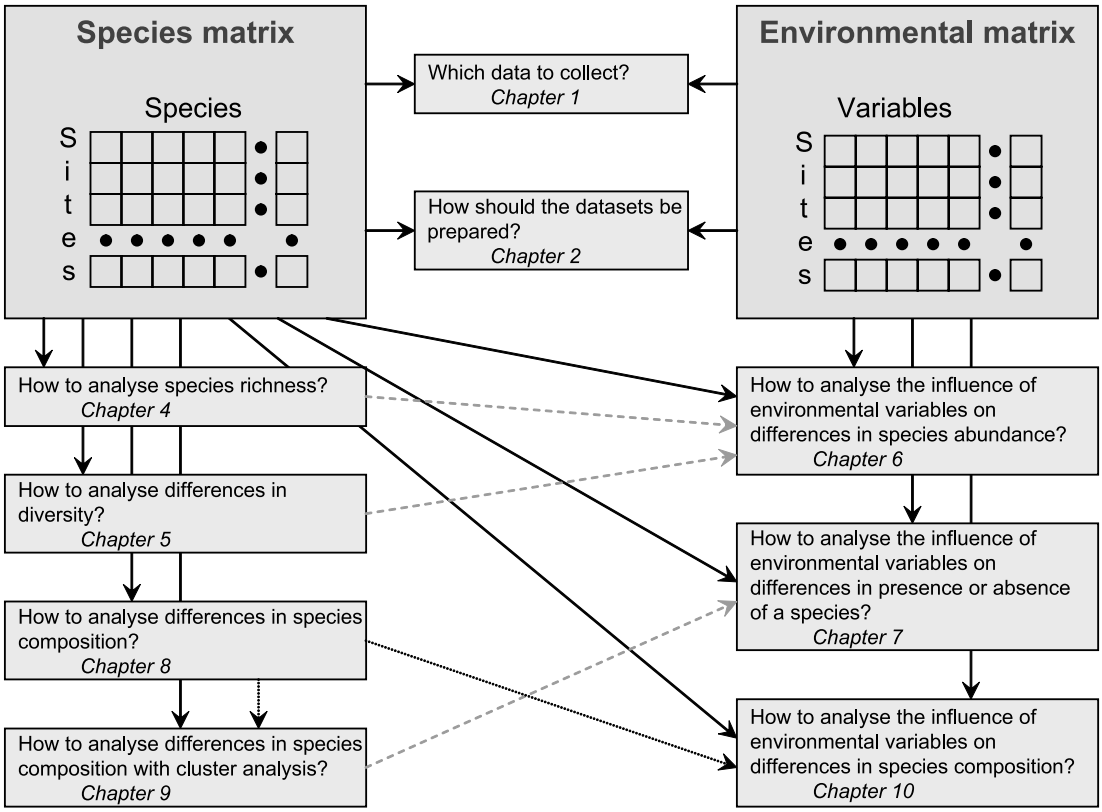
Overview of methods described in this manual

On the following page, a general diagram is provided that describes the data analysis questions that you can ask when analysing biodiversity based on the methodologies that are provided in this manual. Each question is discussed in further detail in the respective chapter. The arrows indicate the types of information that are used in each method. All information is derived from either the species data or the environmental data of the sites. Chapter 2 describes the species and environmental data matrices in greater detail.

Some methods only use information on species. These methods are depicted on the left-hand side of the diagram. They are based on biodiversity statistics that can be used to compare the levels of biodiversity between sites, or to analyse how similar sites are in species composition.

The other methods use information on both species and the environmental variables of the sites. These methods are shown on the right-hand side of the diagram. These methods provide insight into the influence of environmental variables on biodiversity. The analysis methods can reveal how much of the pattern in species diversity can be explained by the influence of the environmental variables. Knowing how much of a pattern is explained will especially be useful if the research was conducted to arrive at options for better management of biodiversity. Note that in this context, 'environmental variables' can include characteristics of the social and economic environment, not only the biophysical environment.

You may have noticed that Chapter 3 did not feature in the diagram. The reason is that this chapter describes how the Biodiversity.R software can be installed and used to conduct all the analyses described in the manual, whereas you may choose to conduct the analysis with different software. For this reason, the commands and menu options for doing the analysis in Biodiversity.R are separated from the descriptions of the methods, and placed at the end of each chapter.


Species matrix
Species
Sites
Which data to collect?
Chapter 1
How should the datasets be prepared?
Chapter 2
Environmental matrix
Variables
Sites
How to analyse species richness?
Chapter 4
How to analyse differences in diversity?
Chapter 5
How to analyse differences in species composition?
Chapter 8
How to analyse differences in species composition with cluster analysis?
Chapter 9
How to analyse the influence of environmental variables on differences in species abundance?
Chapter 6
How to analyse the influence of environmental variables on differences in presence or absence of a species?
Chapter 7
How to analyse the influence of environmental variables on differences in species composition?
Chapter 10

CHAPTER 1

Sampling

Sampling

Choosing a way to sample and collect data can be bewildering. If you find it hard to decide exactly how it should be done then seek help. Questions about sampling are among the questions that are most frequently asked to biometricians and the time to ask for assistance is while the sampling scheme is being designed. Remember: if you go wrong with data analysis it is easy to repeat it, but if you collect data in inappropriate ways you can probably not repeat it, and your research will not meet its objectives.

Although there are some particular methods that you can use for sampling, you will need to make some choices yourself. Sample design is the art of blending theoretical principles with practical realities. It is not possible to provide a catalogue of sampling designs for a series of situations – simply too much depends on the objectives of the survey and the realities in the field.

Sampling design has to be based on specific research objectives and the hypotheses that you want to test. When you are not clear about what it is that you want to find out, it is not possible to design an appropriate sampling scheme.

Research hypotheses

The only way to derive a sampling scheme is to base it on a specific research hypothesis or research objective. What is it that you want to find out? Will it help you or other researchers when you find out that the hypothesis holds true? Will the results of the study point to some management decisions that could be taken?

The research hypotheses should indicate the 3 basic types of information that characterize each piece of data: **where** the data were collected, **when** the data were collected, and **what** type of measurement was taken. The where, when and what are collected for each **sample unit**. A sample unit could be a sample plot in a forest, or a farm in a village. Some sample units are natural units such as fields, farms or forest gaps. Other sample units are subsamples of natural units such as a forest plot that is placed within a forest. Your **sampling scheme** will describe how sample units are defined and which ones are selected for measurement.

The objectives determine **what** data, the **variables** measured on each sampling unit. It is helpful to think of these as response and explanatory variables, as described in the chapter on data preparation. The response variables are the key quantities that your objectives refer to, for example 'tree species richness on small farms'. The explanatory variables are the variables that you expect, or hypothesize, to influence the response. For example, your hypothesis could be that 'tree species richness on small farms is influenced by the level of market integration of the farm enterprise *because* market integration determines which trees are planted and retained'. In this example, species richness is the response variable and level of market integration is an explanatory variable. The hypothesis refers to small farms, so these should be the study units. The 'because…' part of the hypothesis adds much value to the research, and investigating it requires additional information on whether species were planted or retained and why.

Note that **this manual only deals with survey data.** The only way of proving **cause-effect** relationships is by conducting well-designed **experiments** – something that would be rather hard for this example! It is common for ecologists to draw conclusions about causation from relationships founding surveys. This is dangerous, but inevitable when experimentation is not feasible. The risk of making erroneous conclusions is reduced by: (a) making sure other possible explanations have been controlled or allowed for; (b) having a mechanistic theory model that explains why the cause-effect may apply; and (c) finding the same relationship in many different studies. However, in the end the conclusion depends on the argument of the scientist rather than the logic of the research design. Ecology progresses by scientists finding new evidence to improve the inevitably incomplete understanding of cause and effect from earlier studies.

When data are collected is important, both to make sure different observations are comparable and because understanding change – trends, or before and after an intervention – is often part of the objective. Your particular study may not aim at investigating trends, but investigating changes over time may become the objective of a later study. Therefore you should also document when data were collected.

This chapter will mainly deal with **where** data are collected. This includes definition of the **survey area**, of the **size and shape of sample units and plots** and of **how sample plots are located** within the survey area.

Survey area

You need to make a clear statement of the survey area for which you want to test your hypothesis. The survey area should have explicit geographical (and temporal) boundaries. The survey area should be at the ecological scale of your research question. For example, if your research hypothesis is something like 'diversity of trees on farms decreases with distance from Mount Kenya Forest because seed dispersal from forest trees is larger than seed dispersal from farm trees', then it will not be meaningful to sample trees in a strip of 5 metres around the forest boundary and measure the distance of each tree from the forest edge. In this case we can obviously not expect to observe differences given the size of trees (even if we could determine the exact distance from the edge within the small strip). But if the 5 m strip is not a good survey area to study the hypothesis, which area is? You would have to decide that on the basis of other knowledge about seed dispersal, about other factors which dominate the process when you get too far from Mt Kenya forest, and on practical limitations of data collection. You should select the survey area where you expect to **observe the pattern** given the **ecological size** of the phenomenon that you are investigating.

If the research hypothesis was more general, for example 'diversity of trees on East African farms decreases with distance from forests because more seeds are dispersed from forest trees than from farm trees', then we will need a more complex strategy to investigate it. You will certainly have to study more than one forest to be able to conclude this is a general feature of forests, not just Mt Kenya forest. You will therefore have to face questions of what you mean by a 'forest'. The sampling strategy now needs to determine how forests are selected as well as how farms around each forest are sampled.

A common mistake is to restrict data collection to only part of the study area, but assume the results apply to all of it (see Figure 1.1). You can not be sure that the small window actually sampled is representative of the larger study area.

An important idea is that **bias** is avoided. Think of the case in which samples are only located in sites which are easily accessible. If accessibility is associated with diversity (for example because fewer trees are cut in areas that are more difficult to access), then the area that is sampled will not

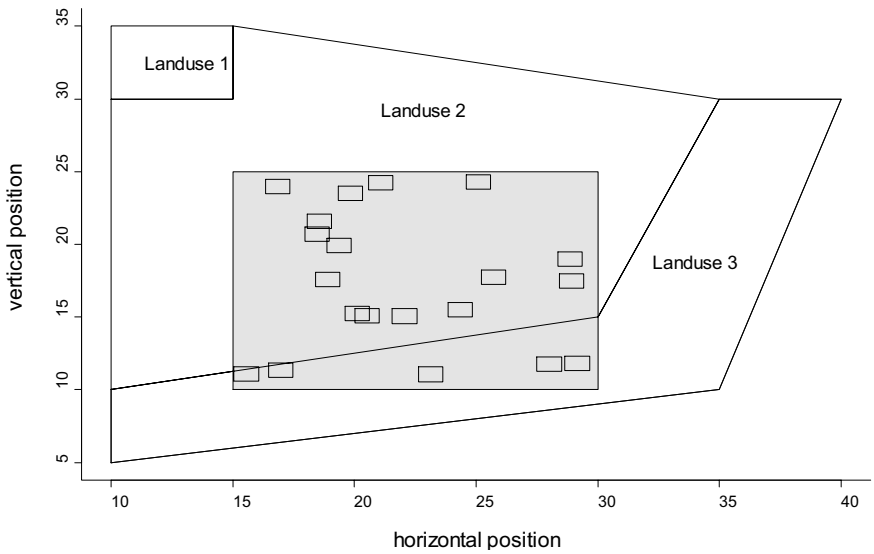


Figure 1.1 When you sample within a smaller window, you may not have sampled the entire range of conditions of your survey area. The sample may therefore not be representative of the entire survey area. The areas shown are three types of landuse and the sample window (with grey background). Sample plots are the small rectangles.

be representative of the entire survey area. An estimate of diversity based only on the accessible sites would give biased estimates of the whole study area. This will especially cause problems if the selection bias is correlated with the factors that you are investigating. For example, if the higher diversity next to the forest is caused by a larger proportion of areas that are difficult to access and you only sample areas that are easy to access, then you may not find evidence for a decreasing trend in diversity with distance from the forest. In this case, the dataset that you collected will generate estimates that are biased since the sites are not representative of the entire survey area, but only of sites that are easy to access.

The sample plots in Figure 1.1 were selected from a sampling window that covers part of the study area. They were selected using a method that allowed any possible plot to potentially be included. Furthermore, the selection was random. This means that inferences based on the data apply to the sampling window. Any particular sample will not give results (such as diversity, or its relationship with distance to forest) which are equal to those from measuring the whole sampling window. But the sampling will not predispose us to under- or overestimate the diversity, and statistical methods will generally allow us to determine just how far from the 'true' answer any result could be.

Size and shape of sample units or plots

A sample unit is the geographical area or plot on which you actually collected the data, and the time when you collected the data. For instance, a sample unit could be a 50 × 10 m^2 quadrat (a rectangular sample plot) in a forest sampled on 9th May 2002. Another sample unit could be all the land that is cultivated by a family, sampled on 10th December 2004. In some cases, the sample plot may be determined by the hypothesis directly. If you are interested in the influence of the wealth of farmers on the number of tree species on their farm, then you could opt to select the farm as the sample plot. Only in cases where the size of this sample plot is not practical would you need to search for an alternative sample plot. In the latter case you would probably use two sample units such as farms (on which you measure wealth) and plots within farms (on which you measure tree species, using the data from plots within a farm to estimate the number of species for the whole farm to relate to wealth).

The **size** of the quadrat will usually influence the results. You will normally find more species and more organisms in quadrats of 100 m^2 than in quadrats of 1 m^2. But 100 dispersed 1 m^2 plots will probably contain more species than a single 100 m^2 plot. If the aim is not to find species but understanding some ecological phenomenon, then either plot size may be appropriate, depending on the scale of the processes being studied.

The **shape** of the quadrat will often influence the results too. For example, it has been observed that more tree species are observed in rectangular quadrats than in square quadrats of the same area. The reason for this phenomenon is that tree species often occur in a clustered pattern, so that more trees of the same species will be observed in square quadrats. When quadrats are rectangular, then the orientation of the quadrat may also become an issue. Orienting the plots parallel or perpendicular to contour lines on sloping land may influence the results, for instance. As deciding whether trees that occur near the edge are inside or outside the sample plot is often difficult, some researchers find circular plots superior since the ratio of edge-to-area is smallest for circles. However marking out a circular plot can be much harder than marking a rectangular one. This is an example of the trade off between what may be theoretically optimal and what is practically best. Balancing the trade off is a matter of practical experience as well as familiarity with the principles.

As size and shape of the sample unit can influence results, it is best to stick to one size and shape for the quadrats within one study. If you want to compare the results with other surveys, then it will be easier if you used the same sizes and shapes of quadrats. Otherwise, you will need to convert results to a common size and shape of quadrat for comparisons. For some variables, such **conversion** can easily be done, but for some others this may be quite tricky. Species richness and diversity are statistics that are influenced by the size of the sample plot. Conversion is even more complicated since different methods can be used to measure sample size, such as area or the number of plants measured (see chapter on species richness). The average number of trees is easily converted to a common sample plot size, for example 1 ha, by multiplying by the appropriate scaling factor. This can not be done for number of species or diversity. Think carefully about conversion, and pay special attention to conversions for species richness and diversity. In some cases, you may not need to convert to a sample size other than the one you used – you may for instance be interested in the average species richness per farm and not in the average species richness in areas of 0.1 ha in farmland. Everything will depend on being clear on the research objectives.

One method that will allow you to do some easy conversions is to split your quadrat into sub-plots of smaller sizes. For example, if your quadrat is 40 × 5 m^2, then you could split this quadrat into eight

5×5 m^2 subplots and record data for each subplot. This procedure will allow you to easily convert to quadrat sizes of 5×5 m^2, 10×5 m^2, 20×5 m^2 and 40×5 m^2, which could make comparisons with other surveys easier.

Determining the size of the quadrat is one of the tricky parts of survey design. A quadrat should be large enough for differences related to the research hypothesis to become apparent. It should also not be too large to become inefficient in terms of cost, recording fatigue, or hours of daylight. As a general rule, several small quadrats will give more information than few large quadrats of the same total area, but will be more costly to identify and measure. Because differences need to be observed, but observation should also use resources efficiently, the type of organism that is being studied will influence the best size for the quadrat. The best size of the quadrat may differ between trees, ferns, mosses, butterflies, birds or large animals. For the same reason, the size of quadrat may differ between vegetation types. When studying trees, quadrat sizes in humid forests could be smaller than quadrat sizes in semi-arid environments.

As some rough indication of the size of the sample unit that you could use, some of the sample sizes that have been used in other surveys are provided next. Some surveys used 100×100 m^2 plots for differences in tree species composition of humid forests (Pyke et al. 2001, Condit et al. 2002), or for studies of forest fragmentation (Laurance et al. 1997). Other researchers used transects (sample plots with much longer length than width) such as 500×5 m^2 transects in western Amazonian forests for studies of differences in species composition for certain groups of species (Tuomisto et al. 2003). Yet other researchers developed methods for rapid inventory such as the method with variable subunits developed at CIFOR that has a maximum size of 40×40 m^2, but smaller sizes when tree densities are larger (Sheil et al. 2003). Many other quadrat sizes can be found in other references. It is clear that there is no common or standard sample size that is being used everywhere. The large range in values emphasizes our earlier point that there is no fixed answer to what the best sampling strategy is. It will depend on the hypotheses, the organisms, the vegetation type, available resources, and on the creativity of the researcher. In some cases, it may be worth using many small sample plots, whereas in other cases it may be better to use fewer larger sample plots. A pilot survey may help you in deciding what size and shape of sample plots to use for the rest of the survey (see below: pilot testing of the sampling protocol). Specific guidelines on the advantages and disadvantages of the various methods is beyond the scope of this chapter (an entire manual could be devoted to sampling issues alone) and the best advise is to consult a biometrician as well as ecologists who have done similar studies.

Simple random sampling

Once you have determined the survey area and the size of your sampling units, then the next question is where to take your samples. There are many different methods by which you can place the samples in your area.

Simple random sampling involves locating plots randomly in the study area. Figure 1.2 gives an example where the coordinates of every sample plot were generated by random numbers. In this method, we randomly selected a horizontal and vertical position. Both positions can be calculated by multiplying a random number between 0 and 1 with the range in positions (maximum – minimum), and adding the result to the minimum position. If the selected position falls outside the area (which is possible if the area is not rectangular), then a new position is selected.

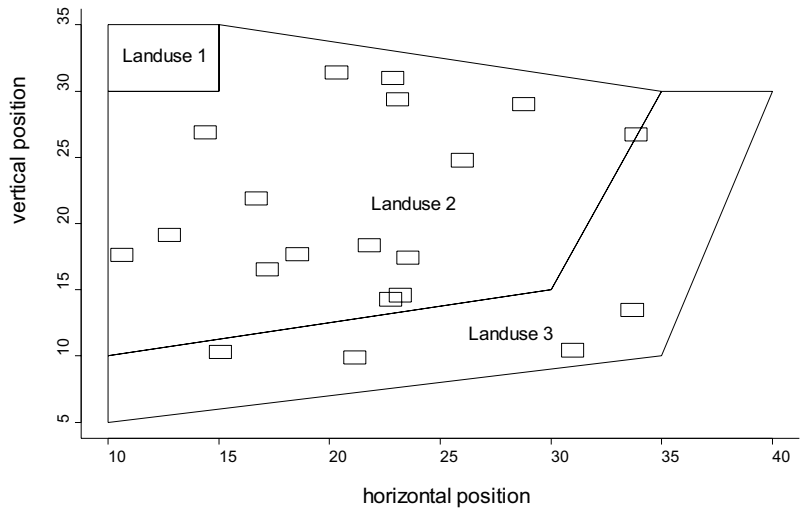


Figure 1.2 Simple random sampling by using random numbers to determine the position of the sample plots. Using this method there is a risk that regions of low area such as that under Landuse 1 are not sampled.

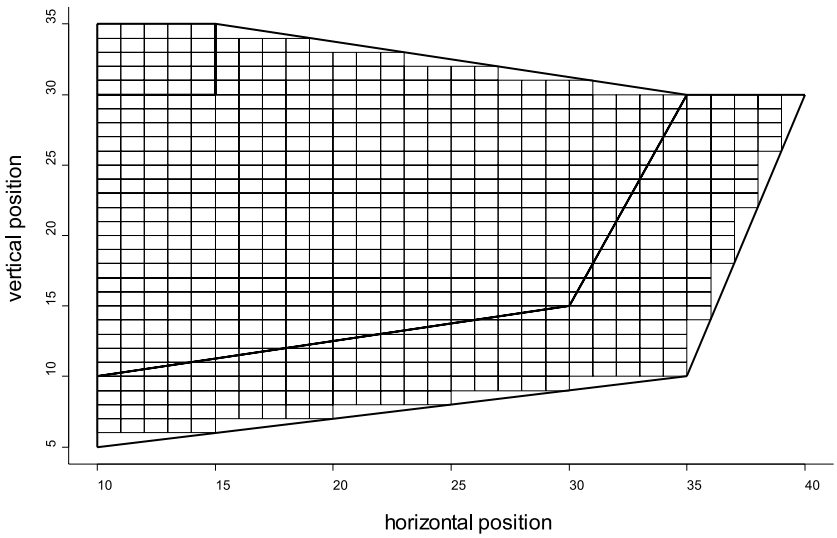


Figure 1.3 For simple random sampling, it is better to first generate a grid of plots that covers the entire area such as the grid shown here.

Simple random sampling is an easy method to select the sampling positions (it is easy to generate random numbers), but it may not be efficient in all cases. Although simple random sampling is the basis for all other sampling methods, it is rarely optimal for biodiversity surveys as described next. Simple random sampling may result in selecting all your samples within areas with the same environmental characteristics, so that you can not test your hypothesis efficiently. If you are testing a hypothesis about a relationship between diversity and landuse, then it is better to stratify by the type of landuse (see below: stratified sampling). You can see in Figure 1.2 that one type of landuse was missed by the random sampling procedure. A procedure that ensures that all types of landuse are included is better than repeating the random sampling procedure until you observe that all the types of landuse were included (which is not simple random sampling any longer).

It may also happen that the method of using random numbers to select the positions of quadrats will cause some of your sample units to be selected in positions that are very close to each other. In the example of Figure 1.2, two sample plots actually overlap. To avoid such problems, it is theoretically better to first generate the population of all the acceptable sample plots, and then take a simple random sample of those. When you use random numbers to generate the positions, the population of all possible sample plots is infinite, and this is not the best approach. It is therefore better to first generate a **grid** of plots that covers the entire survey area, and then select the sample plots at random from the grid.

Figure 1.3 shows the grid of plots from which all the sample plots can be selected. We made the choice to include only grid cells that fell completely into the area. Another option would be to include plots that included boundaries, and only sample the part of the grid cell that falls completely within the survey area – and other options also exist.

Once you have determined the grid, then it becomes relatively easy to randomly select sample plots from the grid, for example by giving all the plots on the grid a sequential number and then randomly selecting the required number of sample plots with a random number. Figure 1.4 shows an example of a random selection of sample plots from the grid. Note that although we avoided ending up with overlapping sample plots, some sample plots were adjacent to each other and one type of landuse was not sampled.

Note also that the difference between selecting points at random and gridding first will only be noticeable when the quadrat size is not negligible compared to the study area. A pragmatic solution to overlapping quadrats selected by simple random sampling of points would be to reject the second sample of the overlapping pair and choose another random location.

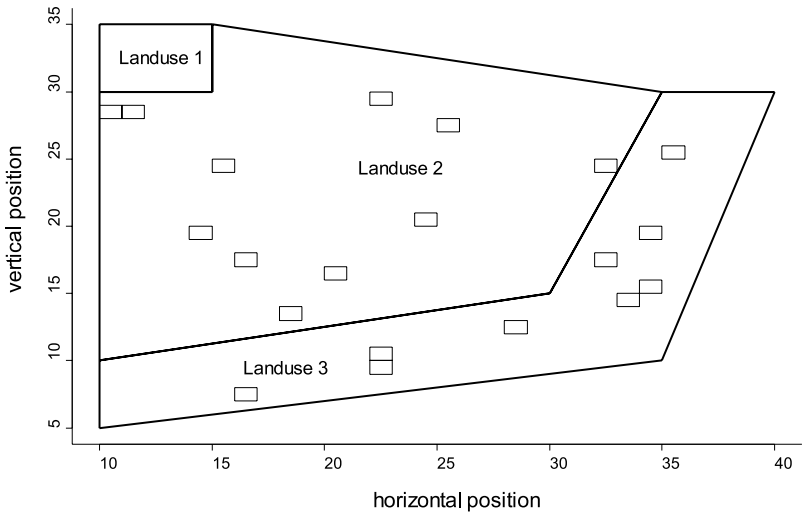


Figure 1.4. Simple random sampling from the grid shown in Figure 1.3.

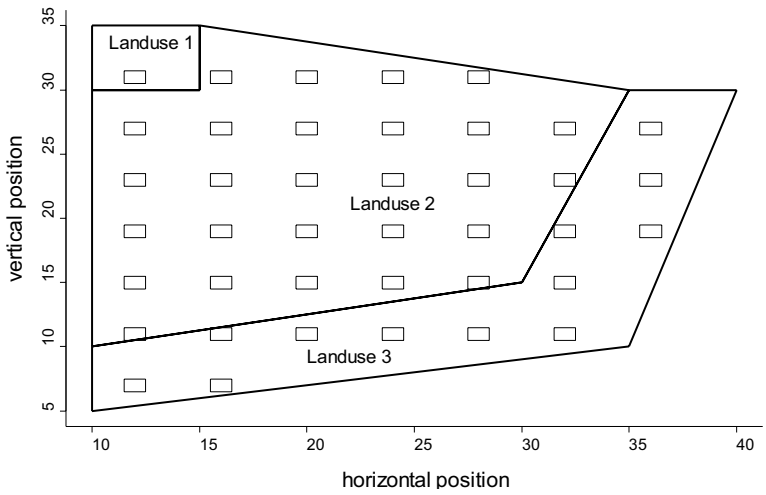


Figure 1.5 Systematic sampling ensures that data are collected from the entire survey area.

Systematic sampling

Systematic or regular sampling selects sample plots at regular intervals. Figure 1.5 provides an example. This has the effect of spreading the sample out evenly through the study area. A square or rectangular grid will also ensure that sample plots are evenly spaced.

Systematic sampling has the advantage over random sampling that it is easy to implement, that the entire area is sampled and that it avoids picking sample plots that are next to each other. The method may be especially useful for finding out where a variable undergoes rapid changes. This may particularly be interesting if you sample along an environmental gradient, such as altitude, rainfall or fertility gradients. For such problems systematic sampling is probably more efficient – but remember that we are not able in this chapter to provide a key to the best sampling method.

You could use the same grid depicted in Figure 1.5 for simple random sampling, rather than the complete set of plots in Figure 1.3. By using this approach, you can guarantee that sample plots will not be selected that are too close together. The grid allows you to control the minimum distance between plots. By selecting only a subset of sample plots from the entire grid, sampling effort is reduced. For some objectives, such combination of simple random sampling and regular sampling intervals will offer the best approach. Figure 1.6 shows a random selection of sample plots from the grid depicted in Figure 1.5.

If data from a systematic sample are analysed as if they came from a random sample, inferences may be invalidated by correlations between neigbouring observations. Some analyses of systematic samples will therefore require an explicitly spatial approach.

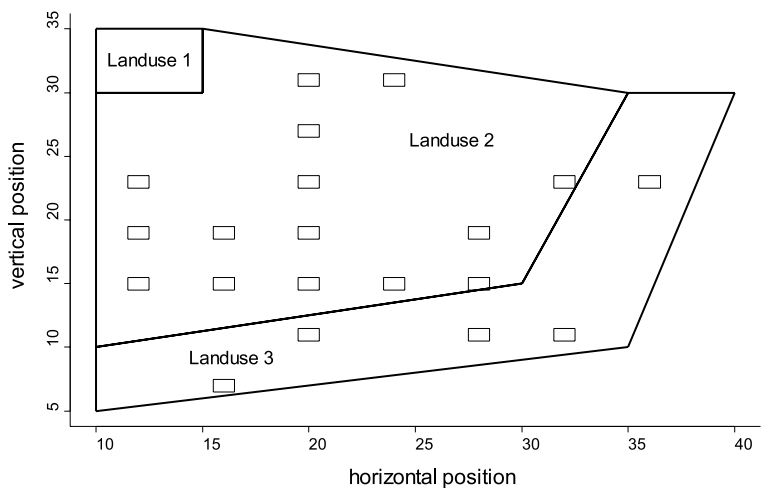


Figure 1.6 Random selection of sample plots from a grid. The same grid was used as in Figure 1.5.

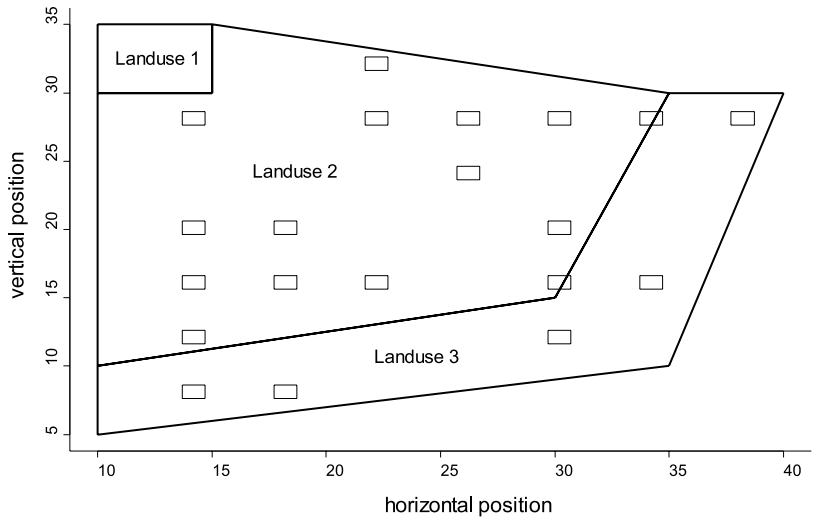


Figure 1.7. Systematic sampling after random selection of the position of the first sample plot.

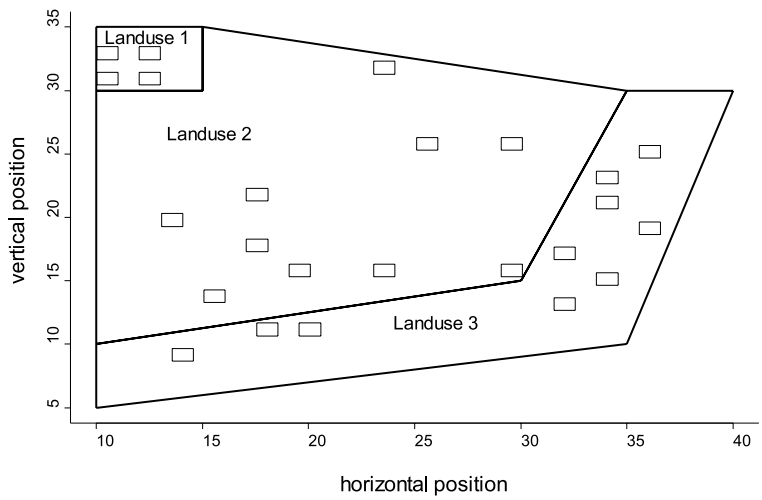


Figure 1.8 Stratified sampling ensures that observations are taken in each stratum. Sample plots are randomly selected for each landuse from a grid.

Another problem that could occur with systematic sampling is that the selected plots coincide with a periodic pattern in the study area. For example, you may only sample in valley bottoms, or you may never sample on boundaries of fields. You should definitely be alert for such patterns when you do the actual sampling. It will usually be obvious if a landscape can have such regular patterns.

Systematic sampling may involve no randomization in selecting sample plots. Some statistical analysis and inference methods are not then suitable. An element of randomization can be introduced in your systematic sampling by **selecting the position of the grid at random**. Figure 1.7 provides an example of selecting sample plots from a sampling grid with a random origin resulting in the same number of sample plots and the same minimum distance between sample plots as in Figure 1.6.

Stratified sampling

Stratified sampling is an approach in which the study area is subdivided into different **strata**, such as the three types of landuses of the example (Landuse 1, Landuse 2 and Landuse 3, figures 1.1-1.9). Strata do not overlap and cover the entire survey area. Within each stratum, a random or systematic sample can be taken. Any of the sampling approaches that were explained earlier can be used, with the only difference that the sampling approach will now be applied to each stratum instead of the entire survey area. Figure 1.8 gives an example of stratified random sampling with random selection of maximum 10 sample plots per stratum from a grid with random origin.

Stratified sampling ensures that data are collected from each stratum. The method will also ensure that enough data are collected from each stratum. If stratified sampling is not used, then a rare stratum could be missed or only provide one observation. If a stratum is very rare, you have a high chance of missing it in the sample. A stratum that only occupies 1% of the survey area will be missed in over 80% of simple random samples of size 20.

Stratified sampling also avoids sample plots being placed on the boundary between the strata so that part of the sample plot is in one stratum and another part is in another stratum. You could have noticed that some sample plots included the boundary between Landuse 3 and Landuse 2 in Figure 1.7. In Figure 1.8, the entire sample plot occurs within one type of landuse.

Stratified sampling can increase the precision of estimated quantities if the strata coincide with some major sources of variation in your area. By using stratified sampling, you will be more certain to have sampled across the variation in your survey area. For example, if you expect that species richness differs with soil type, then you better stratify by soil type.

Stratified sampling is especially useful when your research hypothesis can be described in terms of differences that occur between strata. For example, when your hypothesis is that landuse influences species richness, then you should stratify by landuse. This is the best method of obtaining observations for each category of landuse that will allow you to test the hypothesis.

Stratified sampling is not only useful for testing hypotheses with categorical explanatory variables, but also with **continuous** explanatory variables. Imagine that you wanted to investigate the influence of rainfall on species richness. If you took a simple random sample, then you would probably obtain many observations with near average rainfall and few towards the extremes of the rainfall range. A stratified approach could guarantee that you take plenty of observations at high and low rainfalls, making it easier to detect the influence of rainfall on species richness.

The main disadvantage of stratified sampling is that you need information about the distribution of the strata in your survey area. When this information is not available, then you may need

to do a survey first on the distribution of the strata. An alternative approach is to conduct systematic surveys, and then do some gap-filling afterwards (see below: dealing with covariates and confounding).

A modification of stratified sampling is to use **gradient-oriented transects** or **gradsects** (Gillison and Brewer 1985; Wessels et al. 1998). These are transects (sample plots arranged on a line) that are positioned in a way that steep gradients are sampled. In the example of Figure 1.8, you could place gradsects in directions that ensure that the three landuse categories are included. The advantage of gradsects is that travelling time (cost) can be minimized, but the results may not represent the whole study area well.

Sample size or the number of sample units

Choosing the sample size, the number of sampling units to select and measure, is a key part of planning a survey. If you do not pay attention to this then you run two risks. You may collect far more data than needed to meet your objectives, wasting time and money. Alternatively, and far more common, you may not have enough information to meet your objectives, and your research is inconclusive. Rarely is it possible to determine the exact sample size required, but some attempt at rational choice should be made.

We can see that the sample size required must depend on a number of things. It will depend on the complexity of the objectives – it must take more data to unravel the complex relationships between several response and explanatory variables than it takes to simply compare the mean of two groups. It will depend on the variability of the response being studied – if every sample unit was the same we only need to measure one to have all the information! It will also depend on how precisely you need to know answers – getting a good estimate of a small difference between two strata will require more data than finding out if they are roughly the same.

If the study is going to compare different strata or conditions then clearly we need observations in each stratum, or representing each set of conditions. We then need to plan for repeated observations within a stratum or set of conditions for four main reasons:

1. In any analysis we need to give some indication of the precision of results and this will depend on variances. Hence we need enough observations to estimate relevant variances well.
2. In any analysis, a result estimated from more data will be more precise than one estimated from less data. We can increase precision of results by increasing the number of relevant observations. Hence we need enough observations to get sufficient precision.
3. We need some 'insurance' observations, so that the study still produces results when unexpected things happen, for example some sample units can not be measured or we realize we will have to account for some additional explanatory variables.
4. We need sufficient observations to properly represent the study area, so that results we hope to apply to the whole area really do have support from all the conditions found in the area.

Of these four, 1 and 2 can be quantified in some simple situations. It is worth doing this quantification, even roughly, to make sure that your sample size is at least of the right order of magnitude.

The first, 1, is straightforward. If you can identify the variances you need to know about, then make sure you have enough observations to estimate each. How well you estimate a variance is determined by its degrees of freedom (*df*), and a minimum of 10 *df* is a good working rule. Get help finding the degrees of freedom for your sample design and planned analysis.

The second is also straightforward in simple cases. Often an analysis reduces to comparing means between groups or strata. If it does, then the

mathematical relationship between the number of observations, the variance of the population sampled and the precision of the mean can be exploited. Two approaches are used. You can either specify how well you want a difference in means to be estimated (for example by specifying the width of its confidence interval), or you can think of the hypothesis test of no difference. The former tends to be more useful in applied research, when we are more interested in the size of the difference than simply whether one exists or not. The necessary formulae are encoded in some software products.

An example from R is shown immediately below, providing the number of sample units (n) that will provide evidence for a difference between two strata for given significance and power of the t-test that will be used to test for differences, and given standard deviation and difference between the means. The formulae calculated a fractional number of 16.71 sample units, whereas it is not possible in practice to take 16.71 sample units per group. The calculated fractional number could be rounded up to 17 or 20 sample units. We recommend interpreting the calculated sample size in relative terms, and concluding that 20 samples will probably be enough whereas 100 samples would be too many.

```
Two-sample t test power calculation

              n = 16.71477
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

Sample size in each stratum

A common question is whether the survey should have the same number of observations in each stratum. The correct answer is once again that it all depends. A survey with the same number of observations per stratum will be optimal if the objective is to compare the different strata and if you do not have additional information or hypotheses on other sources of variation. In many other cases, it will not be necessary or practical to ensure that each stratum has the same number of observations.

An alternative that is sometimes useful is to make the number of observations per stratum proportional to the size of the stratum, in our case its area. For example, if the survey area is stratified by landuse and one category of landuse occupies 60% of the total area, then it gets 60% of sample plots. For the examples of sampling given in the figures, landuse 1 occupies 3.6% of the total area (25/687.5), landuse 2 occupies 63.6% (437.5/687.5) and landuse 3 occupies 32.7% (225/687.5). A possible proportional sampling scheme would therefore be to sample 4 plots in Landuse 1, 64 plots in Landuse 2 and 33 plots in Landuse 3.

One advantage of taking sample sizes proportional to stratum sizes is that the average for the entire survey area will be the average of all the sample plots. The sampling is described as **self-weighting**. If you took equal sample size in each stratum and needed to estimate an average for the whole area, you would need to weight each observation by the area of each stratum to arrive at the average of the entire area. The calculations are not very complicated, however.

Some researchers have suggested that taking larger sample sizes in larger strata usually results in capturing more biodiversity. This need not be the case, for example if one landuse which happens to occupy a small area contains much of the diversity. However, most interesting research objectives require more than simply finding the diversity. If the objective is to find as many species as possible, some different sampling schemes could be more effective. It may be better to use an adaptive method where the position of new samples is guided by the results from previous samples.

Simple random sampling will, in the long run, give samples sizes in each stratum proportional to the stratum areas. However this may not happen in any particular selected sample. Furthermore, the strata are often of interest in their own right, and more equal sample sizes per stratum may be more appropriate, as explained earlier. For these reasons it is almost always worth choosing strata and their sample sizes, rather than relying on simple random sampling.

Dealing with covariates and confounding

We indicated at the beginning of this chapter that it is difficult to make conclusions about cause-effect relationships in surveys. The reason that this is difficult is that there may be confounding variables. For example, categories of landuse could be correlated with a gradient in rainfall. If you find differences in species richness in different landuses it is then difficult or impossible to determine whether species richness is influenced by rainfall or by landuse, or both. Landuse and rainfall are said to be confounded.

The solution in such cases is to attempt to break the strong correlation. In the example where landuse is correlated with rainfall, then you could attempt to include some sample plots that have another combination of landuse and rainfall. For example, if most forests have high rainfall and grasslands have low rainfall, you may be able to find some low rainfall forests and high rainfall grasslands to include in the sample. An appropriate sampling scheme would then be to stratify by combinations of both rainfall and landuse (e.g. forest with high, medium or low rainfall or grassland with high, medium or low rainfall) and take a sample from each stratum. If there simply are no high rainfall grasslands or low rainfall forests then accept that it is not possible to understand the separate effects of rainfall and landuse, and modify the objectives accordingly.

An extreme method of breaking confounding is to **match sample plots**. Figure 1.9 gives an example.

The assumption of matching is that confounding variables will have very similar values for paired sample plots. The effects from the confounding variables will thus be filtered from the analysis.

The disadvantage of matching is that you will primarily sample along the edges of categories. You will not obtain a clear picture of the overall biodiversity of a landscape. Remember, however, that matching is an approach that specifically investigates a certain hypothesis.

You could add some observations in the middle of each stratum to check whether sample plots at the edges are very different from sample plots at the edge. Again, it will depend on your hypothesis whether you are interested in finding this out.

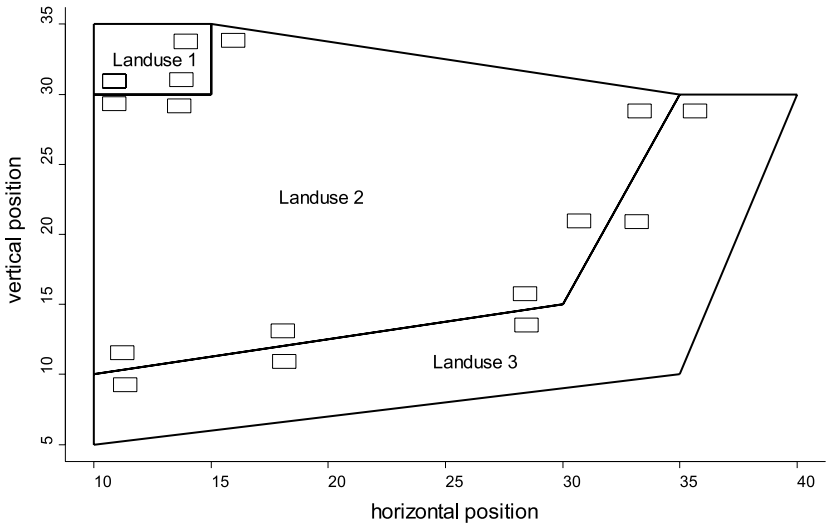


Figure 1.9 Matching of sample plots breaks confounding of other variables.

Pilot testing of the sampling protocol

The best method of choosing the size and shape of your sample unit is to start with a **pilot phase** in your project. During the pilot phase all aspects of the data collection are tested and some preliminary data are obtained.

You can evaluate your sampling protocol after the pilot phase. You can see how much variation there is, and base some modifications on this variation. You could calculate the required sample sizes again. You could also opt to modify the shape, size or selection of sample plots.

You will also get an idea of the time data collection takes per sample unit. Most importantly, you could make a better estimation of whether you will be able to test your hypothesis, or not, by already conducting the analysis with the data that you already have.

Pilot testing is also important for finding out all the non-statistical aspects of survey design and management. These aspects typically also have an important effect on the overall quality of the data that you collect.

References

Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, Foster RB, Nuñez P, Aguilar S, Valencia R, Villa G, Muller-Landau HC, Losos E, and Hubbell SP. 2002. Beta-diversity in tropical forest trees. *Science* 295: 666–669.

Feinsinger P. 2001. *Designing field studies for biodiversity conservation.* Washington: The Nature Conservancy.

Gillison AN and Brewer KRW. 1985. The use of gradient directed transects or gradsects in natural resource surveys. *Journal of Environmental Management* 20: 103-127.

Gotelli NJ and Ellison AM. 2004. *A primer of ecological statistics.* Sunderland: Sinauer Associates. (recommended as first priority for reading)

Hayek LAC and Buzas MA. 1997. *Surveying natural populations.* New York: Columbia University Press.

Laurance WF, Laurance SG, Ferreira LV, Rankin-de Merona JM, Gascon C and Loverjoy TE. 1997. Biomass collapse in Amazonian forest fragments. *Science* 278: 1117-1118.

Pyke CR, Condit R, Aguilar S and Lao S. 2001. Floristic composition across a climatic gradient in a neotropical lowland forest. *Journal of Vegetation Science* 12: 553-566.

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press.

Sheil D, Ducey MJ, Sidiyasa K and Samsoedin I. 2003. A new type of sample unit for the efficient assessment of diverse tree communities in complex forest landscapes. *Journal of Tropical Forest Science* 15: 117-135.

Sutherland WJ. 1996. *Ecological census techniques: a handbook.* Cambridge: Cambridge University Press.

Tuomisto H, Ruokolainen K and Yli-Halla M. 2003. Dispersal, environment and floristic variation of western Amazonian forests. *Science* 299: 241-244.

Underwood AJ. 1997. *Experiments in ecology: their logical design and interpretation using analysis of variance.* Cambridge: Cambridge University Press.

Wessels KJ, Van Jaarsveld AS, Grimbeek JD and Van der Linde MJ. 1998. An evaluation of the gradsect biological survey method. *Biodiversity and Conservation* 7: 1093-1121.

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

See in chapter 3 how Biodiversity.R can be loaded onto your computer.

To load polygons with the research areas:

```
area <- array(c(10,10,15,35,40,35,5,35,35,30,30,10),
 dim=c(6,2))

landuse1 <- array(c(10,10,15,15,30,35,35,30), dim=c(4,2))

landuse2 <- array(c(10,10,15,15,35,30,10,30,30,35,30,15),
 dim=c(6,2))

landuse3 <- array(c(10,10,30,35,40,35,5,10,15,30,30,10),
 dim=c(6,2))

window <- array(c(15,15,30,30,10,25,25,10),dim=c(4,2))
```

To plot the research area:

```
plot(area[,1], area[,2], type="n", xlab="horizontal position",
 ylab="vertical position", lwd=2, bty="l")

polygon(landuse1)

polygon(landuse2)

polygon(landuse3)
```

To randomly select sample plots in a window:

```
spatialsample(window, method="random", n=20, xwidth=1,
 ywidth=1, plotit=T, plothull=T)
```

To randomly select sample plots in the survey area:

```
spatialsample(area, method="random", n=20, xwidth=1, ywidth=1,
 plotit=T, plothull=F)
```

To select sample plots on a grid:

```
spatialsample(area, method="grid", xwidth=1, ywidth=1,
 plotit=T, xleft=10.5, ylower=5.5, xdist=1, ydist=1)

spatialsample(area, method="grid", xwidth=1, ywidth=1,
 plotit=T, xleft=12, ylower=7, xdist=4, ydist=4)
```

To randomly select sample plots from a grid:

```
spatialsample(area, method="random grid", n=20, xwidth=1,
 ywidth=1, plotit=T, xleft=10.5, ylower=5.5, xdist=1, ydist=1)

spatialsample(area, method="random grid", n=20, xwidth=1,
 ywidth=1, plotit=T, xleft=12, ylower=7, xdist=4, ydist=4)
```

To select sample plots from a grid with random start:

```
spatialsample(area, method="random grid", n=20, xwidth=1,
 ywidth=1, plotit=T, xdist=4, ydist=4)
```

To randomly select maximum 10 sample plots from each type of landuse:

```
spatialsample(landuse1, n=10, method="random", plotit=T)

spatialsample(landuse2, n=10, method="random", plotit=T)

spatialsample(landuse3, n=10, method="random", plotit=T)
```

To randomly select sample plots from a grid within each type of landuse. Within each landuse, the grid has a random starting position:

```
spatialsample(landuse1, n=10, method="random grid", xdist=2,
 ydist=2, plotit=T)

spatialsample(landuse2, n=10, method="random grid", xdist=4,
 ydist=4, plotit=T)

spatialsample(landuse3, n=10, method="random grid", xdist=4,
 ydist=4, plotit=T)
```

To calculate sample size requirements:

```
power.t.test(n=NULL, delta=1, sd=1, sig.level=0.05, power=0.8,
 type="two.sample")

power.t.test(n=NULL, delta=0.5, sd=1, sig.level=0.05,
 power=0.8, type="two.sample")

power.anova.test(n=NULL, groups=4, between.var=1, within.var=1,
 power=0.8)

power.anova.test(n=NULL, groups=4, between.var=2, within.var=1,
 power=0.8)
```

To calculate the area of a polygon:

```
areapl(landuse1)
```

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.

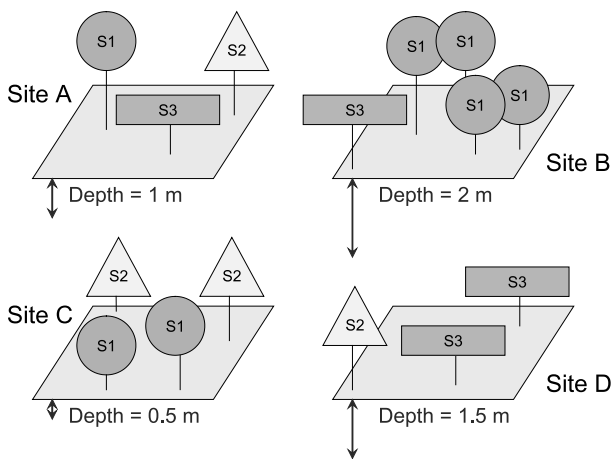


Figure 2.1 A simplified example of information recorded on species and environmental data.

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).

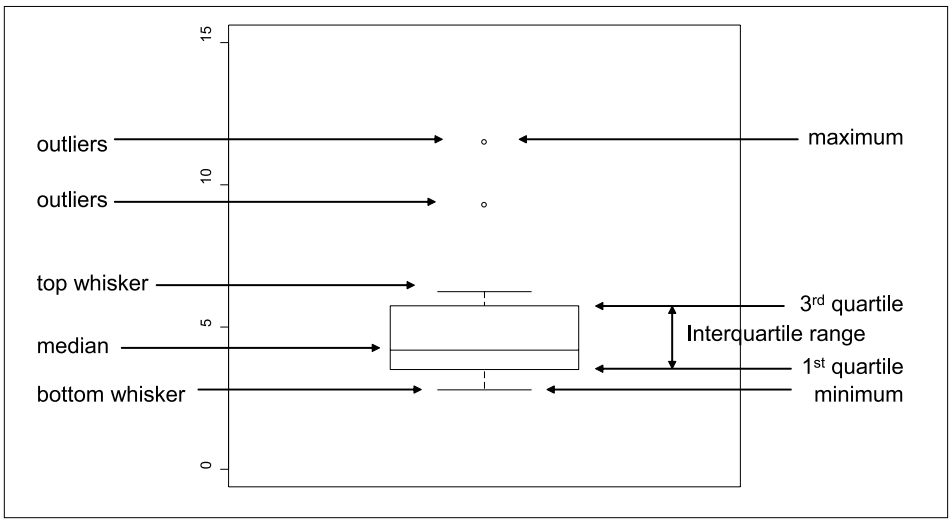


Figure 2.2 Summary of a quantitative variable as a boxplot. The variable that is summarized is the thickness of the A1 horizon of Table 2.2.

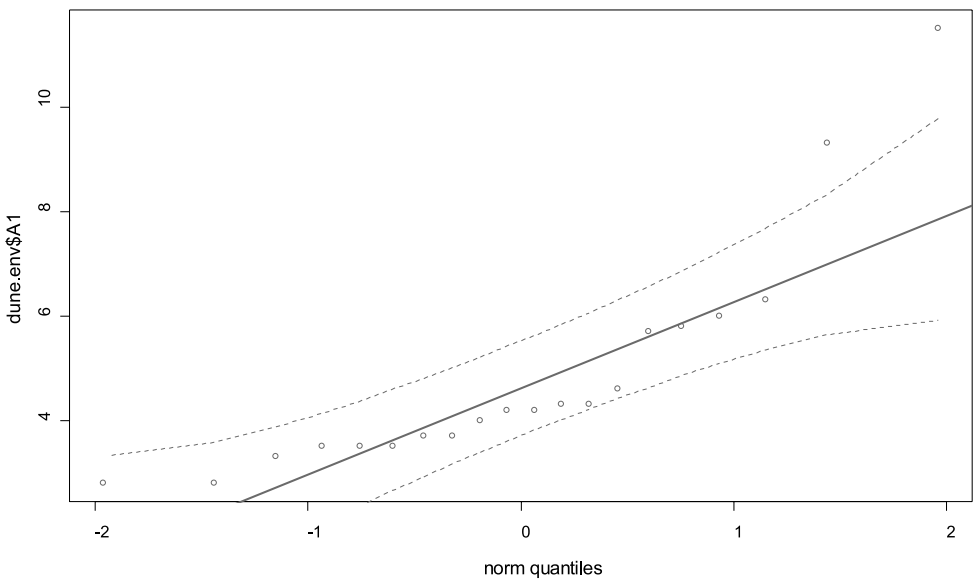


Figure 2.3 Summary of a quantitative variable as a Q-Q plot. The variable that is summarized is the thickness of the A1 horizon of Table 2.2. The two outliers (upper right-hand side) correspond to the outliers of Figure 2.2.

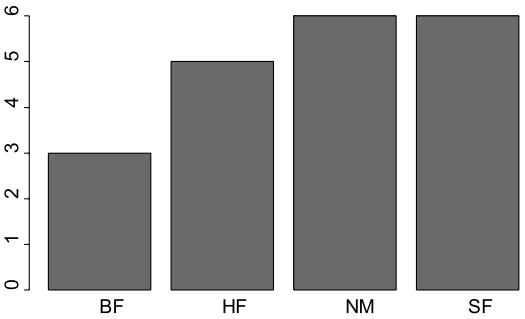


Figure 2.4 Summary of a categorical variable by a bar plot. The management of Table 2.2 is summarized.

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.

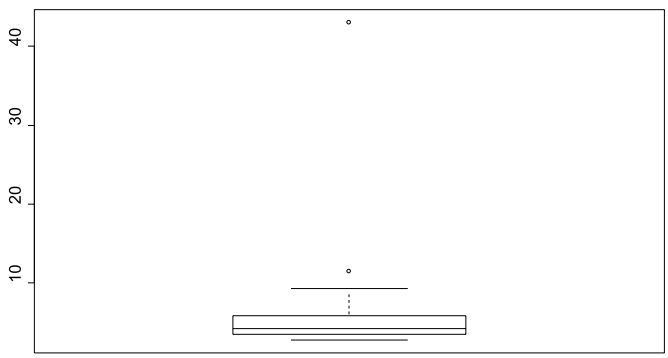


Figure 2.5 Checking for exceptional observations.

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.

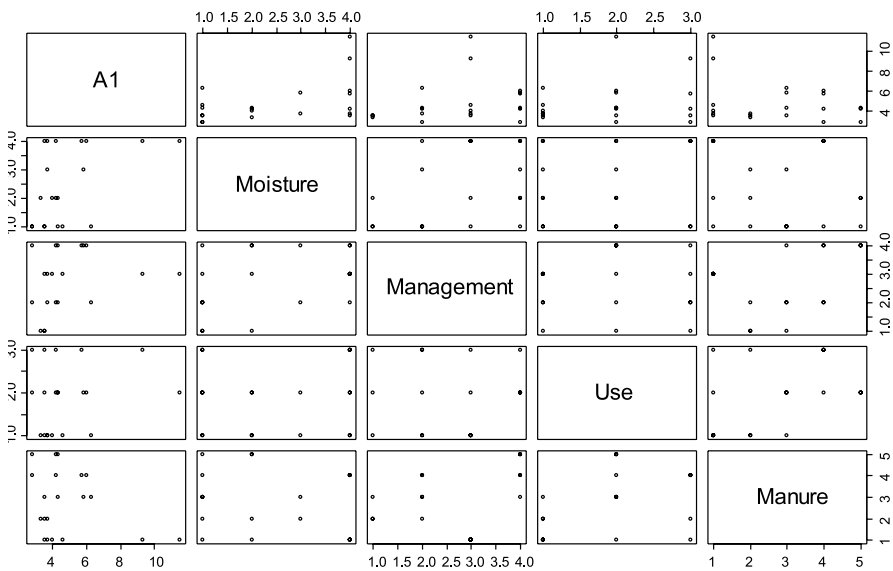


Figure 2.6 Checking for exceptional data by pairwise comparisons of the variables of Table 2.2.

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)
```

CHAPTER 3

Doing biodiversity analysis with Biodiversity.R

Doing biodiversity analysis with Biodiversity.R

This chapter describes how the analyses presented in this manual can be performed with the Biodiversity.R software.

This special attention to Biodiversity.R does not mean that other software packages can not be used for biodiversity analysis. In 2003, a Biodiversity Analysis Package CD-ROM was produced to provide several software packages that are very good for biodiversity analysis. However, some of the software packages had a more limited scope in the analyses that they supported. Some of the software packages had not developed a graphical user interface, which caused some problems for teaching the analysis. Some types of analysis could not be performed in the software provided on the CD-ROM. For these reasons, the Biodiversity.R software was developed, including a graphical user interface. All the analyses described in this manual can be conducted with Biodiversity.R.

What is Biodiversity.R?

Biodiversity.R is software that does all the biodiversity analyses described in this manual. It needs to be loaded into the R statistical software. R is a software that was developed to allow for many different types of statistical analysis. It is very similar to the S and S-Plus statistical software. It is free software, as is Biodiversity.R. The software is also open, so that you can check how calculations are done, and the graphics are quite advanced.

How do I run Biodiversity.R?

The CD-ROM that is provided with this manual contains an installed version of Biodiversity.R. You can run Biodiversity.R from the CD-ROM by clicking on the file *Run-Biodiversity.bat* in the Biodiversity.R folder on the CD-ROM. Alternatively you need to install Biodiversity.R on another location on your computer first.

Each time that you run R, you need to load the *Rcmdr* package (see below: how do I run Biodiversity.R) to access the Biodiversity.R graphical user interface.

How do I install Biodiversity.R?

Two options are provided for installing Biodiversity.R. When you click on *Install-Biodiversity.bat*, then all files will be installed under the *Program Files* folder of your C drive.

The alternative method is to install Biodiversity.R in steps. The files that are used during installation are all listed on the CD-ROM in the *Biodiversity.R/installation files* folder (Figure 3.1, see next page).

The first thing that you need to do is to install the R software. You can also obtain the software from the following website: http://cran.r-project.org. If you are using Windows, then you can download R from: http://cran.r-project.org/bin/windows/base. You will find a file there with a name similar to **rw2010-1.exe**. You need to run the file to install R. Note that it may be safer to close all other programs when you install R.

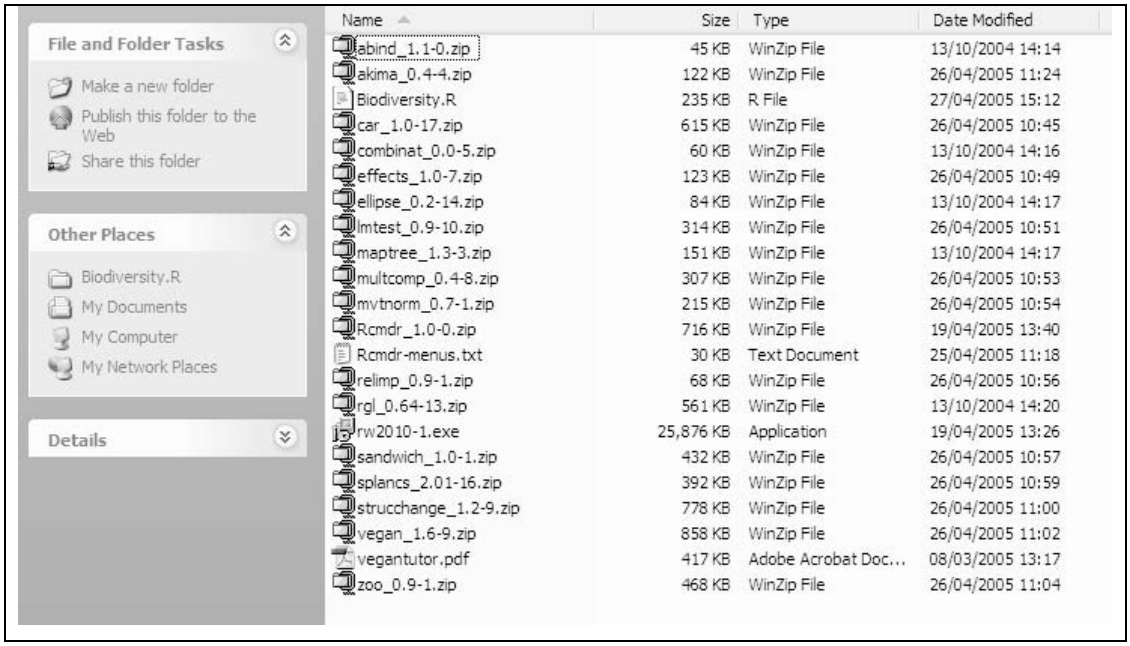


Figure 3.1. All the files used during installation are provided in the *Installation files* folder

During the installation, make sure that you opt to install the support files for library(tcltk).

After you have installed R, then you can run it. R is a language that is run by typing in commands. It does not have an extensive user interface.

Some additional software can be loaded onto R. These addins are called packages or libraries. Some packages already come with the installation version of R. Some other packages need to be downloaded.

You have various options to install these additional packages. The first option is install the additional packages from the CD-ROM. Within the R program, you need to go to the top menu, choose Packages, and then choose menu option: Install packages from local zip files…. Figure 3.2 shows how this can be done.

Note that we will not put many images of menus in this manual, so that some space will be saved. We will describe the selection shown in Figure 3.2 as: "You select the R menu option: Packages > Install package(s) from local zip files…".

You will obtain a list of available packages under the *Installation files* folder of the CD-ROM. You can install all the packages by selecting all of them at the same time (click on the CTRL and A keys simultaneously).

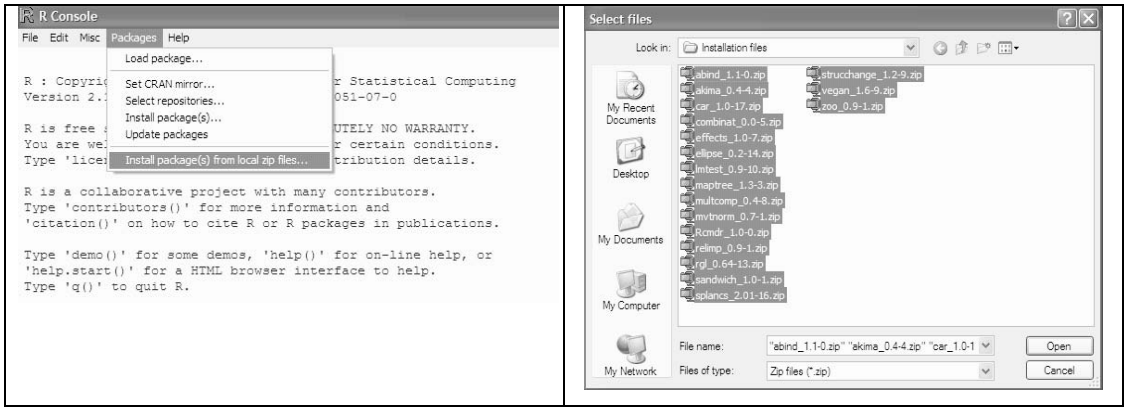


Figure 3.2. Installing other packages onto R. This menu option is described in the text as: "Packages > Install package(s) from local zip files...".

The alternatives are to download the packages from R and then install them manually afterwards, or to install the packages directly from the CRAN website. You can download the files for the different packages from the same website you downloaded the R package itself from: http://cran.r-project.org. You will see a link to "contributed packages". If you are using Windows, then you can download the packages from: http://cran.r-project.org/bin/windows/contrib.

You only need to install a package once after installing R. When you run R, you will not be able to access the functions of contributing packages, unless you follow the menu option of: Packages > Load package…, or if you give the appropriate command in R of loading the package. For the `vegan` library, you need to select this package after menu option: Packages > Load package… or you need to type `library(vegan)` each time that you started using R and wanted to access the functions of this library.

Note that we will use a different font for results and commands in R and Biodiversity. R. For example, `library(vegan)` points to a command in R. Any information in this font will appear in the R console.

The packages that you need to install into R are:

- abind
- akima
- car
- combinat
- effects
- ellipse
- lmtest
- maptree
- multcomp
- mvtnorm
- rcmdr
- relimp
- rgl
- sandwich
- splancs
- strucchange
- vegan
- zoo

These are quite a few packages, but this also means that various types of analysis can be done in R. Especially important is `vegan` as it allows

for many of the analyses of Biodiversity.R. `Rcmdr` is also an important package as it allows for the R-commander graphical interface. (For the advanced user: The R-commander runs best under Single-Document Interface. You can set this option by setting "MDI = no" in file *C:\Program Files\R\rw2010\etc\Rconsole.)*

The final (and probably most complicated) step in installing Biodiversity.R is to copy two files into the *library\Rcmdr\etc* folder of R. In case that you installed R under the program files of your C drive, then you can find the *Rcmdr* directory under *C:\Program files\R\rw2010\library\Rcmdr\etc*. You need to put the following files into this directory: *Biodiversity.R* and *Rcmdr-menus.txt*. The *Rcmdr-menus.txt* will already exist in the library, so you need to **replace** the *Rcmdr-menus.txt* with the file that is provided on the CD-ROM. You could use a program such as the Windows Explorer to copy the files.

When you will have completed all these steps, Biodiversity.R will have been installed.

In short, you need to follow these steps to install and run Biodiversity.R:

1. Install R
2. Install the required contributing packages
3. Copy *Biodiversity.R* and *Rcmdr-menus.txt* into *library\Rcmdr\etc*

How do I run Biodiversity.R?

Obviously you need to run R first.

To run Biodiversity.R with the menu options, you will need to load the R-commander, either by the command `library(Rcmdr)` or by menu option: Packages > Load package…

Each time that you want to use Biodiversity.R, you need to load the `Rcmdr` package again after launching R.

The species and environmental datasets

After you have loaded Biodiversity.R, you will not be able to conduct any type of analysis before you have chosen the species and environmental datasets. As we saw in chapter 2 on data preparation, the analyses in this manual will be conducted with these two datasets.

The R-commander was designed to only use one dataset. This dataset is called the active dataset. Biodiversity.R uses two datasets, the species and environmental datasets. When developing it the decision was taken to make the environmental dataset of Biodiversity.R always to be the active dataset of the R-commander. When a new dataset is chosen to be the new active dataset for the R-commander, then it also becomes the new environmental dataset of Biodiversity.R. With the menu options of the R-commander, various manipulations can be done on the active dataset, including importing an active dataset and saving an active dataset. You can do the same manipulations to the species dataset, but first you need to set the species dataset to also be the active dataset (and thus also the environmental dataset).

An example of an analysis in Biodiversity.R

Descriptions of the way in which each type of analysis can be done in R are provided at the end of each chapter. As an example and a test whether you installed Biodiversity.R correctly, you could run following analysis. This example calculates a species accumulation curve for the dune dataset. This dataset is provided with `vegan`. The result that you will obtain is shown in the chapter on species richness.

You need to follow these menu options:

Biodiversity > Environmental Matrix > Select environmental matrix

→Select the dune.env dataset

Biodiversity > Community Matrix > Select community matrix

→Select the dune dataset

Biodiversity > Analysis of biodiversity > Species accumulation curves

→Accumulation method: exact

→Press on the "OK" button

→Press on the "Plot" button

The last window that allows you to calculate a species accumulation curve is shown in Figure 3.3.

Note that there are separate "OK" and "Plot" options for different menus. When you press "OK", then the analysis is calculated and the results are saved under name the specified for the result. The "Plot" options apply to the name of the result on top. It is not necessary to redo the analysis to plot results, and you can also plot information from earlier analyses by changing the name of the result on top.

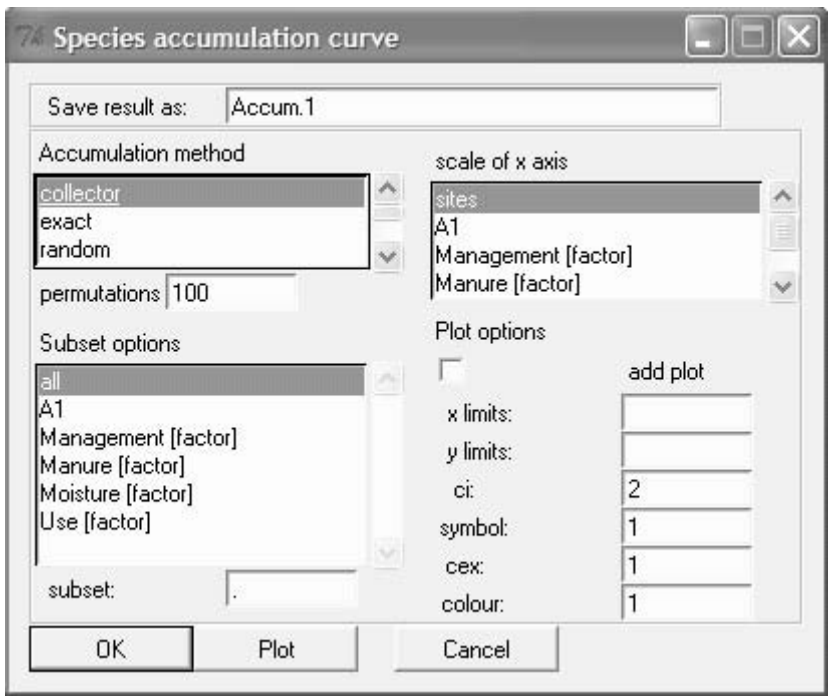


Figure 3.3. The window interface for the calculation of species accumulation curves

The alternative option to using the menu interface is to type in the following commands:

```
Accum.1 <- accumresult(dune,
  method='exact')

Accum.1

plot(Accum.1)
```

Note that R is case sensitive: `Plot(accum.1)` will not work!

If you chose the first option, then you will see that the same commands were generated by the R-commander. The R-commander is actually an interface that allows you to find appropriate commands that can be used in R. If you are very familiar with the commands, then you could perform some analyses more quickly by directly typing them in R.

You can see that R commands consist of functions (such as `accumresult` or `specaccum`), followed by arguments that are put between brackets. The command `specaccum(dune)` will calculate a species accumulation for the dune dataset, for instance.

Importing files into Biodiversity.R

Files of various formats can be imported into R. We usually prefer to use text files (files with an extension of .txt), where the columns are separated by tabs. You can create such files from Excel by choosing the menu option in Excel: File > Save as… and then putting "text (tab delimited)".

Be aware that text fields (also names of variables and sites) can only contain one continuous text without spaces. If you have words separated by spaces (such as the genus and species of a plant name), then you need to change these into a continuous text such as changed_into_continuous_text, changedintocontinuoustext or changed.into.continuous.text.

Datasets in R have one particular feature in not having a name for the column that contains the row names of the data. So if your first column of data determines the name of each site, then you should not have any contents in the first row of data with the column names. As a result, the number of fields in the first row of the file will be one less than the other rows of the file. You can check for this feature by looking at the files dune.txt and dune.env.txt, provided under the *Data* folder on the CD-ROM, or look at figures 3.4 and 3.5. If the first row has the same number of columns, then R will conclude that you have not chosen to give row names to your data.

	Achmil	Agrsto	Airpra	Alogen	Antodo	Belper	Brarut	Brohor	Calcus	Chealb
X1	1	0	0	0	0	0	0	0	0	0
X2	3	0	0	2	0	3	0	4	0	0
X3	0	4	0	7	0	2	2	0	0	0
X4	0	8	0	2	0	2	2	3	0	0
X5	2	0	0	0	4	2	2	2	0	0

Figure 3.4. Part of the dune.txt file with the species data for the dune meadow dataset. Note that the column with the names of the sites has no name.

	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

Figure 3.5. Part of the dune.env.txt file with the environmental data for the dune meadow dataset. Note that the column with the names of the sites has no name.

Once you have created the file, then you can import it by menu option: Data > Import data > From text file… and then browse to the right place of your file. You can look at your data by clicking on the "View Dataset" button below the top menu of the R-commander.

Learning more about the commands of R and the contributed packages

As we saw above, R is statistical software that is driven by functions. You may therefore increase the utility of this software by learning more about its functions, especially finding out options, calculation methods and related functions.

There are various options of learning about the functions.

One option is to type in the name of the function by R menu option: Help > R functions (text)… and then typing in "specaccum", for example.

The second option is to go to R menu option: Help > Html help. It may especially be useful to check then under the Packages and the Search Engine.

The third option is to type a ? before the function in R. For example `?specaccum`

The fourth option is to just type the function in R, for example `specaccum`. In that case, you will obtain the actual programming code that was used for the function.

Note that some new functions were created for Biodiversity.R. Documentation for these functions is not provided through the R console, but you can access these files from the CD-ROM that comes with this manual under the *Biodiversity.R* folder (click on Biodiversity.R help.html)

Citing R

When you use R for data analysis, you need to cite using the base package and the additional packages in the publications where you report the results of the data analysis. Information on citation is provided by the `citation` function:

```
> citation()
To cite R in publications use:
  R Development Core Team (2005). R: A language and environment for
  statistical computing.
  R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-
  900051-07-0,
  URL http://www.R-project.org.

> citation("vegan")
To cite package 'vegan' in publications use:
  Oksanen, J., Kindt, R. & O'Hara, R.B. (2005). vegan: Community
  Ecology Package
  version 1.6-9. http://cc.oulu.fi/~jarioksa/.
```

CHAPTER 4

Analysis of species richness

Analysis of species richness

Species richness is the number of species that has been recorded for a specific group of organisms during a specific time period. It also refers to a specific area. That may be the study area, it could be a plot or sampling unit, or an assemblage of sampling units.

Although it is relatively easy to calculate the number of species identified in a set of sample plots, it is trickier to find estimates that can be used to compare the species richness between various subsets in a particular dataset.

How can species richness be calculated?

It is relatively easy to calculate the number of species if the species identity of every individual tree is known. It is good practice to document how species were identified, and ideally where herbarium specimens were deposited. If you do not know the identities of each and every species, then you could opt to only base your analysis on the species with confirmed identities. Some researchers include morphospecies in the species matrix that they analyse: these are species that can be confidently categorized as being different from the other species in the species matrix, but can not be identified.

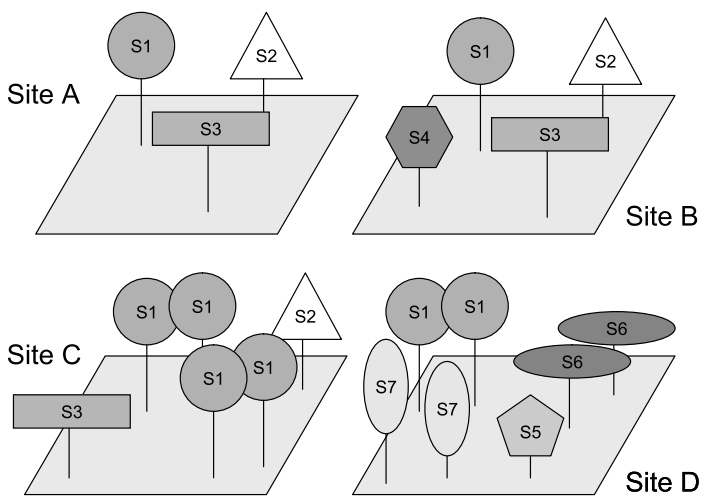


Figure 4.1 Species richness depends on the scale at which it is measured in a landscape. In this example, the number of species on site A is 3, the average number of species per site is 3.5 and the total number of species in the survey area (equaling 4 sites since this is a simplified example) is 7.

Always mention where (and possibly when) you counted the species. Figure 4.1 shows that you can count 3 or 4 species within a single site and 7 species in the total survey. You should always mention the sample size for which you counted the species.

Calculating the total number of species or total species richness

The total number of species can be calculated relatively easy. The total number of species for the dune dataset is 30. If you prepared your species matrix correctly, then the total species richness is the same as the number of columns of your species matrix. In Figure 4.1, the total number of species is 7.

Calculating the total number of species for various subsets in the data

You may be interested in finding differences in the total number of species in various subsets in your data. For example, you could calculate the total number of species for each site. You could also be interested in differences in the species richness for subsets based on the type of management for the dune datasets (see chapter on data preparation). You could for instance hypothesize that the management will influence the species richness. You will obtain the following result:

```
Management n richness
        BF 3       16
        HF 5       21
        NM 6       21
        SF 6       21
```

You could conclude that species richness is lowest for biological farming (BF) since the species richness calculated was 16, and it was 21 for the other categories of management. The `n` in the output stands for the number of sample units of each category. When you check for differences in `n` between the various categories, then you also notice that BF has the lowest value. Since both species richness and the sample size are different, we are not in a good position to compare management classes. The difference could be caused either by an actual difference in species richness, or a perceived difference because of sample size differences.

The reason is that species richness is dependent on the sample size. Study Figure 4.1, for example. You can see that the entire area portrayed contains 7 different species, or a species richness of 7. The 4 sites separately each have a species richness of 3 or 4 species. The average species richness of a site is 3.5.

Calculating the average number of species for various subsets in the data

Instead of being interested in the **total** number of species in each subset of the data, you may be interested in finding out differences in the **average** number of species per site in various subsets in your data. For example, you could compare the average number of species on farms located next to forests with the average number of species on farms that are not adjacent to forests. In this case, the research hypothesis specified that differences were expected in the average number of species per farm and not necessarily in the total number of species.

To investigate differences in the average number of species, regression techniques as described in Chapter 6 offer the best approach to analysis. This approach does not require that sample sizes are equal, as is required to compare the total number of species. Note however that the size of the sample unit should also be the same, and that the size of

the sample unit can be measured in different ways as shown below (section: changing the scale of the horizontal axis for species accumulation curves). To investigate whether differences in species richness between farms close and far away from the forest could be caused by differences in farm size, you could construct species accumulation curves for the farms that are close and farms that are far and compare the curves (see below: using species accumulation curves to compare total species richness of various subsets of data).

Constructing species accumulation curves

It is important to realize that the total number of species of two sites combined is not always equal to the sum of the species on site 1 plus the species on site 2. This will only be the case if site 1 and site 2 do not share any species. If they share some species, then the total number of species on the two sites combined will be smaller than the sum of the richness of each site. Because of this difficulty, we need special types of calculation to find the number of species for combinations of sites.

Species accumulation curves show the species richness for combinations of sites. These curves portray the **average pooled species richness** when 1, 2, … , all sites are combined together.

The reason that the **average** pooled species richness is calculated is that different combinations will have different species richness. For example, for the dune dataset, the species richness of site X1 is 5, for sites X2 and X3 it is 10. In other words, there is not a single value for the species richness of 1 site. For this reason, we calculate the average value. The average species richness for 1 site is 9.85.

When we combine sites X1 and X2, then we obtain a total combined species richness of 10. When we combine sites X3 and X4, then we obtain a total combined species richness of 13. The average species richness for combinations of two sites is 15.11. Note that it not so easy to calculate the average pooled species richness for 2 sites because many different pairs can be formed (20 * 19 / 2 = 190 actually).

Species accumulation curves calculate the average species richness for larger combinations of sites, such as combinations of 3 or combinations of 16 sites, until all possible sizes of combinations were investigated. The result for the dune dataset are shown below.

The output shows that the average richness for all possible combinations of 14 sites is 28.85, whereas for all possible combinations of 17 sites it is 29.51. The `sd` indicates the standard deviation observed as, by chance, some combinations of sites contain small numbers of species, and other combinations contain large numbers of species. Species accumulation patterns are usually

```
Sites    1.000000  2.000000  3.000000  4.000000  5.000000  6.000000  7.000000
Richness 9.850000 15.110526 18.510526 20.937461 22.754321 24.149587 25.239564
sd       2.351064  1.876385  1.572271  1.446958  1.390159  1.353035  1.316480

Sites     8.000000  9.000000 10.000000 11.000000 12.000000 13.0000000
Richness 26.103548 26.798244 27.364962 27.833984 28.227546 28.5620356
sd        1.274903  1.228201  1.176341  1.119344  1.056454  0.9874094

Sites    14.0000000 15.000000 16.000000 17.0000000 18.000000 19.0000000 20
Richness 28.8496388 29.099587 29.319092 29.5140351 29.689474 29.8500000 30
sd        0.9115998  0.828689  0.738092  0.6333903  0.513971  0.3570714  0
```

shown graphically by species accumulation curves. Figure 4.2 shows the species accumulation curve for the dune dataset. It is important to realize that both horizontal and vertical axes show results of pooling or combination: on the horizontal axis the sites are combined (10 = 10 sites combined), whereas on the vertical axis the species are combined (20 = 20 species combined).

The results below are based on exact calculations of the average species richness for combinations of sites. Another approach to approximate this average species richness is a randomisation approach. For example, to find an estimate of the average richness for 3 sites, select 3 sites at random from all 30 sites. Calculate the species richness. Then select another 3 sites at random, calculating a second richness. Repeat maybe 1000 times. Then the overall estimate is the average of the 1000 values. Although this randomisation approach is the classic approach to calculate species accumulation curves and it is based directly on the idea of calculating the average species richness of a series of randomly combined sites, it is better to use the exact method which will usually be faster and more precise.

The results for species accumulation curve based on randomisation (calculated from 1,000 randomisations) for the dune dataset are shown below.

Note that you will obtain slightly different results when you do this analysis again, since selections

```
Sites    1.00000  2.000000  3.000  4.000000  5.000000  6.000000  7.000000
Richness 9.85100 15.163000 18.654 21.012000 22.829000 24.181000 25.276000
sd       2.42088  2.227868  2.271  2.188297  2.053771  1.911516  1.762081

Sites     8.000000  9.000000 10.000000 11.000000 12.000000 13.000000 14.0000000
Richness 26.096000 26.754000 27.323000 27.788000 28.196000 28.509000 28.7980000
sd        1.587862  1.450375  1.358866  1.246840  1.112124  1.015357  0.9219794

Sites    15.0000000 16.0000000 17.0000000 18.0000000 19.0000000 20
Richness 29.0630000 29.2890000 29.4860000 29.6820000 29.8420000 30
sd        0.8376949  0.7470188  0.6404797  0.5149186  0.3649235  0
```

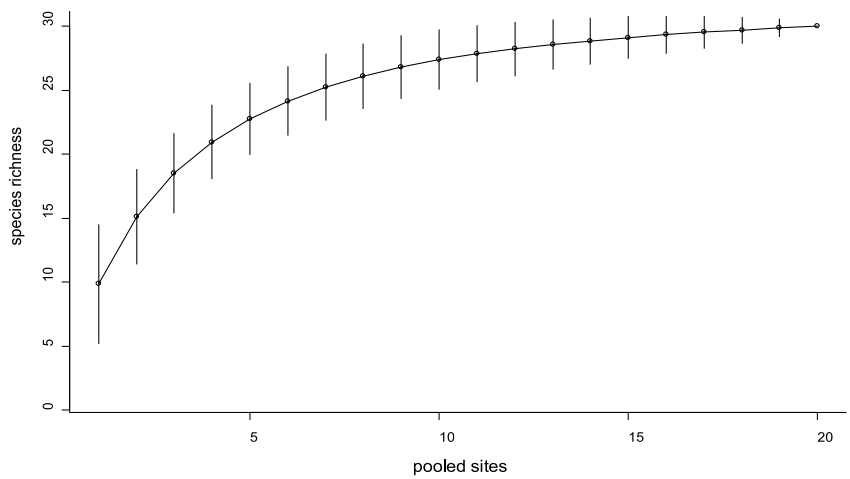


Figure 4.2 Species accumulation curve for the dune meadow dataset. The bars indicate +2 and -2 standard deviations.

are done at random. The `sd` again indicates the standard deviation observed. You can see that the results are similar to the exact results provided earlier. Many randomisations are required to obtain results that will be very close to the exact results. It would not be easy to spot differences between plots of both results, however.

Using species accumulation curves to compare total species richness of various subsets of data

Species accumulation curves are especially useful when comparing species richness for subsets in the data when sample sizes of subsets are different.

When you calculate species accumulation curves for each category of management of the dune dataset separately, then you obtain Figure 4.3. You can observe the total species richness of 16 for a sample size of 3 sites for biological farming, and the total species richness of 21 for a sample size of 5 sites for hobby farming. The results now allow comparing species richness at the same sample size. For example, we can compare the 4 categories at sample size 3. We can observe that hobby farming has substantially larger species richness (19.10) than the three other categories (16, 16.2 and 16.25). More importantly, we can see that there is little evidence now that nature management and standard farming have more species than biological farming, although the total number of species that was observed for the latter categories was larger.

These results emphasize that species richness depends on – among other factors - sample size. Therefore, a value of species richness without indication of sample size is meaningless. You need to provide both sample size and species richness. For example, you could state that 21 species were recorded for nature management for a sample size of 6 sites of 4 m^2. A shorter way of communicating this information would be: "21 species were recorded for nature management *(n=6)*". The same goes for the entire survey. You should indicate that the total species richness of 30 was recorded for 20 sites. You should definitely indicate your sample size when you describe your methodology.

Note that we did not make any comparisons

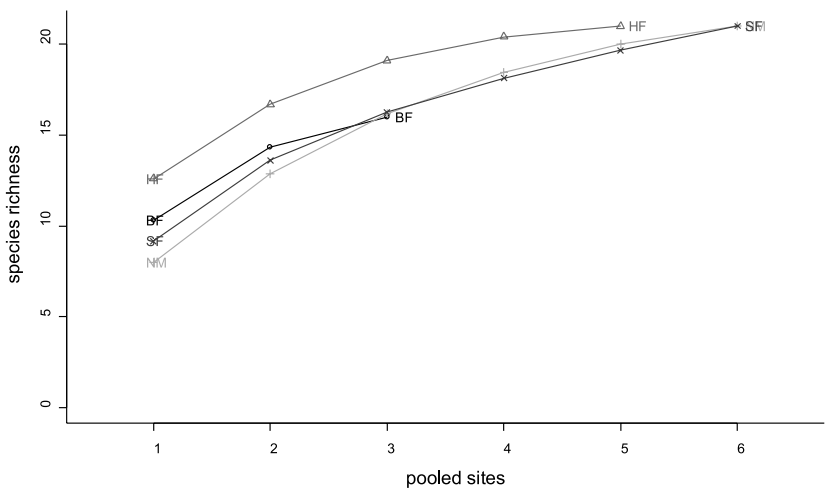


Figure 4.3 Species accumulation curves for various subsets of the dune meadow dataset based on management.

of species richness at the scale of the entire landscape, but only for areas of equal sample size. It is possible that patterns are different in the entire landscape, since we only have data from a fraction of the total landscape. For example, it could be possible that more species occur in the entire landscape in standard farming if 95% of the total area is under this type of management. Although species accumulate more quickly in hobby farming (Figure 4.3), a **smaller total area** of farmland under hobby farming could potentially result in smaller **total richness** for hobby farming for the entire landscape. It is difficult to compare different subsets at the level of the total landscape since it is extremely difficult to extrapolate species richness beyond the area that was sampled (see below: estimating the total number of species of the survey area). You could mention important differences in coverage of different subsets of data that could potentially influence patterns at the scale of the entire landscapes that you are studying.

It is also important to realize that similarity or difference in species richness does not mean that the identities of the species are the same: in the example above, it could be that standard farming has a subset of the species that occur in hobby farming, or it could be that most species under standard farming are different from those that occur in hobby farming. The species accumulation curves do not provide information on overlap in species composition. If the objectives of your study require understanding differences in species composition of the various categories of landuse then look at the methods of chapter 8 and beyond.

Species accumulation curves based on the number of plants surveyed

The species accumulation curves that were calculated earlier gave values for combinations of 1, 2 … ,all sites. This is only one method by which species accumulation curves can be calculated.

An alternative method of calculating a species accumulation curve is not to calculate the average number of species for 1, 2, …, all **sites**, but for 1, 2, …, all **sampled plants** (individuals). If you check Figure 4.1 again, you could notice that some sites have 3 trees whereas other sites have 6 trees. As all the sites have the same number of species, you can deduct that some sites have fewer species per tree.

For example, assume that the values of the species matrix of the dune dataset indicate the number of plants observed for every site and for each species. The total number of plants is now 685. A species accumulation curve can now be calculated by selecting the first plant at random from the 685, the second plant at random from the remaining 684 plants, …, until all plants were selected. An alternative species accumulation curve will be obtained.

This approach has been described as **individual-based species accumulation**. The approach where sites are combined can be referred to as **sample-based species accumulation**.

For some datasets, information is not collected from sites, but from observed individuals. Often surveys on vertebrates are obtained in this manner, recording the species identity for the first observed animal, then the identity for the second observed animal and so on. In such situations, the only accumulation curve that can be calculated is an individual-based species accumulation. The individual-based species accumulation results for the dune dataset are shown on the next page.

```
Indiv    34.000000 68.000000 103.000000 137.000000 171.000000 206.000000 240.000000
Richness 17.111627 22.239417  24.702790  26.049452  26.907559  27.520440  27.963881
sd        1.797432  1.733433   1.535193   1.374937   1.253959   1.157158   1.078629

Indiv    274.000000 308.0000000 342.0000000 377.0000000 411.0000000 445.000000
Richness  28.313305  28.5986770  28.8374254  29.0458951  29.2195268  29.369823
sd         1.008308   0.9423203   0.8785947   0.8141377   0.7518479   0.689217

Indiv    480.0000000 514.0000000 548.0000000 582.0000000 616.0000000 651.0000000 685
Richness  29.5041434  29.6177792  29.7170766  29.8037462  29.8790427  29.9455652  30
sd         0.6236441   0.5579764   0.4891322   0.4151685   0.3324024   0.2278851   0
```

Figure 4.4 shows the results. You could have noticed that the results do not indicate all 685 possible combinations of individuals, although the species richness can be calculated for any number of individuals. The results were only reported for multiples of the average number of individuals of a site. In the case of the dune meadow dataset, the average number of individuals per site is 685/20 = 34.25. The number of accumulated individuals **closest** to multiples of the average number of individuals are listed. These numbers are the 34, 68, 103, …, 651, 685 printed above for `Indiv`. An integer number is chosen as the individual-based species accumulation curve can not be calculated for fractional numbers of combined individuals.

Since 34.25 individuals occur on a site on average, the results of an individual-based species accumulation curve can also be expressed at the scale of the expected number of combined sites rather than the number of combined individuals. The scale can easily be changed by dividing the number of combined individuals by the average number of individuals per site. For the dune meadow dataset, these numbers of sites equal

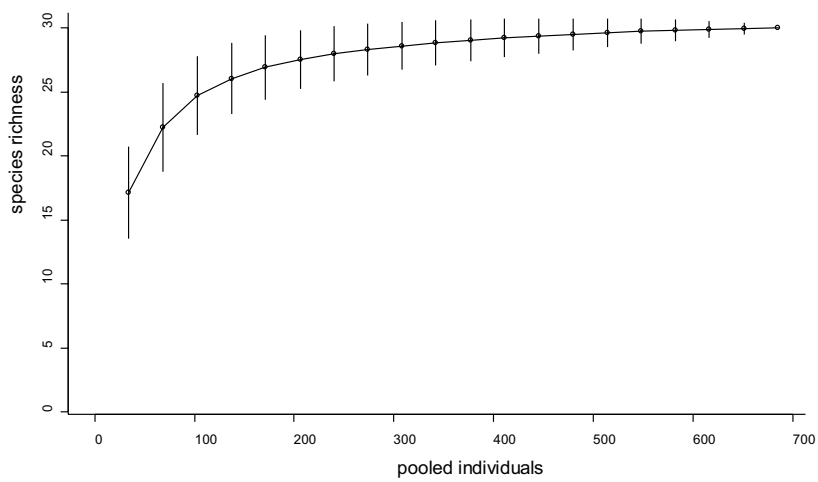


Figure 4.4 Species accumulation curve for the dune meadow dataset, based on individual-based species accumulation.

0.9927007 (= 34/34.25), 1.985401 (= 68/34.25), 3.007299 (= 103/34.25), …, 19.0072993 (= 651/34.25) and 20 (= 685/34.25). Since the average number of individuals has decimals, it is not possible to calculate the individual-based species accumulation for exact (integer) numbers of accumulated sites.

One of the reasons for only reporting individual-based species accumulation curves at the scale of the expected number of sites is that results become very lengthy for large numbers of individuals otherwise (in our example reporting the average number of species for 1, 2, 3, 4, …, 683, 684, 685 individuals – many surveys record substantially larger numbers of individuals and would require lengthier outputs). The most important reason for reporting only for multiples of the average number of individuals is that the sample-based and individual-based species accumulation curves can then be plotted on the same graph. Figure 4.5 gives an example for the dune meadow dataset. The reason that the curves can be plotted together is that we obtain multiples of the average number of individuals when we combine sites at random in sample-based species accumulation curves, so that both curves can be plotted at the same scale (Gotelli and Colwell 2001).

Since we expect multiples of the average number of individuals per site, we can also choose to scale the horizontal axis by the average number of individuals. Figure 4.6 gives an example. Note that the shapes of the curves of Figures 4.5 and 4.6 are the same, as it is only the scale of the horizontal axis that is different.

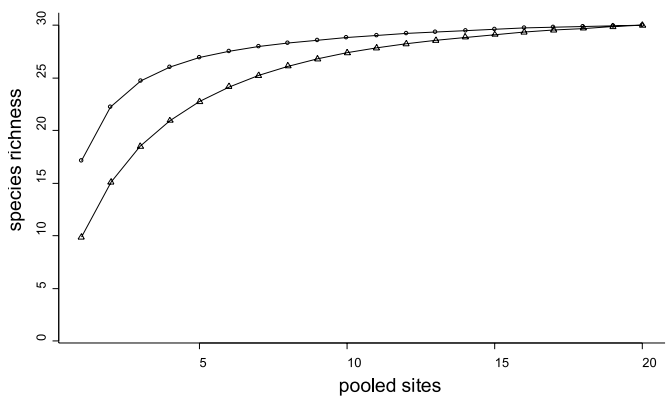


Figure 4.5 Plot of sample-based (Δ) and individual-based (O) species accumulation for the dune meadow dataset.

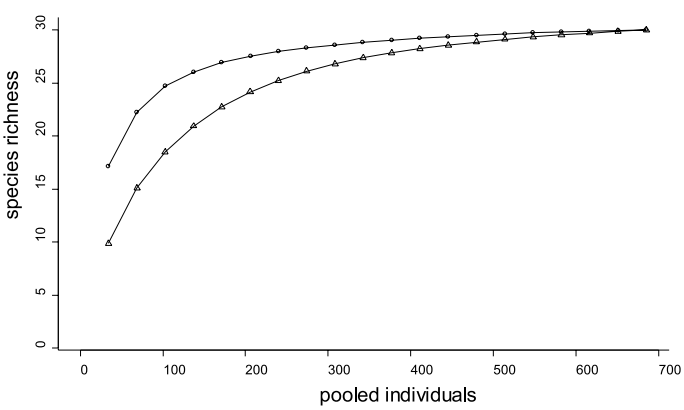


Figure 4.6 Plot of sample-based (Δ) and individual-based (O) species accumulation for the dune meadow dataset. The horizontal axis is scaled by the average number of combined individuals.

The difference between the individual-based and sample-based species accumulation (as can be seen in figures 4.5 and 4.6) is a result of the fact that species are not distributed randomly over the sites. Remember that individual-based species accumulation is based on randomly selecting individuals from the entire set of individuals not taking site information into account. If both curves have similar patterns, you can conclude that species are distributed at random over the sites.

The difference is especially big for the first site of the species accumulation curves. Figure 4.6 shows that you find more species when taking 34 individuals at random (average species richness = 17.1) than found on one site selected at random (average species richness = 9.8, average number of individuals = 34.25). This means that a single site is more likely to contain individuals of the same species than expected from the frequencies of species in the whole survey. These results thus show that plants of the same species tend to be clustered.

Changing the scale of the horizontal axis for species accumulation curves

We can change the scale of the horizontal axis, by multiplying or dividing the number of sites by a specific number. We can make this choice of scaling for the horizontal axis both for the sample-based and individual-based species accumulation curve. The shapes of the curves will remain the same.

However, when comparing different subsets in the data, then different patterns will be obtained by changing between number of sites and the number of individuals on the horizontal axis. Figure 4.7 gives an example. The subsets based on management of the dune meadow datasets were used again, and a sample-based species accumulation pattern was calculated, but in contrast with Figure 4.3, the scale of the horizontal axis was now based on the number of individuals. Because the total number of individuals is different, the relative positions of the curves will change.

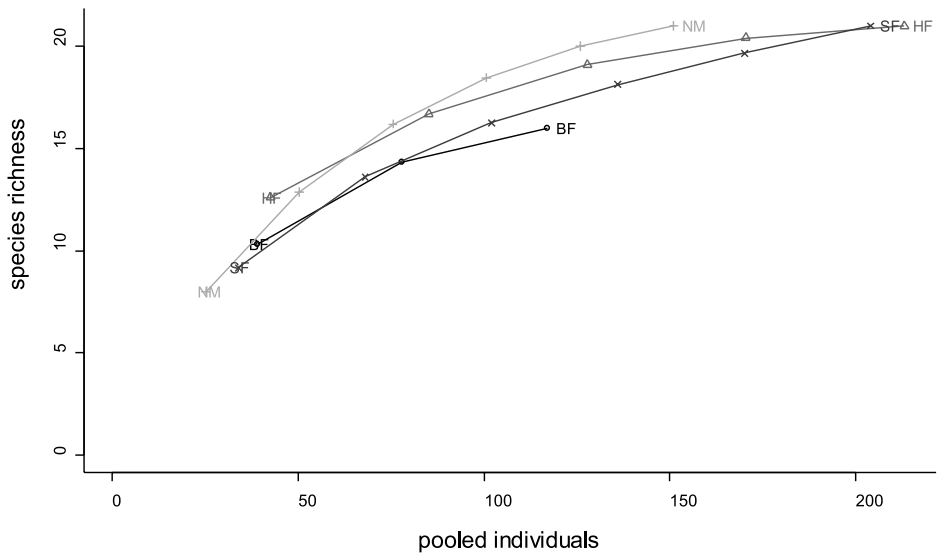


Figure 4.7 Sample-based species accumulation curves for various subsets of the dune meadow dataset based on management. The horizontal axis is scaled by the average number of combined individuals.

You can see now that hobby farming does not have a curve above the other curves as was seen in Figure 4.3. The reason for this pattern is that hobby farming had the largest number of individuals of the 4 categories. In this situation, we can conclude that the pattern observed for sample-based species accumulation curves (Figure 4.3) are mainly a result from differences in density between management systems. It will often be a good idea to compare sample-based with individual-based species accumulation to get a feel of the influence of density on sample-based results. And of course it is also important to summarize what these differences in plant density are. They may turn out to be a key aspect of the differences.

You can choose for other methods of scaling the horizontal axis. You could for instance scale the horizontal axis by the average size of the site. This would result in a different pattern in case sizes of sites are not constant. Since different patterns can be obtained, it is important that you started your study with clear objectives on how the results will be compared.

To summarize, you have two options to construct species accumulation curves. The first option is to either accumulate sites or to accumulate individuals. The second option concerns how you scale the horizontal axis of the species accumulation curve. The horizontal axis needs to be scaled by an average for all the sites. This average could be the average number of sites per site (=1 site/site), the average number of individuals per site, the average area per site, or you may think of others!

The important message here is that since species richness depends on sample size, and since sample size can be measured by different methods (number of sites, size of sample area, number of individuals, …), then different results can be obtained for different sample sizes. You should always mention the method that you used to determine sample size, and if possible compare results for different sample sizes. You should make sure that the method that you use meets your initial objectives for conducting the survey, since calculation of species accumulation curves should never be an end in itself.

Estimating the total number of species of the survey area

In most situations, we record the number of species for a number of sites. Normally we will not cover the entire area that we are interested in. We will not know the species composition of the area that we did not sample. However, since species richness depends on sample size, we can expect that we will not have recorded all the species that occur in the survey area.

Some formulae were developed to estimate the total number of species of a survey area (e.g. Longino et al. 2002). In terms of species accumulation curves, we need to extrapolate the curve until the point on the x axis corresponding to the whole study area. The estimation methods are based on different mathematical methods for extrapolation. These formulae depend on the fact that the sites were selected at random from the entire survey area. Different formulae make different assumptions of how species accumulate beyond the sampled sites. Probably one method will be more precise in one situation, and another method will be better for a second situation. Since you have no data to check for the accuracy of an extrapolation, you can not determine the best extrapolation method for your data. Any results from these methods have to be treated with caution.

If you want to extrapolate, then the best approach is to report the range in values obtained with the different methods and expect that the correct total richness lays somewhere within that range. For the dune meadow dataset, some predictions of total species richness (using the first- and second-order Jackknife, Chao and bootstrap formulae) are:

```
    Jack.1
all  32.85
    Jack.2
all  33.84
     Chao
all 32.25
     Boot
all 31.54
```

You can see that these estimates are quite similar, giving a range of 31-34 species. This is not surprising given the shape of the species accumulation curves, which flatten off for large numbers of sites, suggesting that the sampling has captured nearly all the species in the study area. In other situations, the estimates will differ more strongly. This will be the case if the species accumulation curves are still climbing at the right hand end. For such curves there are many ways they can be extrapolated, hence a wide range of answers. The specialist software *EstimateS* (http://viceroy.eeb.uconn.edu/EstimateS) was developed especially to calculate estimations of total diversity.

Calculating genus richness or family richness

You can calculate accumulation curves for the number of genera, number of families, or any other taxonomic level for which you have appropriate data. You could also calculate accumulation curves at intraspecific levels, such as for specific alleles. To be able to calculate those curves, you need to modify the species matrix so that the columns will refer to genera or families. This will be easy with a good database where you keep your data (see Chapter 2). The rest of the procedure is the same as for the calculation of species accumulation curves.

Some other methods of species accumulation

The methods described earlier were based on random accumulation of sample plots. Some other methods use the spatial configuration of sample plots.

One spatial method constructs a species accumulation curve based on a complete inventory of an area. This means that there are no unsampled areas within the area. This method has been used for 50-ha forest surveys, for instance (Plotkin et al. 2000). The total area is then subdivided, and the average species richness for half the area is estimated. The average species richness is then calculated for further subdivisions. For example, the average species richness is calculated for the two 25-ha halves, for the four 12.5-ha quarters and for the eight 6.25-ha subdivisions.

Another spatial method is also based on the species richness for sample plots of different sizes. Randomly in an area, sample plots of different size classes are placed. Afterwards, the average species richness for each size class is calculated.

Yet another method is to select the first sample plot at random, then add the second sample plot that is nearest, then the third that is second nearest and so on.

You can use these methods when you are specifically interested in the influence of spatial proximity of sites on the species accumulation curves. You could compare the curve that uses the spatial configuration with the curve that is based on random combination of sites to get an idea of the influence of spatial proximity. The approach is similar to the comparison of the sample-based and individual-based species accumulation curves. As always, you should only conduct this type of analysis if you have a clear objective for doing so.

References

Brose U, Martinez ND and Williams RJ. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology* 84: 2364-2377.

Fonseca CR and Ganade G. 2001. Species functional redundancy, random extinctions and the stability of ecosystems. *Journal of Ecology* 89: 118-125.

Gotelli NJ and Colwell RK. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4: 379-391 (recommended as first priority for reading)

Griffiths D. 1999. On investigating local-regional species richness relationships. *Journal of Animal Ecology* 68: 1051-1055.

Harte J, Kinzig A and Green J. 1999. Self-similarity in the distribution and abundance of species. *Science* 284: 334-336.

Hayek L-AC and Buzas MA. 1997. *Surveying natural populations.* New York: Columbia University Press.

He F and Legendre P. 2002. Species diversity patterns derived from species-area models. *Ecology* 83: 1185-1198.

Hurlbert SH. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52: 577-586.

Kindt R. 2002. *Methodology for tree species diversification planning for African agroecosystems.* Ph. D. thesis. Ghent: University of Ghent, Belgium.

Kindt R, Van Damme P and Simons AJ. in press. Patterns of species richness at varying scales in western Kenya: planning for agroecosystem diversification. Biodiversity and Conservation.

Krebs CJ. 1994. *Ecological methodology*. Second edition. Menlo Park: Benjamin Kummings.

Legendre P and Legendre L. 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV.

Lomolino MV. 2000. Ecology's most general, yet protean pattern: the species-area relationship. *Journal of Biogeography* 27: 17-26.

Longino JT, Coddington JA and Colwell RK. 2002. The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology* 83: 689-702.

Loreau M. 2000. Are communities saturated? On the relationship between α, ß and γ diversity. *Ecology Letters* 3: 73-76.

Magurran AE. 1988. *Ecological diversity and its measurement.* Princeton, N.J.: Princeton University Press.

Moreno CE and Halffter G. 2001. On the measure of sampling effort used in species accumulation curves. *Journal of Applied Ecology* 38: 487-490.

Pielou EC. 1969. *An introduction to mathematical ecology.* New York: Wiley - Interscience.

Plotkin JB, Potts MD, Yu DW, Bunyavejchewin S, Condit R, Foster R, Hubbell S, LaFrankie J, Manokaran N, Seng LH, Sukumar R, Nowak MA and Ashton PS. 2000. Predicting species diversity in tropical forests. *Proceedings of the National Academy of Sciences* 97 (20), 10850-10854

Rosenzweig ML. 1995. *Species diversity in space and time.* Cambridge: Cambridge University Press.

Rosenzweig ML. 1999. Heeding the warning of biodiversity's basic law. *Science* 284: 276-277.

Rosenzweig ML, Turner WR, Cox JG and Ricketts TH. 2003. Estimating diversity in unsampled habitats of a biogeographical province. *Conservation Biology* 17: 964-874.

Ugland K I, Gray JS and Ellingsen KE. 2003. The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology* 72: 888-897.

Whittaker RJ, Willis KJ and Field R. 2001. Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography* 28: 453-470.

Doing the analyses with the menu options of Biodiversity.R

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 calculate the total number of species:

Biodiversity > Analysis of diversity > Diversity indices…
→ Diversity index: richness
→ Calculation method: all

To calculate the total species richness for separate sites:

Biodiversity > Analysis of diversity > Diversity indices…
→ Diversity index: richness
→ Calculation method: separate per site

To compare the total number of species for various subsets of data:

Biodiversity > Analysis of diversity > Diversity indices…
→ Diversity index: richness
→ Calculation method: all
→ Subset options: Management
→ Subset: .

To calculate a sample-based species accumulation curve

Biodiversity > Analysis of diversity > Species accumulation curves…
→ Accumulation method: exact
→ Accumulation method: random (alternative to the exact method)

To calculate a sample-based species accumulation curve scaled by the number of accumulated individual plants:

Biodiversity > Analysis of diversity > Diversity indices…
→ Diversity index: abundance
→ Calculation method: separate per site
→ Output options: save results
Biodiversity > Analysis of diversity > Species accumulation curves…
→ Accumulation method: exact
→ Scale of x axis: abundance

To calculate an individual-based species accumulation curve scaled by the number of accumulated individual plants:

Biodiversity > Analysis of diversity > Diversity indices…
→ Diversity index: abundance
→ Calculation method: separate per site
→ Output options: save results
Biodiversity > Analysis of diversity > Species accumulation curves…
→ Accumulation method: rarefaction
→ Scale of x axis: abundance

To compare species richness between various subsets in the data using species accumulation curves

Biodiversity > Analysis of diversity > Species accumulation curves…
→ Accumulation method: exact
→ Subset options: Management
→ Subset: .

Calculating the expected species richness for the entire survey area:

Biodiversity > Analysis of diversity > Diversity indices…
→ Diversity index: Jack.1
→ Calculation method: all

Doing the analyses with the command options of Biodiversity.R

To calculate the total number of species:

```
Diversity.1 <- diversityresult(dune, index='richness')
Diversity.1
```

To calculate the total species richness for separate sites:

```
Diversity.2 <-  diversityresult(dune, index='richness',
  method='s')
Diversity.2
summary(Diversity.2)
Diversity.3 <- diversityresult(dune[1:2,], index='richness')
Diversity.3
```

To compare the total number of species for various subsets of data:

```
Diversity.4 <- diversitycomp(dune, y=dune.env,
  factor1='Management', index='richness' ,method='all')
Diversity.4
```

To calculate a sample-based species accumulation curve

```
Accum.1 <- accumresult(dune, method='exact')
Accum.1
accumplot(Accum.1)
Accum.2 <- accumresult(dune, method='random',
  permutations=1000)
Accum.2
accumplot(Accum.2)
```

To calculate a sample-based species accumulation curve scaled by the number of accumulated individual plants

```
dune.env$site.totals <- apply(dune,1,sum)
Accum.3 <- accumresult(dune, y=dune.env, scale='site.totals',
  method='exact')
Accum.3
accumplot(Accum.3, xlab='pooled individuals')
```

To calculate an individual-based species accumulation curve (first scaled by the number of sites in Accum.4, then by the number of accumulated plants in Accum.5)

```
Accum.4 <- accumresult(dune, method='rarefaction')
Accum.4
accumplot(Accum.4)
dune.env$site.totals <- apply(dune,1,sum)
Accum.5 <- Accum.5 <- accumresult(dune, y=dune.env,
  scale='site.totals', method='rarefaction')
Accum.5
accumplot(Accum.5, xlab='pooled individuals')
```

To compare species richness between various subsets in the data using species accumulation curves

```
Accum.6 <- accumcomp(dune, y=dune.env, factor='Management',
method='exact')
Accum.6
dune.env$site.totals <- apply(dune,1,sum)
Accum.7 <- accumcomp(dune, y=dune.env, factor='Management',
  scale='site.totals', method='exact', xlab='pooled
  individuals')
Accum.7
```

To calculate a collector's curve

```
Accum.8 <- accumresult(dune, method='collector')
Accum.8
accumplot(Accum.8)
```

Calculating the expected species richness for the entire survey area

```
Diversity.5 <- diversityresult(dune, index='Jack.1')
Diversity.5
Diversity.6 <- diversityresult(dune, index='Jack.2')
Diversity.6
Diversity.7 <- diversityresult(dune, index='Chao')
Diversity.7
Diversity.8 <- diversityresult(dune, index='Boot')
Diversity.8
```

CHAPTER 5

Analysis of diversity

Analysis of diversity

This chapter describes some methods of investigating diversity. First, the concept of diversity is introduced. Then some methods of calculating and comparing diversity are discussed.

Diversity entails richness (or the number of species) and evenness (or equality in the number of individuals for every species).

Where is diversity the highest?

In this chapter, some methods are introduced to calculate the diversity of a specific site (a sample plot in a forest, a farm, a village, …). Consider Figure 5.1, for instance. You may be interested in finding out whether the diversity is higher in site B than in site A. You could for instance have a hypothesis that the diversity in site B is greater because temperatures are higher in site B. This chapter will only describe the **methods of calculating the diversity** of a site. To test a hypothesis for the relationship between some explanatory factors of diversity and diversity of a site, you will need to use a regression method as described in chapter 6 and two sites will definitely not be a large enough sample size to investigate such hypothesis.

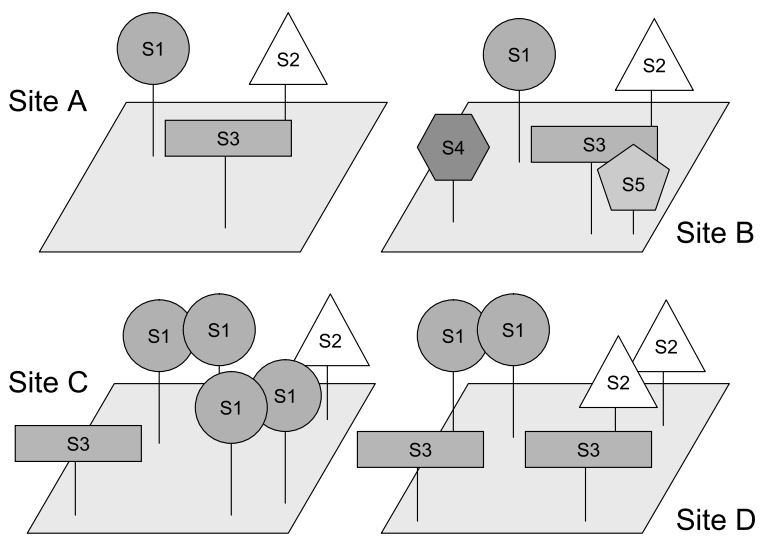


Figure 5.1 (a) Various sites differ in the number of species and in the number of trees of each species.

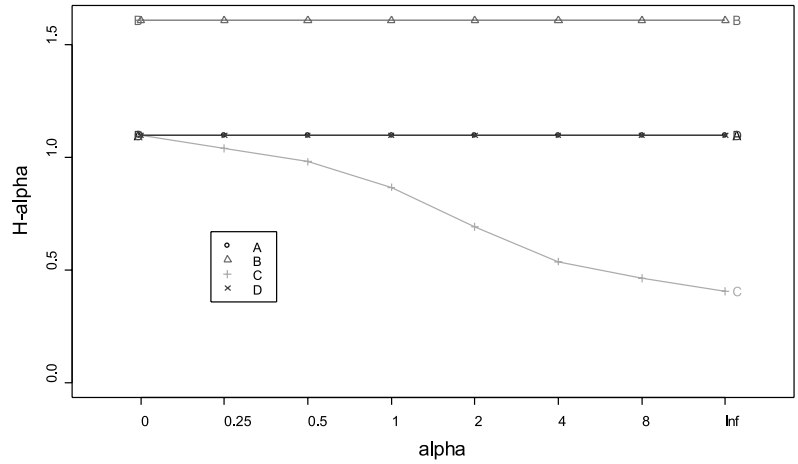


Figure 5.1 (b) Sites that are more diverse have a profile (the line in the diagram) that is higher, therefore the diversity ordering is: B > A = D > C.

What is diversity?

In general, diversity refers to the number of categories that can be differentiated, and to the proportions (or relative abundances) of the number of objects in each category. When we study tree species diversity, the categories refer to different species, whereas the objects are the trees that are counted.

Imagine that you have 2 sites: site A has 3 tree species, whereas site B has 5 tree species. In this situation, site B has the largest **species richness**. This situation is depicted in Figure 5.1.

Imagine another situation where both site C and site D contain 3 species. However, site C is dominated by one species that has 4 trees out of the total number of 6 trees on the entire site (or a proportion of 4/6). The other two species have proportions of 1/6. In site D, each species has the same number of trees (or proportions of 2/6). In this situation, site D has the largest **evenness**, which means that the proportions of the individual species are more similar. In this situation, the proportions are actually all the same for site D, so evenness is maximum for this site. This situation is also shown in Figure 5.1.

Sites A and D (Figure 5.1) have the same proportions. Since both sites have the same proportions and the same number of species, they have the same diversity. Diversity does not depend on density or total abundance.

Sites of maximum evenness will have proportions of 1/*S* for each species, where *S* is the number of species (the species richness). For example, a plot with 5 species of maximum evenness will have proportions of 1/5 for each species, whereas a farm with 10 species of maximum evenness will have proportions of 1/10 for each species. If 100 trees were recorded in total, then 5 species will be most evenly distributed when each species has 100/5 = 20 trees.

On the other hand, a site of minimum evenness will have only 1 tree for the *S*-1 less frequent species and *Tot* - (*S*-1) trees for the dominant category, if *Tot* indicates the total number of trees. If a site contains 6 trees and 3 species, the minimum evenness will be where 2 species contain 1 tree and the remaining species contains 6-2=4 trees. If a site contains 100 trees and 5 species, then with minimum evenness the dominant species will contain 100-4=96 trees.

In most situations, the evenness will be in between the maximum and minimum evenness.

Since diversity refers to richness and evenness, both these facets need to be considered when comparing diversity. If evenness is the same for the sites (sites, farms, sample plots) that you are comparing, then differences in richness will correspond to differences in diversity. If the richness is the same, then differences in evenness will correspond to differences in diversity. There will be situations, however, where one site has larger richness but lower evenness than another site. In these situations, it is not always possible to rank one site as higher in diversity than the other site.

Rank-abundance curves

Rank-abundance curves are conceptually the easiest method of analysing patterns of diversity.

First, the total number of individuals is calculated for each species. Second, species are ranked from the most abundant to the least abundant. Finally a plot is constructed with the rank number on the horizontal axis, and the abundance on the vertical axis.

For example, the rank-abundance pattern for the dune meadow dataset (we treated the values in the cells of the species matrix as if they were counts of individuals) is:

	rank	abundance	proportion
Poatri	1	63	9.2
Lolper	2	58	8.5
Leoaut	3	54	7.9
Brarut	4	49	7.2
Agrsto	5	48	7.0
Poapra	6	48	7.0
Trirep	7	47	6.9
Alogen	8	36	5.3
Elyrep	9	26	3.8
Plalan	10	26	3.8
Elepal	11	25	3.6
Antodo	12	21	3.1
Sagpro	13	20	2.9
Junart	14	18	2.6
Rumace	15	18	2.6
Achmil	16	16	2.3
Brohor	17	15	2.2
Ranfla	18	14	2.0
Belper	19	13	1.9
Junbuf	20	13	1.9
Salrep	21	11	1.6
Calcus	22	10	1.5
Hyprad	23	9	1.3
Tripra	24	9	1.3
Airpra	25	5	0.7
Potpal	26	4	0.6
Viclat	27	4	0.6
Cirarv	28	2	0.3
Empnig	29	2	0.3
Chealb	30	1	0.1

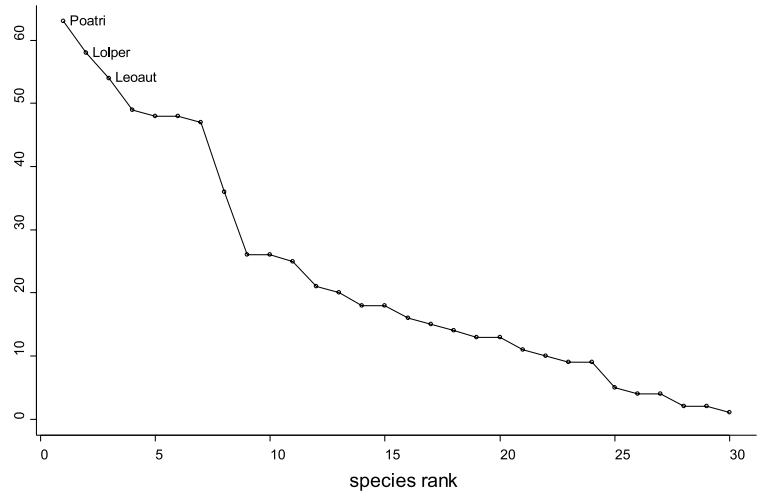


Figure 5.2 Rank-abundance curve for the dune meadow dataset.

The results given before are normally provided as a rank-abundance curve such as Figure 5.2.

You can see that *Poa trivialis* was ranked 1 as this species had the largest total abundance of 63, and that *Chenopodium album* was ranked 30 since this species had the lowest total abundance of 1. Figure 5.2 shows the rank-abundance curve for this dataset. You could create an alternative rank-abundance curve by plotting the proportion instead of the abundance on the vertical axis. The shape of the curve would remain the same, since only the scaling of the vertical axis would be different (note also that the proportion was given as percentage by multiplying all proportions with 100%). Other alternatives include plotting the logarithm of abundance on the vertical axis – this could produce better graphs when a few species are highly dominant.

The interpretation of a rank-abundance curve in terms of diversity, i.e. richness and evenness, is as follows. On the horizontal axis, species richness is provided by the width of the curve. A wider curve will indicate higher species richness. The shape of the rank-abundance curve is an indication of the evenness. A completely horizontal curve is an indication of a completely evenly distributed system. The steeper the curve, the less evenly species are distributed.

Figure 5.3 provides the rank-abundance curves for the four sites shown in Figure 5.1. The proportion is plotted on the vertical axis. Note that sites A and D have the same rank-abundance curve when scaled to proportion (each species has proportion = 1/3). You can see from the widths of the rank-abundance curves of Figure 5.3 that one site has species richness of 5, whereas the other sites have richness 3. You can also see that three sites have completely horizontal profiles or completely evenly distributed species. You can notice one site with a declining profile, indicating that some species have higher abundance than others. In other words, species are not evenly distributed for this last site. Based on this information, you could classify site B as the most diverse (highest richness [=widest] and evenness [=most horizontal]), and site C as the least diverse (lowest richness [=narrowest] and evenness [=least horizontal]).

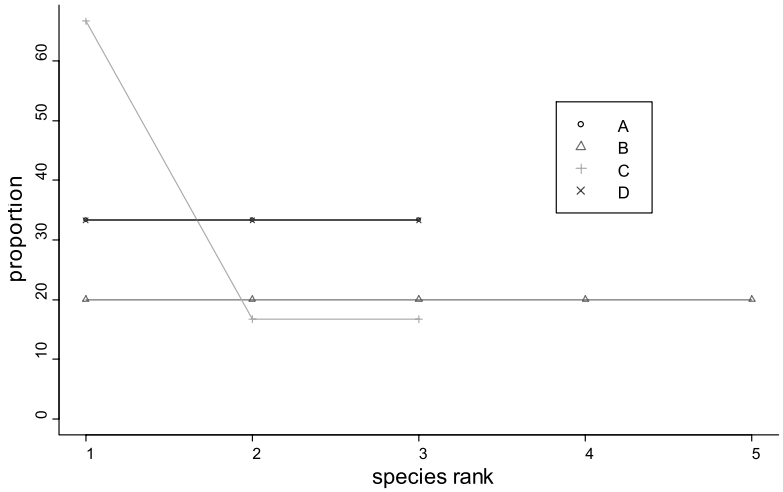


Figure 5.3 Rank-abundance curves for the 4 sites of Figure 5.1. Abundance is proportional abundance (percentage of each species of total abundance).

Models for rank-abundance curves

Various studies have been conducted to model specific rank-abundance distributions. By fitting a model, the shape of a particular rank-abundance distribution may be summarized by a few parameters. As some models are derived from theoretical assumptions about the ways in which species could coexist, the observation of a rank-abundance distribution that conforms to a particular model provides some evidence that the conditions that generate the model could apply to a particular survey. This could be important information since the question why species differ in abundance has been an important topic of biodiversity research.

A thorough discussion of rank-abundance distribution models is beyond the scope of this manual. The interested reader could consult a specialized text such as Hubbell (2001). Models should not be fitted because it is possible, but because the information that they provide is useful. If you only want to summarize the rank-abundance distribution, you should reconsider whether you would not provide better information by just providing the actual rank-abundance curve.

Different models have been formulated to describe rank-abundance distributions, including the log-normal, log series and geometric distributions. Fitting these distributions to data is not difficult but it is often difficult to choose the one model that provides the best fit to the data.

When you attempt to fit several models to the dune meadow dataset, you will obtain following results:

```
RAD models, family poisson
No. of species 30, total abundance 685

Warning: NAs introduced by coercion
              par1    par2    par3 Deviance     AIC     BIC
Preemption 0.09674                   16.852 155.570 156.972
Lognormal        3       1           38.217 178.936 181.738
Veiled.LN        3       1       1   38.217 180.936 185.140
Zipf       0.14817      -1          106.934 247.653 250.455
Mandelbrot     Inf -463703 4621740   15.204 157.923 162.127
```

It is easy to compare the fits graphically in this example as shown in Figure 5.4. Figure 5.4 shows the fit of the various models. Visually comparing the difference in the actual values and the predicted values allows you to choose a model that might fit your purpose. Of those in Figure 5.4, none capture the curvature of the observed distribution at high and particularly the low abundance end of the distribution. Over much of the range the veiled lognormal curve seems to fit best. Note that a logarithmic scale was used on the vertical axis.

An additional method of choosing the best distribution that you could use together with the graphical evaluation is to choose the model with the lowest AIC or BIC (Akaike's Information Criterion or the Bayesian Information Criterion; these are statistics that indicate goodness-of-fit of a model – lower values indicate better fits). For the dune meadow dataset, we might thus prefer the pre-emption model. But note that the statistics such as deviance, AIC and BIC only measure some aspects of fit, and they might not be the aspects you are most interested in.

Rényi diversity profiles

Rényi diversity profiles are curves that also provide information on richness and evenness, as rank-abundance curves do. Rényi diversity profiles have the advantage over rank-abundance curves that ordering from lowest to highest diversity is easier. For this reason, a Rényi diversity profile is one of several **diversity ordering techniques** (Tóthmérész 1995). The disadvantage of these curves is that information on the proportions of each species is not provided any longer.

Figure 5.1b provides the Rényi profiles for the same sites shown in Figure 5.3. The interpretation of a profile is as follows. The shape of the profile is an indication of the evenness. A horizontal profile indicates that all species have the same evenness – the same situation as for rank-abundance curves. The less horizontal a profile is, the less evenly species are distributed. In Figure 5.1b, we see that 3 sites have horizontal profiles, which means that species are completely evenly distributed for these

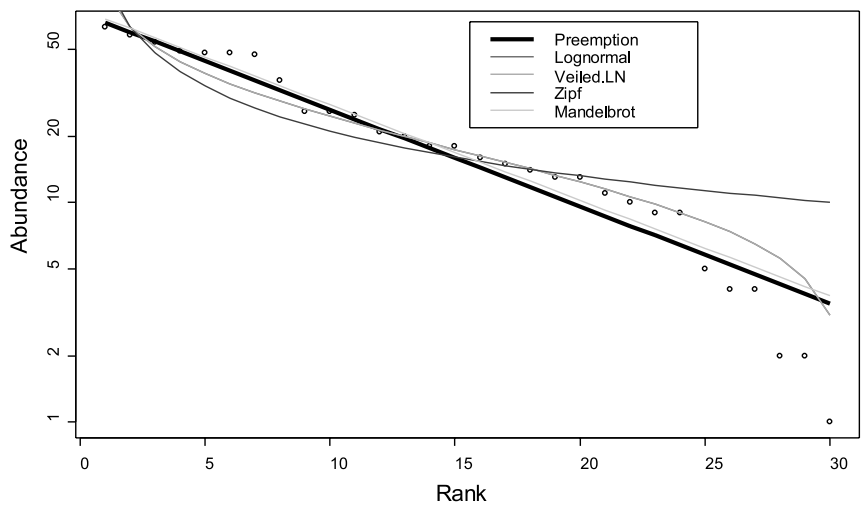


Figure 5.4 Fits of various models to the rank-abundance distribution of the dune meadow dataset.

sites. Site C has a profile that declines from left to right. This indicates that species are not evenly distributed for this site. The starting position at the left-hand side of the profile is an indication of the species richness. Profiles that start at a higher level have higher richness.

The major advantage of Rényi diversity profiles is that sites can easily be ordered from high to low diversity. If the profile for one site is everywhere above the profile for another site, then this means that the site with the highest profile is the more diverse of the two. From Figure 5.1b, you can for instance see that site B is the most diverse, and site C is the least diverse. If the profiles intersect, it is not possible to order the sites from lowest to highest diversity. It is possible that one site has larger species richness, but lower species evenness, although this is not a necessary condition for intersecting profiles.

If one diversity profile is higher than another, then the corresponding cumulated proportions of the rank-abundance curve will be lower. The cumulated proportions are calculated as the sum of the proportions of the rank-abundance curve for 1, 2, 3, …, all species. Figure 5.5 shows the cumulated proportions for the rank-abundance curves of Figure 5.3. The curve that is lowest everywhere in the figure corresponds to the site with the highest diversity – in this example, this is site B. You can verify that the ranking of diversity that is portrayed in Figure 5.5 (B > A = D > C) is the same as the ranking of Figure 5.1b.

Since the shape of the Rényi diversity profiles is influenced by evenness, you can compare the evenness of various sites by only looking at the shape of these curves. **Rényi evenness profiles** are a more direct method of comparing evenness (Ricotta 2003, Kindt et al. in press). These evenness profiles only reflect differences in evenness. The way that the evenness profiles are interpreted is similar to the way that diversity profiles should be interpreted, and the only difference is that a graphical comparison of evenness rather than of diversity is provided. An area of larger evenness will have an evenness profile that is everywhere above the evenness profile of an area of lower evenness. Intersecting evenness profiles means that no ranking in evenness can be provided. Figure 5.6

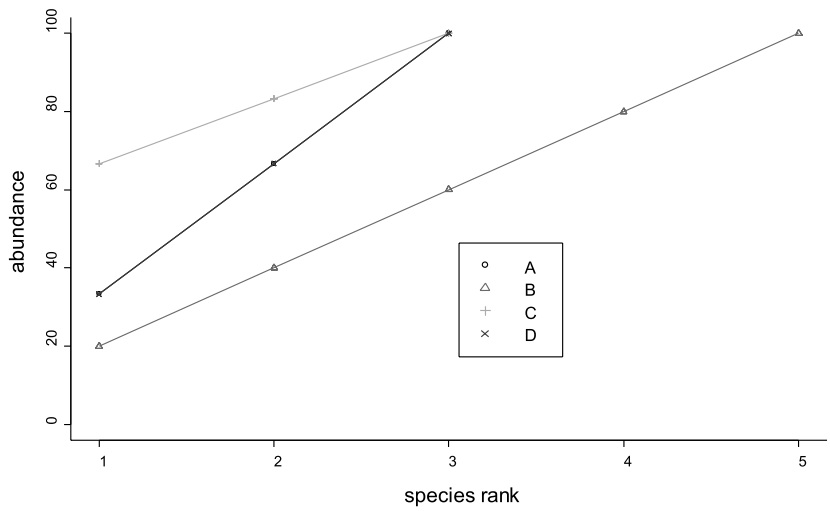


Figure 5.5 Cumulated proportions (%) for the 4 sites of Figure 5.1. This is an alternative diversity ordering technique to the Rényi diversity profile, with the lowest curve over the entire range indicating the site of highest diversity.

provides the evenness profiles for the same sites as Figure 5.1b. You can see from Figure 5.6 that three sites have the same and complete evenness (horizontal profiles), and one site has unevenly distributed species (site C).

You can calculate one Rényi diversity profile for an entire dataset, or separate profiles for each site. Figure 5.7 provides the Rényi diversity profile for the separate sites of the dune meadow dataset.

You can observe in Figure 5.7 that many profiles are intersecting. This means that many sites can not be ranked from highest to lowest diversity. Since the profile for site X1 is lowest over its entire range, this is clearly the site with the lowest diversity. There is not a site with the highest diversity of all sites. Site X5 has the largest richness, but the profile for this site intersects with some other profiles as those for X6 and X8. X5 can thus not be classified as the most diverse site.

Figure 5.8 shows the evenness profiles for the separate sites. These curves show that the evenness is the largest for site X20, and the lowest for site X13. The intersections of the profiles (for instance for X1 and X4) indicate that it is not possible to rank those sites from lowest to highest evenness.

Interpretation of some Rényi diversity profile values

Some of the values of the Rényi diversity profile provide some specific details on the corresponding site. The values used to construct Figure 5.1b are for instance:

	0	0.25	0.5	1	2	4	8	Inf
A	1.098612	1.098612	1.0986123	1.0986123	1.0986123	1.0986123	1.0986123	1.0986123
B	1.609438	1.609438	1.6094379	1.6094379	1.6094379	1.6094379	1.6094379	1.6094379
C	1.098612	1.040010	0.9808293	0.8675632	0.6931472	0.5380261	0.4633843	0.4054651
D	1.098612	1.098612	1.0986123	1.0986123	1.0986123	1.0986123	1.0986123	1.0986123

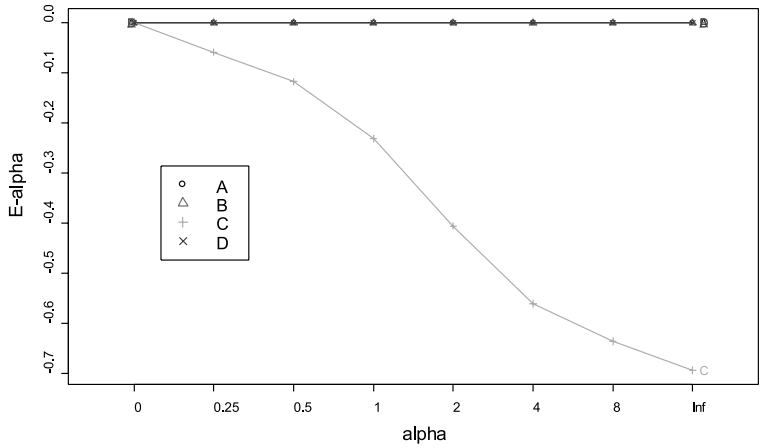


Figure 5.6 Rényi evenness profiles for the 4 sites of Figure 5.1.

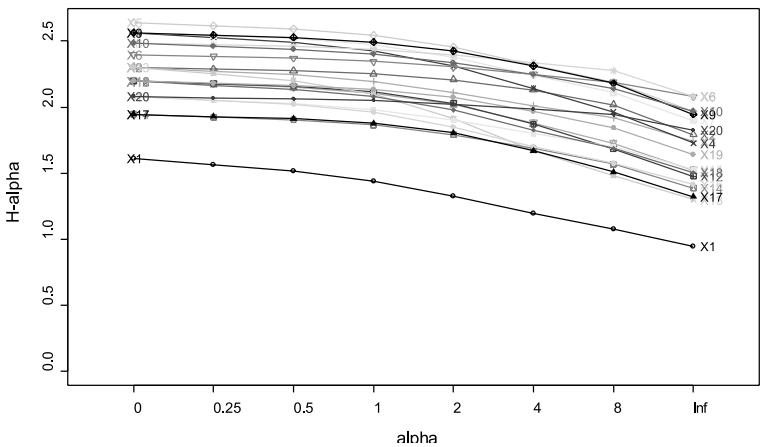


Figure 5.7 Rényi diversity profiles for the separate sites of the dune meadow dataset.

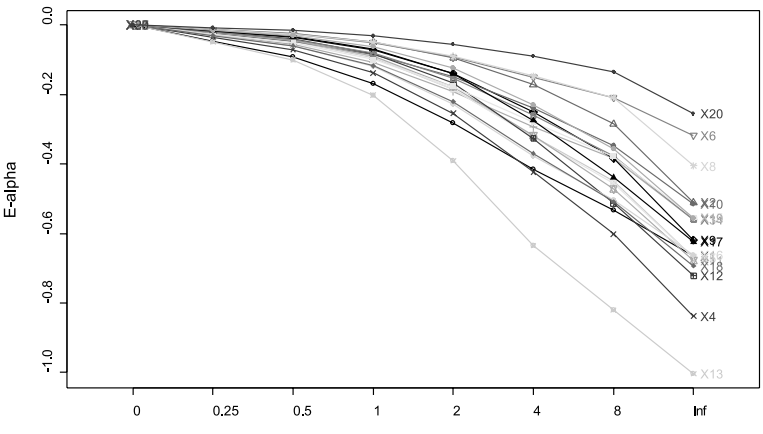


Figure 5.8 Rényi evenness profiles for the separate sites of the dune meadow dataset.

Each value of the Rényi diversity profile is based on a parameter 'alpha'. Box 5.1 shows how the profile value is calculated from the species proportions and the alpha parameter.

The profile values for alpha=0 provide information on species richness. The profile value is the logarithm of the species richness. For site A, 1.098612 = ln(3). Thus, if you take the anti-logarithm (y=exp(x)) of a Rényi diversity profile, you will obtain the species richness at alpha=0. For site B, species richness = 5 = exp(1.609438). This is the reason that profiles that start at a higher level correspond to sites that are richer.

The profile value for alpha=infinity provides information on the proportion of the most abundant species. The profile value for alpha=infinity for system C equals 0.4054651. When you take the anti-logarithm (y=exp(x)), and then take the reciprocal value, then you will obtain the proportion of the most dominant species. For system C, the proportion of the most dominant species = 4/6 = 1/exp(0.4054651). As a consequence, profiles that are higher at alpha=infinity have a lower proportion of the dominant species. A larger evenness thus corresponds with lower proportions of the dominant species.

The profile value for alpha=1 is the Shannon diversity index (discussed in the next section). The profile value for alpha=2 is the logarithm of the reciprocal Simpson diversity index (see next section).

Box 5.1 How to calculate the Rényi diversity profile?

The formula to calculate the diversity profile is:

$$H_\alpha = \frac{\ln(\sum_{i=1}^{S} p_i^\alpha)}{1-\alpha}$$

The p_i values of the above formula are the proportions of each species. For site A (Figure 1), these are $p_1 = p_2 = p_3 = 1/3$.

To calculate the profile value for α=8, these proportions are raised to power α (= 8) = $(1/3)^8$ = 0.000152 and added together for all species as indicated by the summation symbol ∑ (0.000152 + 0.000152 + 0.000152 = 0.000457). Next the logarithm is taken (ln(0.000457) = -7.690). Finally, the obtained value is divided by (1-α = -7) (-7.690/-7 = 1.0986).

A profile is calculated by changing the value of α from 0 to infinity. In Biodiversity.R, the standard values for α are: 0, 0.25, 0.5, 1, 2, 4, 8, and infinity. Although the formula can not be used directly to calculate the profile value at scales 1 and infinity, other formulas can be used to calculate these values since it is known that these values are related to the Shannon and Berger-Parker diversity indices (see main text). You can directly calculate the diversity profile by the `renyi` function (see at the end of this chapter).

Diversity indices

Diversity indices provide a summary of richness and evenness by combining these two facets of diversity into a single statistic. There are many ways by which richness and evenness can be combined, and this has resulted in many different diversity indices. Some of the common diversity indices are the Shannon, Simpson, and log series alpha diversity indices. A larger Simpson index will indicate lower diversity, hence it is better to analyse the reciprocal value of the Simpson index. An alternative approach is to report 1-Simpson index.

For the dune meadow dataset, you will obtain following results for the Shannon and 1-Simpson diversity indices (for each diversity index, the sites are sorted in increasing order) (shown on the right-hand side):

Diversity indices are a more compact method of comparing diversity. However, one diversity index will often not provide sufficient information to order sites from high to low diversity. Only diversity ordering techniques such as the Rényi diversity profiles will provide enough information that will allow you to conclude that one site is more diverse than another site. The reason for this phenomenon is actually that not all entities can be ordered from lowest to highest diversity (as shown earlier for the dune meadow dataset, see Figure 5.7).

It is true that if site A is more diverse than site B, that then the diversity index of site A will be larger. It is however not necessarily true that if the diversity index of site A is larger than the diversity index of site B, that then the diversity of site A is larger.

You could see in this dataset that the Shannon index of site X13 > X20, but that for the Simpson index X13 < X20. This is one illustration of the fact that a single diversity index may not provide sufficient information for diversity ordering. In Figure 5.7, you can see the intersection in the profiles of X13 and X20.

```
    Shannon
X1    1.440
X14   1.864
X17   1.876
X16   1.960
X15   1.979
X20   2.048
X18   2.079
X13   2.100
X11   2.106
X12   2.114
X19   2.134
X3    2.194
X2    2.253
X6    2.346
X10   2.399
X4    2.427
X8    2.435
X7    2.472
X9    2.494
X5    2.544

    inverseSimpson
X1           3.767
X14          6.000
X17          6.081
X16          6.368
X15          6.696
X13          6.764
X18          7.218
X11          7.529
X20          7.567
X12          7.609
X19          7.942
X3           8.247
X2           9.093
X6          10.017
X4          10.075
X10         10.330
X7          10.811
X8          10.959
X9          11.308
X5          11.629
```

Comparing the total diversity of different subsets of the dataset

Similarly to comparisons of species richness, you need to be cautious if you want to compare the total diversity of various subsets in your data when these subsets have different sample sizes. As for species richness, diversity indices will also change when sample size is increased. This is to be expected since diversity indices provide information on richness and evenness – so if richness changes with sample size, then diversity will change too.

If you want to compare the total diversity of subsets in your data, then you need to calculate the diversity for subsets in your data that have the same sample size. A procedure similar to the randomisation approach discussed for species accumulation curves can be used. This procedure involves taking random subsets of the data and calculating a diversity profile for the subset. By randomized resampling of the subsets, average values of the diversity profiles can be obtained. Figure 5.9 shows an accumulation pattern for the average Rényi diversity profile for the dune meadow dataset.

If we want to compare the diversity of the different management categories of the dune meadow dataset for example, then we need to compare the diversity at the same sample size for the various categories. In this dataset, the largest sample size at which this is possible is 3, the number of sites of hobby farming. The results of a comparison at sample size 3 for the dune meadow dataset is shown in Figure 5.10. From this figure, you could conclude that hobby farming is the

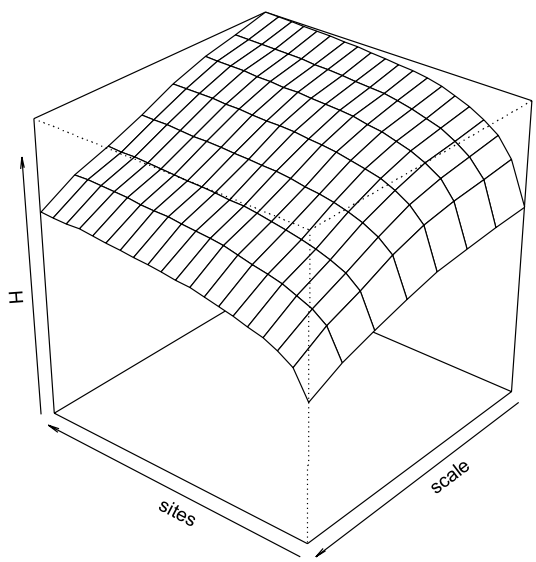


Figure 5.9 Accumulation pattern for the average Rényi diversity profiles for the dune meadow dataset.

most diverse category when comparing average diversity profiles for combinations of 3 sites.

As for species accumulation curves, there are various options to measure the same sample size. You could either choose the number of sites as a measure of sample size, or the number of plants, or the area that was sampled – and some other measures could theoretically be chosen too.

As we saw in the previous chapter, it is not necessarily so that hobby farming will be the most diverse at the scale of the entire landscape, although it is the most diverse at sample size 3. It could be possible that only 5% of the entire landscape is under hobby farming whereas standard farming could be 80% of the landscape. In such situations, standard farming could be more diverse at the scale of the entire landscape. Since we have only sampled a fraction of the landscape and since extrapolation is difficult, we have no evidence that standard farming or hobby farming is more diverse at the landscape level. When your primary interest is to understand the distribution of biodiversity, then it may also be worthwhile to investigate differences in species composition (Chapter 8 and beyond) rather than differences in total diversity of different subsets in your data.

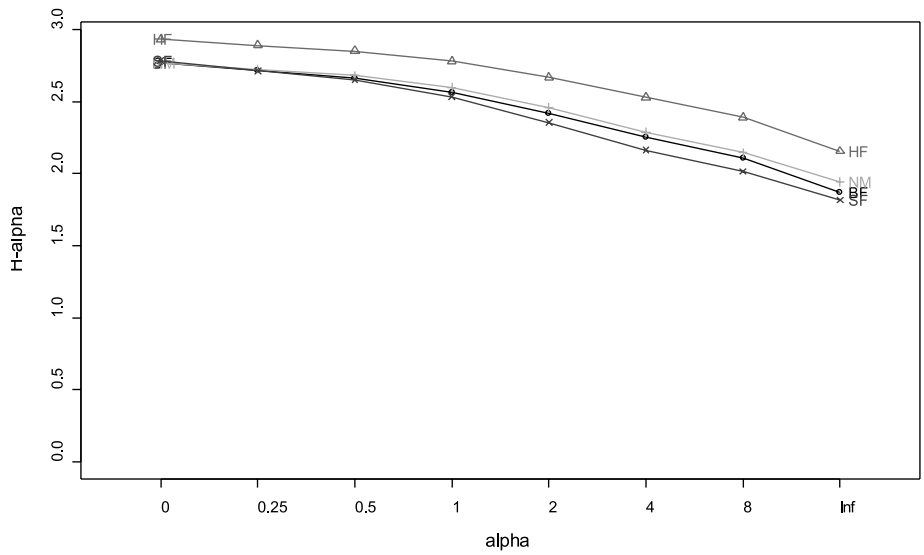


Figure 5.10 Comparison of diversity for the management categories of the dune meadow dataset. Results are based on 100 randomisations.

References

Feinsinger P. 2001. *Designing field studies for biodiversity conservation.* Washington: The Nature Conservancy.

Hayek L-AC and Buzas MA. 1997. *Surveying natural populations.* New York: Columbia University Press.

He F and Legendre P. 2002. Species diversity patterns derived from species-area models. *Ecology* 83: 1185-1198.

Hubble SP. 2001. *The unified neutral theory of biodiversity and biogeography.* Princeton: Princeton University Press.

Kempton RA. 2002. Species diversity. In: El-Shaarawi, AH and Piegrosch WW. *Encyclopedia of environmetrics.* Chichester: John Wiley and Sons.

Kent M and Coker P. 1992. *Vegetation description and analysis: a practical approach.* London: Belhaven Press.

Kindt R, Van Damme P and Simons AJ. (in press). Tree diversity in western Kenya: using profiles to characterise richness and evenness. *Biodiversity and Conservation.*

Krebs CJ. 1994. Ecological methodology. Second edition. Menlo Park: Benjamin Kummings.

Legendre P and Legendre L. 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV.

Magnussen S and Boyle TJB. 1995. Estimating sample size for inference about the Shannon-Weaver and the Simpson indices of species diversity. *Forest Ecology and Management* 78: 71-84.

Magurran AE. 1988. *Ecological diversity and its measurement.* Princeton, N.J: Princeton University Press. (recommended as first priority for reading)

Magurran AE and Henderson PA. 2003. Explaining the excess of rare species in natural species abundance distributions. *Nature* 422: 714-716.

Pielou EC. 1969. *An introduction to mathematical ecology.* New York: Wiley - Interscience.

Purvis A and Hector A. 2000. Getting the measure of biodiversity. *Nature* 405: 212-218.

Ricotta C. 2003. On parametric evenness measures. *Journal of Theoretical Biology* 222: 189-197.

Rousseau D, Van Hecke P, Nijssen D, and Bogaert J. 1999. The relationship between diversity profiles, evenness and species richness based on partial ordering. *Environmental and Ecological Statistics* 6: 211-223.

Shaw PJA. 2003. *Multivariate statistics for the environmental sciences.* London: Hodder Arnold.

Tóthmérész B. 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science* 6: 283-290.

Doing the analyses with the menu options of Biodiversity.R

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 calculate and plot a rank-abundance curve:

Biodiversity > Analysis of diversity > Rank abundance…

To model a rank-abundance curve:

Biodiversity > Analysis of diversity > Rank abundance…
→ Plot options: fit RAD

To calculate and plot a Rényi diversity profile:

Biodiversity > Analysis of diversity > Renyi profile…
→ Calculation method: all

To calculate and plot a Rényi diversity profile for each site separately:

Biodiversity > Analysis of diversity > Renyi profile…
→ Calculation method: separate per site

To calculate diversity indices for each site:

Biodiversity > Analysis of diversity > Diversity indices…
→ Diversity index: Shannon
→ Calculation method: separate per site

To compare diversity between subsets of the dataset:

Biodiversity > Analysis of diversity > Renyi profile…
→ Calculation method: separate per site
→ Subset options: Management
→ Subset: .

To calculate accumulation patterns for the Rényi diversity profile

Biodiversity > Analysis of diversity > Renyi profile…
→ Calculation method: accumulation

Doing the analyses with the command options of Biodiversity.R

To calculate and plot a rank-abundance curve:

```
RankAbun.1 <- rankabundance(dune)
RankAbun.1
rankabunplot(RankAbun.1, scale='abundance')
rankabunplot(RankAbun.1, scale='proportion')
```

To model a rank-abundance curve:

```
radfitresult(dune)
```

To calculate and plot a Rényi diversity profile:

```
Renyi.1 <- renyiresult(dune)
Renyi.1
renyiplot(Renyi.1)
renyiplot(Renyi.1, evenness=TRUE)
```

To calculate and plot a Rényi diversity profile for each site separately:

```
Renyi.2 <- renyiresult(dune, method='s')
Renyi.2
renyiplot(Renyi.2)
renyiplot(Renyi.2, evenness=TRUE)
```

To calculate diversity indices for each site:

```
Diversity.1 <- diversityresult(dune, index='Shannon'
  ,method='s')
Diversity.1
Diversity.2 <- diversityresult(dune, index='Simpson'
  ,method='s')
Diversity.2
Diversity.3 <- diversityresult(dune, index='Logalpha'
  ,method='s')
Diversity.3
```

To compare diversity between subsets of the dataset:

```
Renyi.3 <- renyicomp(dune, y=dune.env, factor='Management',
  permutations=100)
Renyi.3
```

To calculate accumulation patterns for the Rényi diversity profile

```
Renyi.4 <- renyiaccum(dune, permutations=100)
Renyi.4
renyiaccumplot(Renyi.4)
```

CHAPTER 6

Analysis of counts of trees

Analysis of counts of trees

One way by which patterns in the species matrix can be analysed is to analyse for each species separately. In this manual, we describe two methods by which species can be analysed separately: (i) analysing the counts obtained throughout the survey; and (ii) analysis of species presence-absence data. The latter method is described in the next chapter.

This section describes the analysis of species abundance as the number of individuals. The methods could also be used to analyse the total number of individuals per site, or the total number of species per site. Other measures of species abundance such as cross-sectional area or cover percentages could also be analysed.

Regression models are introduced and used in this chapter. They are a basis for much statistical analysis and there is much that could be said about them. Here we can only point in few directions.

What is a regression model?

Regression analysis is a method by which the pattern in one **response variable** is predicted from the pattern of one or several **explanatory variables**. The better the predictions explain the pattern, the better the model represents the data.

Imagine that you want to analyse the influence of elevation (the explanatory variable) on the abundance of a certain *Acacia* species (the response variable). To analyse the relationship, you measured the abundance of the species and the elevation on 11 sample plots, taking samples at regular elevation intervals of 100 m in between 500 and 1500 m. If you want to examine the relationship between abundance and elevation, then start by plotting the data and looking for a pattern. If there appears to be a linear relationship between elevation and abundance, then you could describe this relationship with a linear regression model (we will see that many types of regression models exist). The linear regression model will model the relationship between elevation and abundance as a straight line. To predict the position of the straight line, the linear regression model will estimate the parameters a and b of the following regression model: $Abundance = a + b \times elevation + deviation$. Once the parameters a and b are estimated, the straight line that predicts the expected abundance for a specific elevation can be calculated. The performance of the model can be measured by how well it describes the variation in abundance. When the differences between the measured abundance and the predicted abundance are small, then the variance explained by the model will be large. It is also possible to test the hypothesis of no relationship. The significance level (P) is sometimes used to decide whether there is evidence for a relationship or not.

Figure 6.1 gives examples of observations and predictions from a linear regression model for four species. The variance explained by the model decreases from Species 1 to Species 4, which can be observed from the larger differences between the observed abundances (circles) and the predicted abundances (straight line). The significance level for the test of no relationship between elevation and abundance also increases from Species 1 to Species 4. For Species 4, there is no evidence for a linear relationship between elevation and abundance ($P = 0.14$)

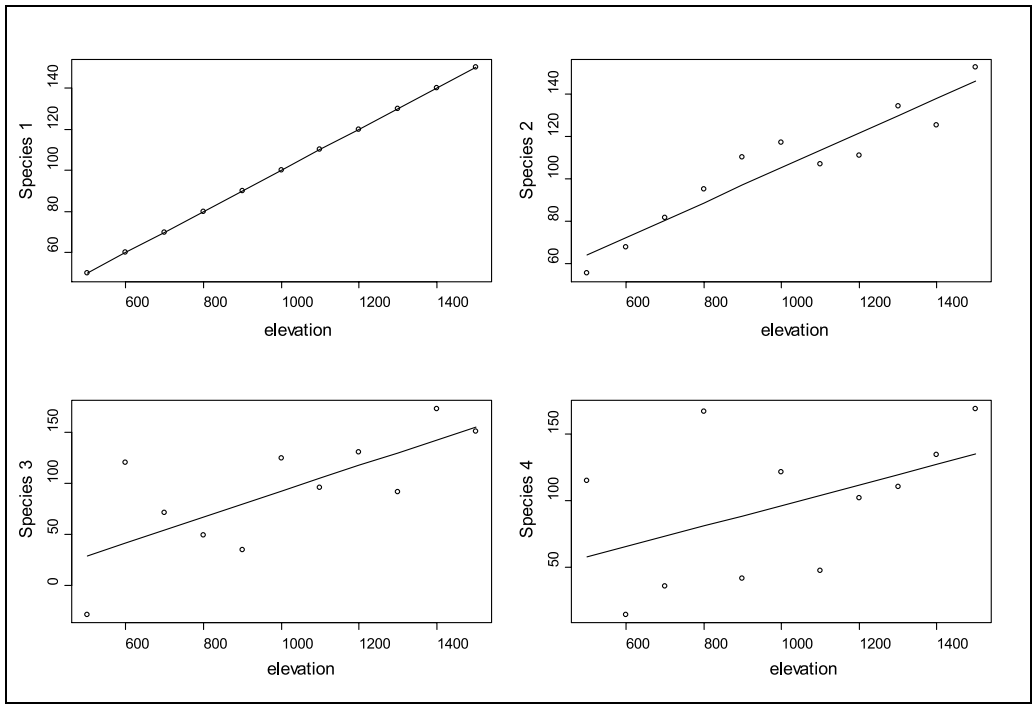


Figure 6.1 Observed values (circles) and linear regression model predictions (lines) for the relationship between elevation and abundance of four species measured on 11 sites. Variance explained by the regression model is 100% for Species 1 ($P < 0.001$), 90% for Species 2 ($P < 0.001$), 53% for Species 3 ($P = 0.01$) and 23% for Species 4 ($P = 0.14$). In each case, visual investigation of the graphs indicated that the linear regression is a sensible model to use. If it were not (for example, the possible relationships were curved), then the values that were calculated for P would be wrong.

A regression model makes a clear differentiation between an explanatory variable and a response variable. You need to specify which variable is the response variable. Regression models considered here only have one response variable.

Using a simple method of analysing species data: linear regression

Regression analysis is a method by which the pattern in one **response variable** (or **dependent variable**) is modelled based on the patterns observed in one or several **explanatory variables** (or **independent variables** or **predictor variables**).

Imagine that we want to investigate the abundance of species *Faramea occidentalis* of the forest surveys conducted in Barro Colorado Island of Panama (Condit et al. 2000; Pyke et al. 2001). Table 6.1 shows part of the data that were collected for the species. We used a subset of the 1-ha plots that belonged to larger sample plots (sites with coding B, C or S), since otherwise the larger plots would have dominated the dataset (see the discussion on mixed models in section: generalized mixed models). The observations on abundance are shown in the same table as the environmental matrix (Table 6.1). Typical for ecological surveys are the many zeros observed for the species abundance. Note the missing data for environmental variables for sites $p40$ and $p41$ – a good statistical program will remove these observations when fitting the model.

Table 6.1 Values of environmental variables and the abundance of *Faramea occidentalis* for various forest plots in Panama (NA indicates missing data)

Site	Precipitation	Elevation	Age	Age.cat	Geology	Faramea.occidentalis
B0	2530.0	120	3	c3	Tb	14
B49	2530.0	120	3	c3	Tb	7
p1	2993.2	20	2	c2	Tc	0
p2	3072.0	100	3	c3	Tc	0
p3	3007.4	180	1	c1	Tc	2
p4	2999.8	180	1	c1	Tc	1
p5	2414.3	40	2	c2	Tgo	0
p6	2393.7	30	2	c2	Tgo	0
p7	2438.4	60	1	c1	Tgo	2
p8	2455.5	50	3	c3	pT	0
p9	2889.3	410	3	c3	pT	0
p10	2529.3	90	3	c3	Tcm	9
p11	2515.5	60	3	c3	Tcm	7
p12	2496.8	10	2	c2	Tbo	1
p13	2576.3	55	2	c2	Tcm	2
p14	2534.7	60	3	c3	Tcm	5
p15	2455.0	70	3	c3	Tgo	0
p16	2501.8	160	3	c3	pT	3
p17	2470.6	120	3	c3	pT	0
p18	2510.8	58	2	c2	Tcm	3
p19	2687.7	160	1	c1	pT	0
p20	2657.5	160	1	c1	pT	0
p21	2411.4	110	1	c1	Tgo	12
p22	2513.7	180	1	c1	Tb	42
p23	2247.5	30	2	c2	Tc	15
p24	2279.8	50	2	c2	Tc	7
p25	2334.3	110	2	c2	pT	0
p26	2251.9	50	2	c2	pT	0
p27	2305.1	180	1	c1	Tl	4
p28	2293.7	160	1	c1	Tl	22
p29	1968.5	100	1	c1	Tb	8
p30	2096.3	180	1	c1	Tb	0
p31	3291.7	343	3	c3	pT	0
p32	3293.1	363	3	c3	pT	0
p33	3615.3	600	3	c3	pT	0
p34	3106.8	210	3	c3	pT	0
p35	4001.7	830	3	c3	pT	0
p36	3029.4	200	3	c3	pT	0
p37	3133.9	600	3	c3	pT	0
p38	2517.4	810	1	c1	pT	0
p39	2400.5	660	1	c1	pT	0
p40	NA	NA	NA	NA	NA	0
p41	NA	NA	NA	NA	NA	0
C1	1887.5	50	1	c1	pT	1
S0	3026.4	140	2	c2	Tc	0

Since we want to investigate a hypothesis about the relationship between precipitation and the abundance of *Faramea occidentalis*, it is good data practice to start with a graph that shows the observations. Figure 6.2 provides this graph. When you have a closer look at the graph, then you notice that sites with precipitation above 3010 mm have none of the trees. Only two sites have some trees in between 2600 and 3000 mm. Below 2600 mm, some sites did not have the species, but many do. We can also spot some sites with abundances that are very high compared to the other abundances. The initial investigation of the graph shows that it is worthwhile to further investigate whether there is a linear relationship between precipitation and the species, since we could see a decreasing trend in abundance with increasing precipitation. The initial investigation also shows that a linear regression model will not explain all variance, as wherever a straight line is placed some observations would not be well predicted because of their scatter. Precipitation can therefore not be the only explanatory variable for differences in abundance.

A linear regression analysis with precipitation as the explanatory variable for differences in abundance provides the following results:

```
lm(formula = Faramea.occidentalis ~ Precipitation, data = faramea,
    na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5606 -4.1558 -1.6295  0.8387 37.4855

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.742059   7.328442   2.285   0.0276 *
Precipitation -0.004864   0.002738  -1.777   0.0830 .

Residual standard error: 7.57 on 41 degrees of freedom
Multiple R-Squared: 0.07149,   Adjusted R-squared: 0.04885
F-statistic: 3.157 on 1 and 41 DF,  p-value: 0.08303


Analysis of Variance Table

Response: Faramea.occidentalis
              Df  Sum Sq Mean Sq F value  Pr(>F)
Precipitation  1  180.91  180.91  3.1569 0.08303 .
Residuals     41 2349.51   57.31

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

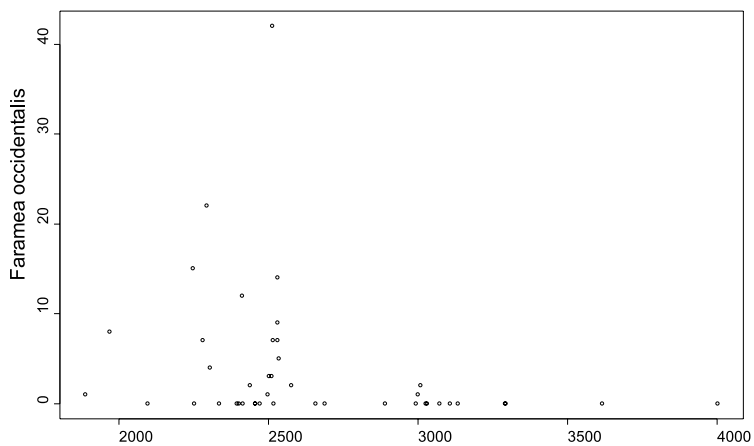


Figure 6.2 Observed values (circles) of the abundance of Faramea occidentalis on precipitation.

What do these results mean? We will take you step-by-step through the output to indicate the interpretation.

The formula shows that in this model, the number of trees of *Faramea occidentalis* is the response variable (at the left of the ~) and precipitation is the explanatory variable (at the right of the ~).

This formula is R language for the model that we fitted. This model has the form of:

Faramea occidentalis = $a + b \times$ precipitation + deviation

In this model, the values for *Faramea occidentalis* and precipitation are the variables of Table 6.1, with observed values for each site, except *p*40 and *p*41.

The *a* and *b* parameters are the **coefficients** or parameters of the model. The software will estimate values for these coefficients. Another name for parameter *a* is the **intercept** and for *b* is the **slope**. Once you have the estimates of the coefficients, then you can calculate the **predicted** or **expected** value for each site, based on the explanatory variables. You can then calculate the predicted abundance of $a + b \times 2530$ for site *B0* (since precipitation at B0 is 2530 mm), and $a + b \times 2993.2$ for site *p1* (since precipitation at *p1* is 2993.2 mm).

Imagine for instance that the model estimated that coefficient $a = 1$ and coefficient $b = 0.02$. In this case, we expect an abundance of $1 + 0.02 \times 2530 = 51.6$ for site *B0* and an abundance of $1 + 0.02 \times 2993.2 = 60.864$ for site *p1*. In case the model calculated $a = 0$ and $b = 0.01$, then we expect 25.3 for site *B0* and 29.932 for site *p1*.

The model will be estimated in a way that the predicted values will be as close as possible to the observed values. The **residual** is the difference between the observed and the predicted value. For the model that calculated an expected abundance of 51.6 for site *B0*, the residual is thus $14 - 51.6 = -37.6$. The model is estimated in a way that will minimize the sum of the squared residuals.

Next in the output after the model, the distribution of residuals is provided. The information gives the minimum, maximum and first and third quartile of the residuals (check chapter 2 how these statistics are calculated). These values allow getting a quick view of the quality of the model. The smaller the residuals are, the better the quality of the model. We can especially notice the large maximum value of 37.4855. Checking how residuals are distributed for a good model is discussed in further detail below (section: checking the residuals of the linear model).

Next in the output, we get the estimated values of the coefficients. The model estimated values of 16.742059 for coefficient *a* and –0.004864 for coefficient *b*. These coefficients allow you to calculate the predicted abundance. For site *B0*, this means that the predicted abundance equals $16.742059 - 0.004864 \times 2530 = 4.43$, and the residual equals $14 - 4.43 = 9.57$.

The standard errors describe the precision with which parameters are estimated. The *t* values and probability values were calculated to test whether the coefficients could be equal to 0.

The test for the coefficient for precipitation, *b*, is built on the idea that behind the sample data is a relationship between abundance and rainfall, that we could find exactly if we took a large enough sample of sites. The test examines the hypothesis that the underlying value of *b* in this hypothetical model is zero. The low (but not small) value of $P = 0.083$ can be interpreted as providing some, but not strong, evidence that the regression coefficient is not zero.

Statistical significance tests such as this t-test are widely used and useful in analysis of experimental and observational data. However they are also often misunderstood and hence misused. All statistical tests depend on assumptions about the data (even so-called 'non-parametric' tests) which can be phrased as a model. In most cases the model describes some pattern, relationship or differences which are of interest. Many of the tests carried out examine whether the data support the notion of

a relationship, or whether the data are consistent with a simpler ('null') model in which there are no differences or relationships. The tests look at how 'likely' the observed data are if the underlying null model describes the real world. If the data are likely to occur even in the absence of the hypothesized relationship, they do not support the hypothesis. If, on the other hand, they are unlikely to occur if the null model is true, this is taken as evidence that the null model does not reflect the real world, and some relationship or differences exist. The significance level (*P*) is the measure of 'how likely'. A low significance level, close to zero, is interpreted as the data not supporting the null hypothesis. A larger significance level is interpreted as no evidence in the data against the null hypothesis.

Three common problems with using statistical significance tests are:

1. Not realizing that the tests' results depend entirely on the statistical models behind the tests being realistic and suitable for the data. It is therefore necessary to understand what these models are and how to check if they are appropriate.
2. Not understanding that the significance level *P* provides a measure of 'strength of evidence'. There is no cut-off value above which we can say 'not significant', even though, for historical reasons, the value of P=0.05 is still sometimes treated that way.
3. Neither the null or alternative models are demonstrated to be 'true' by the results of a test. As a simple example, a test may compare diversity in two land uses. If there is 'no significant difference' it does NOT mean that diversity does not differ between land uses. It means your data have not been sufficient to demonstrate a difference. This might be because there is no difference, or it might be because your data are not adequate for detecting the differences which are there.

Regarding the last point, there are methods available to help decide if you have sufficient data to detect the sort of effects you are interested in. A 'power analysis' can be carried out, or effects estimated together with a confidence interval. Ask a biometrician for help!

The **multiple R-squared value** gives the fraction of variance that is explained by the model. If this fraction is close to 1, then the model explains almost all of the variance. This means that the residuals will be very small, and the predicted and observed values will be close together. The multiple R-squared is thus an expression of the goodness-of-fit or quality of the model (under the assumption that the measured values were realistic). For this model, the value of R-squared = 0.071 (or 7.1%) means that the model only explains a small fraction of the total variance.

The adjusted R-squared value adjusts the multiple R-squared value to the degrees of freedom of the regression model. The purpose of the correction is to enable comparisons between regressions with different datasets, but some researchers have found out that the statistic is not very good at doing that. We are also of the opinion that the adjusted R-squared value provides little extra information than provided by the multiple R-squared value and the results of the F-test (discussed in the next paragraph).

The F statistic tests whether there is no evidence that the model explains some of the variance. You could notice also that the significance level (P = 0.08) is the same as for the coefficient of precipitation given earlier. This is to be expected since the model had only one explanatory variable.

Finally, an **analysis of variance** or **ANOVA** table is given. Analysis of variance and generalizations of it, are widely used in assessing and understanding statistical models. They are basic statistical tools which you will need to become familiar with and perhaps refer to a more detailed text. Dalgaard (2002) contains a simple description of the ideas and the tools within R software. Analysis of variance arises from thinking of statistical analysis and modelling as trying to explain variation in the

response variable. If there was no variation (all the values of the response equal) there is no analysis to do. If the response is related to an explanatory variable, x, then variation in x will lead to variation in the response. Hence a relationship may 'explain' some of the variation. If the explanatory variable is a category or grouping variable, it explains variation in the response if the values of the response within a group tend to be more similar than those in different groups.

The ANOVA table splits up the total variance in the response into components that are explained by each explanatory variable and a residual. This is calculated through sums-of-squares that are then divided by the degrees of freedom, resulting in a mean square. Mean squares are actually variances, since this is the way by which variance is calculated. ANOVA tables give important information on the magnitudes of variance explained by the explanatory variables. The magnitude of variance is an expression of the importance of the explanatory variable in explaining the linear pattern of the response variable.

The significance level values provided by the ANOVA table also relate to tests about the parameters of an underlying model being zero. We can see once again that this significance level is the same as the significance level calculated by earlier tests (P = 0.08), indicating some, but not strong, evidence that precipitation contributes to explaining abundance.

The results of the model can also be analysed in a graphical way. Since the linear regression model fits a straight line that attempts to get as close as possible to the observed values, we can compare the predictions with the actual observations. Figure 6.3 shows this comparison. The dashed lines added to Figure 6.3 show where we are 95% certain that the regression line is. The dotted line of Figure 6.3 corresponds to the area where we expect that 95% of the new observations will be (actually where we expect where the first new observation will be 95% of cases, conditional on the value of precipitation for this observation). These lines are constructed from the probability distribution functions that are assumed to have generated the data.

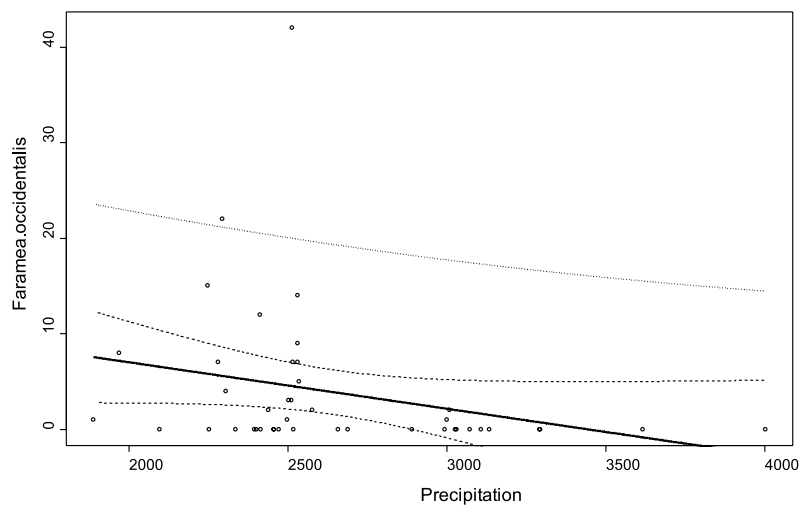


Figure 6.3 Observed values (circles) and predicted values (connected by line) for the linear regression model of the abundance of *Faramea occidentalis* on precipitation. The dashed lines show the 95% confidence interval for the mean, the dotted line the upper 95% prediction interval for new observations.

Checking the residuals of the linear model

Some of the plots that you can request from a decent statistical package are diagnostic plots. These plots help you in evaluating whether some of the assumptions behind the regression analysis are realistic, and hence whether it is safe to accept the results of regression model or not.

What could be the problem? The problem is that a linear regression analysis, like any other statistical analysis method, makes several assumptions about the data to arrive at its results. The most important assumptions are that the effects from various explanatory variables are additive and that there are linear relationships between explanatory variables and response variables – in essence that the formula of the regression model makes sense. Second-order assumptions are that the observations are independent and that the variance of the residuals is constant. A third-order assumption is that the residuals are normally distributed.

A regression analysis can be seen as a model that determines how much pattern can be filtered from a dataset. This concept can be described as:

Data = pattern + residuals

The residuals are the part of the data that were not modelled. Simple regression analysis makes the assumption that no systematic patterns can be observed in the residuals. If the residuals show some patterns, then the model has not explained all the predictable variance.

Since residuals should not show patterns, some diagnostic plots therefore investigate patterns in the residuals. One method is to plot the residuals against the predicted values as shown in Figure 6.4 in the two plots on the left-hand side. You can see that the spread of residuals increases with increasing levels of the predicted value. The assumption of

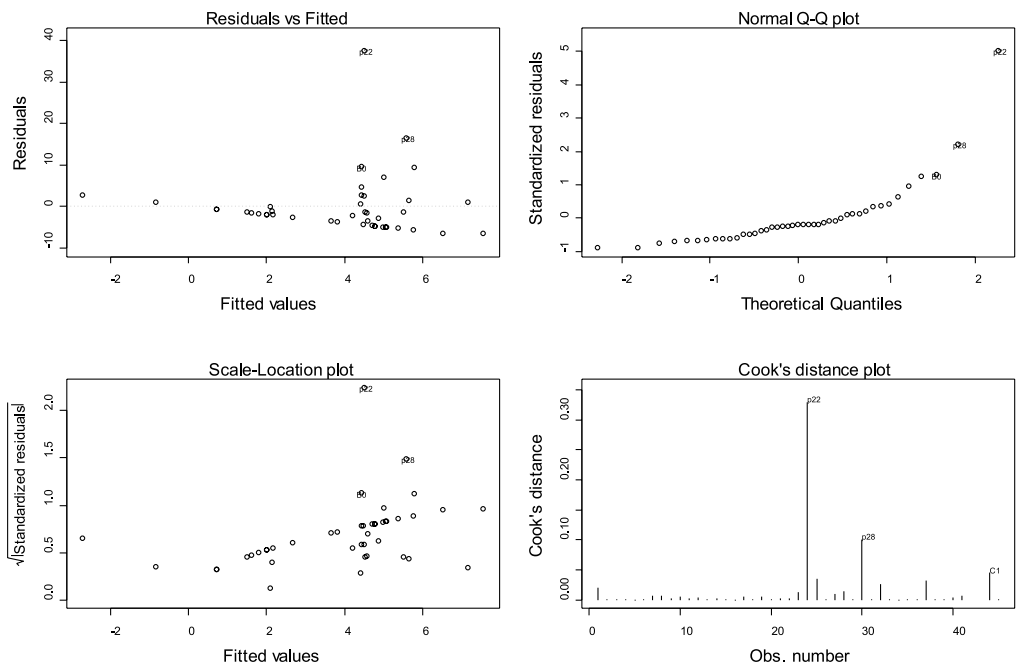


Figure 6.4 Diagnostic plots for a linear regression model.

constant variance is not very realistic. As seen in chapter 2 (Figure 2.3), a variable that is normally distributed will result in a straight line in a Q-Q plot. This provides another method for checking the residuals, since residuals are expected to be normally distributed. The Q-Q plot that is shown on the upper right-hand side shows that residuals are not normally distributed.

The Cook's distance graph (lower-right) shows observations that have an important influence on the results. If site p22 was removed from the dataset, then different results would be obtained. How such observations can be handled is discussed in further detail at the end of this chapter.

The model assumes the residuals are just random, unpredictable deviations. Biodiversity data are collected in space, so an obvious pattern to look for is coherent geographical variation not accounted for by the environmental variables in the model. This can be investigated graphically by using the spatial location of the sample plot as the plotting positions. There should not be any spatial pattern that you can observe, for instance having negative residuals in the north and positive residuals in the south, or patches of high and low residuals. This is important to check since the model assumes that all patterns in abundance can be explained by precipitation.

One method of investigating spatial patterns in the residuals is to directly plot the residuals using the spatial coordinates of each sample plot (Figure 6.5a). A more sophisticated method of investigating the spatial distribution of residuals is to construct a trend-surface that models the changes in the residuals over the spatial coordinates. The trend-surface will model the changes in the response variable (the residuals) as a landscape with hills and valleys for higher and lower values of the variable, respectively (Figure 6.5b). The trend-surface model makes it easier to spot any trends in the residuals.

The output shows that there is a noticeable trend in the residuals: the residuals decrease from an area in the centre to areas in the north and south. If some explanatory variables other than precipitation show the same trend, then these could be important explanatory variables to add to the regression model to further explain the variance in abundance. On the other hand, we may have a patch of the species which can not be explained by forcing variables. Ecological models can show that patchy distributions can arise simply from the

Figure 6.5 (a) Spatial distribution of residuals of the linear regression model of the abundance of *Faramea occidentalis* on precipitation: sign (positive = circle, negative = square) and size (size of circle or square) of residuals at a particular spatial location (+);

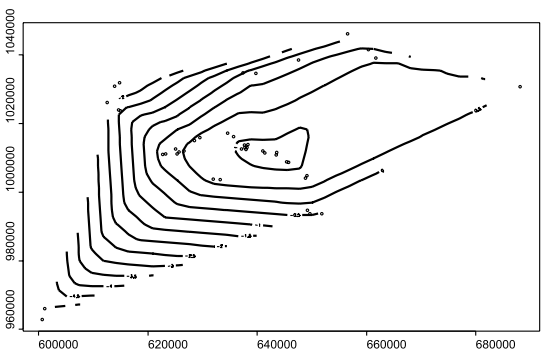


Figure 6.5 (b) Second-order polynomial trend surface for the residuals.

natural dynamics of the species and do not have to be driven by environmental variables.

When patterns in the residuals show that the assumptions of the model are unreasonable (as seen above), then we should look for better models (such as those listed later in this chapter). It is very important that the assumptions of the model apply when you want to make any conclusions from the results of the model.

Linear regression with a categorical explanatory variable

Imagine that you did not want to investigate the influence of altitude on abundance, but the influence of geology on abundance. The dataset includes a categorical variable ("geology") that classifies sites according to various categories of rock types (see Chapter 2). You can construct a model in a similar way as done above, and you would obtain the following results:

```
lm(formula = Faramea.occidentalis ~ Geology, data = char, na.action = na.exclude)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2222     1.5552   0.143 0.887172
GeologyTb    13.9778     3.3355   4.191 0.000172 ***
GeologyTbo    0.7778     6.7788   0.115 0.909292
GeologyTc     3.3492     2.9390   1.140 0.261989
GeologyTcm    4.9778     3.3355   1.492 0.144313
GeologyTgo    2.5778     3.3355   0.773 0.444663
GeologyTl    12.7778     4.9179   2.598 0.013494 *

Residual standard error: 6.598 on 36 degrees of freedom
Multiple R-Squared: 0.3806,   Adjusted R-squared: 0.2774
F-statistic: 3.688 on 6 and 36 DF,  p-value: 0.005868


Analysis of Variance Table

Response: Faramea.occidentalis
          Df  Sum Sq Mean Sq F value   Pr(>F)
Geology    6  963.19  160.53  3.6875 0.005868 **
Residuals 36 1567.23   43.53

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

These results are similar as the results shown earlier for the continuous variable. What is different now is that the model includes a categorical explanatory variable. A coefficient is estimated for each level of the categorical variable.

The regression coefficient for the most common category of geology (*pT*) is fixed to be zero. You could choose to fix the coefficient for another category to zero, but always one coefficient needs to be zero. Since the regression coefficient is zero, it is not provided in the output. It is taken as a reference level against which others are compared.

Imagine that you had a simple dataset with three observations on tree abundance, one for each level of the categorical variable Landuse. You fit the following model: *Abundance* = a + b × *landuse1indicator* + c × *landuse2indicator* + d × *landuse3indicator* + *deviation*. The landuse indicators are variables that have values 1 or 0. When a site has landuse1, then *landuse1indicator* = 1, *landuse2indicator* = 0 and *landuse3indicator* = 0 for this site. If you observed 6, 7 and 8 trees

for *landuse1*, *landuse2* and *landuse3* respectively, then a perfect fit would be provided by estimating a=5, b=1, c=2 and d=3. However, another perfect fit would be provided by estimating a=4, b=2, c=3 and d=4. From the infinite number of possible combinations, combination a=6, b=0, c=1 and d=2 is selected by setting b=0.

The same choice was made in the regression model by opting that the regression coefficient of *pT* was fixed to be zero. For our example, we thus predict an abundance of 0.2222 + 0 = 0.2222 for category *pT* and an abundance of 0.2222 + 13.9778 = 14.20 for category *Tb*.

The interpretation of the results is analogous to the interpretation of regression model with a quantitative explanatory variable. For example, there is evidence that sites with geology *Tb* contain more of the species than sites of geology *pT* since a low significance level was estimated ($P < 0.001$). There is no evidence that sites of geology *Tbo* contain more trees than sites of geology *pT* since a high significance level was estimated ($P = 0.909$).

The multiple R-squared value shows that the model explains 38% (0.3806) of variance. The significance level of the F-test indicates that the model provides evidence for a linear relationship between geology and abundance ($P = 0.0058$). The ANOVA table provides the same information since we only had one explanatory variable.

The graphical presentation of the model is given in Figure 6.6. The observed values are presented as circles, the predicted values for each category of geology by the full lines. You could check that the predicted values correspond to those calculated earlier based on the regression coefficients. In this simple case they are actually just the means for each category. The dashed and dotted thick lines show where we are 95% certain that the average and next value for each category will be.

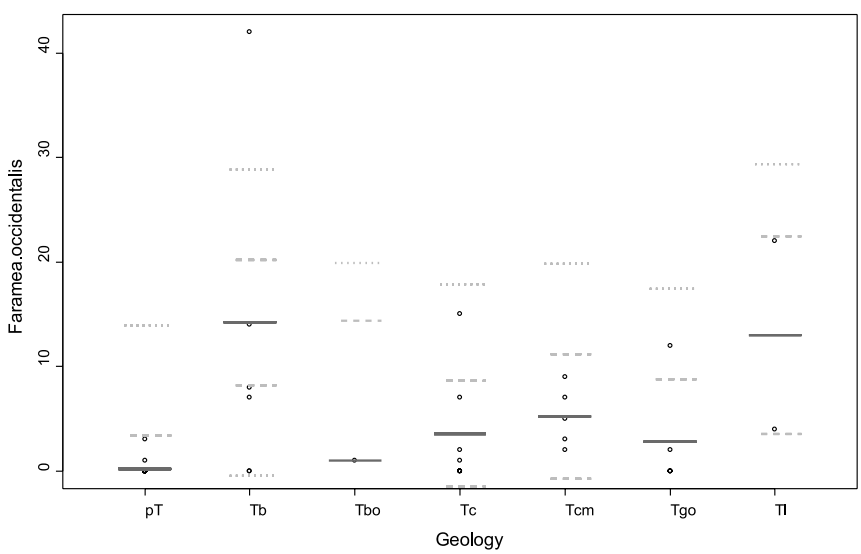


Figure 6.6 Observed values (circles) and predicted values (full line) for the linear regression model of the abundance of *Faramea occidentalis* on geology. The dashed lines show the 95% confidence interval for the mean, the dotted lines the 95% prediction interval for new observations.

As with a continuous variable, we should proceed with checking diagnostic plots for the patterns in the residuals. In this manual, the diagnostic plots will only be given for the first regression analysis in order to save some space, however. Remember to always check the diagnostic plots for a regression analysis.

Transforming the response variable when the residuals are not normally distributed

One way of overcoming the problem with the increasing variance of residuals for larger values of the response variables is to transform the response variable. Some common transformations used with count data are the logarithmic transformation and the square-root transformation. These transformations are not tricks that hide the actual patterns, but are useful tools to reveal patterns.

The major disadvantage of using transformations is that you are not modelling the patterns of the observations that you made, but patterns of the transformed observations. When you interpret your results, you are actually interpreting the transformed observations. In some cases, you may not feel comfortable in interpreting transformed observations. In other cases, such interpretation can seem logical. For instance, the pH scale to measure acidity is actually a logarithmic scale.

Note that the log transformation can not be used if the response variable includes zeros. Log(0) is not defined. A practical solution in such cases is to calculate the transformed value of n as $\log(n+1)$ or $\log(n+0.1)$ instead of $\log(n)$. The results depend on what is added, introducing a certain arbitrariness to the analysis.

Because of the problems with transformations, we advise to use more modern approaches to data analysis such as the GLM methods that are introduced immediately below.

Using generalized models when the residuals of a linear regression are not normally distributed: the Poisson, quasi-Poisson and negative binomial models

Generalized linear models (GLM) were invented to deal with situation where observations are not normally distributed or where other aspects of the linear regression model are not appropriate. They fit a wider, more general class of models that can cope with other situations. In the case of counts data, you know that values should never be negative, but a simple linear regression model provides no guarantee that you would obtain such results. For instance, in Figure 6.3 you can see that the model predicts negative abundances when precipitation is larger than 3500 mm. This is not a realistic result, as we know that abundance can not be negative. By choosing the appropriate GLM, you will only obtain realistic values.

There are many types of GLM that can be constructed. A GLM is characterized by two functions. One function (the link function) describes how the mean of the response variable depends on the linear predictors (the explanatory variables). The second function (the variance function) captures how the variance of the response variable depends on the mean. The same information is usually provided by the following formulae (μ: mean of the response variable y; x: explanatory variable; a, b: regression coefficients; var: variance; θ: dispersion parameter):

Link function: $g(\mu) = a + b_1 \times x_1 + b_2 \times x_2 + b_3 \times x_3 + \ldots$

Variance function: $var(y) = \theta \times V(\mu)$

The types of GLM that are shown here use the log link and the Poisson, quasi-Poisson and negative binomial variance functions. These are types of GLM that are sometimes appropriate for counts data.

The Poisson GLM (with log link) is the

simplest GLM model to use. It uses a logarithmic link function (log(μ)) and a Poisson variance (var(y) = μ).

When you use this type of GLM on the *Faramea occidentalis* abundance data used earlier, then you obtain the results shown below.

The results are similar to those from a linear regression, but with some important differences.

When we look at the output, then we first get the model. The difference with the linear regression is that the variance and link functions are mentioned.

The coefficients that are provided next are the coefficients that were calculated for the model. You can use these coefficients again to calculate the expected abundance of *Faramea occidentalis* at a given precipitation. The coefficients predict the logarithm of the abundance, however, since a log link was used. Thus, to know the expected abundance, you need to take the anti-logarithm. For site *B0* with precipitation of 2530 mm, the predicted abundance thus equals exp(5.6668474 – 0.0017098 × 2530) = 3.82.

The output continues with mentioning that the dispersion parameter was taken to be 1, which means that the model assumed that var(y) = μ. This is one of the assumptions that the Poisson model makes. This is what would be expected if the individuals were randomly located in space. When individuals are clumped, the dispersion parameter will be larger than 1 (var(y) = θ × μ with θ > 1). When individuals are more regularly distributed than random, the dispersion parameter will be smaller than 1.

The null and residual deviances are similar to the total variance and residual variance of a simple linear model. If the difference between them is large, the model will explain much of the variance (deviance). If the difference is small, then the model is not very effective in explaining the response variable. In our case, the model only explains ((414.81 – 357.67) / 414.81) or 13.7% of total deviance. We can thus use the null and residual deviance to calculate a parameter that can

```
glm(formula = Faramea.occidentalis ~ Precipitation, family = poisson(link = log),
    data = faramea, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0071  -2.6373  -1.4418  -0.1339  11.0835

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.6668474  0.6047651   9.370  < 2e-16 ***
Precipitation -0.0017098  0.0002478  -6.901 5.18e-12 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 414.81  on 42  degrees of freedom
Residual deviance: 357.67  on 41  degrees of freedom
AIC: 431.55


Analysis of Deviance Table

             Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                           42     414.81
Precipitation  1    57.14        41     357.67 57.142 4.054e-14 ***

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

be interpreted in the same way as the multiple-R-squared value of a simple linear regression.

The ANOVA table also does not mention 'variances' but reports 'deviances' instead. The similar name already indicates that deviances can be interpreted in the same way as the variances of a GLM. Although the model is not very efficient in explaining the abundance of *Faramea occidentalis*, there is evidence that precipitation explains some of the deviance since the F-test has a small significance level ($P < 0.001$). Note that evidence for explaining some of the deviance does not mean that much deviance is explained. It is therefore necessary to check for both the significance level (to judge whether there is an effect) and the deviance explained (to judge how important the effect is). You could check that the deviance explained by precipitation and the residual deviance sum up to the null deviance.

The graphical representation of the model is provided in Figure 6.7. You can see once more that there is big difference between the actual abundances and the modelled abundances, or large residual deviance.

You could verify that the Poisson model will never predict negative values. You can see that predictions for elevations above 3500 mm are still positive, contrary to the simple linear model that we saw before. This is a good feature of such model for count data, such as abundances of species. Therefore, if you know that your data are counts, it is often better to use a suitable GLM rather than a linear model.

A **quasi-Poisson GLM** is very similar to a Poisson model (as the name suggests), but uses a different variance function. This model is better when the dispersion is not close to 1, an assumption that is used by the Poisson model (see above). The quasi-Poisson makes the assumption that dispersion is not 1 and fits a dispersion parameter to the dataset.

When you fit a quasi-Poisson GLM (with log link) to our data, then you obtain the result shown on the next page.

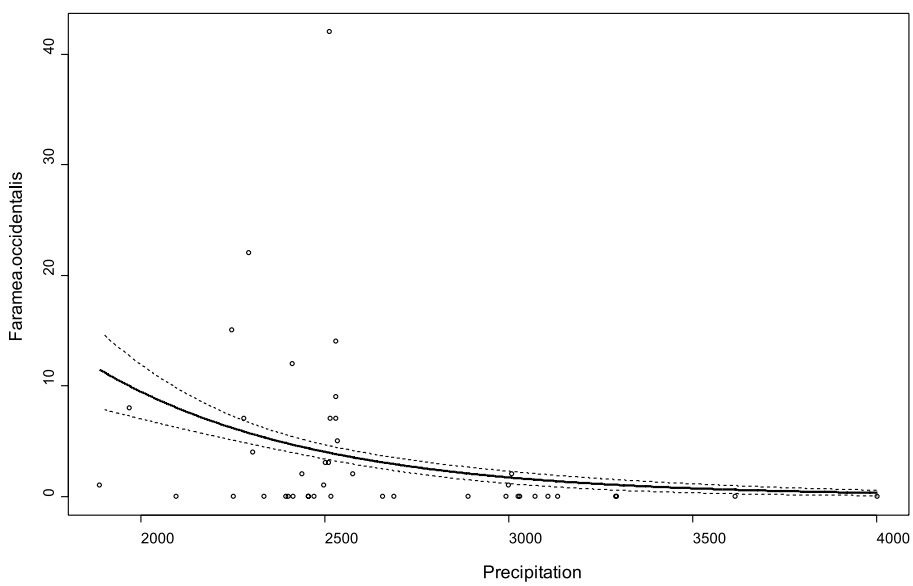


Figure 6.7 Observed values (circles) and predicted values (connected by line) for the Poisson GLM with log link of the abundance of *Faramea occidentalis* on precipitation.

```
glm(formula = Faramea.occidentalis ~ Precipitation, family = quasipoisson(link = log),
    data = faramea, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0071  -2.6373  -1.4418  -0.1339  11.0835

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.6668474  2.2452852   2.524   0.0156 *
Precipitation -0.0017098  0.0009199  -1.859   0.0703 .

(Dispersion parameter for quasipoisson family taken to be 13.78382)

    Null deviance: 414.81  on 42  degrees of freedom
Residual deviance: 357.67  on 41  degrees of freedom
AIC: NA


Analysis of Deviance Table

             Df Deviance Resid. Df Resid. Dev      F  Pr(>F)
NULL                             42     414.81
Precipitation  1    57.14        41     357.67 4.1456 0.04824 *

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

When you compare the output of the quasi-Poisson with the output of the Poisson model, then you can see that the regression coefficients and the deviance are the same. What is different are the standard errors and significant levels of the tests. For the Poisson model, the significance level for the coefficient for precipitation is small ($P < 0.001$), whereas for the quasi-Poisson model it is larger ($P = 0.07$) and indicates some, but not strong evidence for an effect. By not assuming that the dispersion parameter was 1 as in the Poisson model, we thus reach different conclusions that there is only some (not strong!) evidence for an effect of precipitation on abundance. Since the dispersion parameter was estimated to be 13.78, there is an indication that the individuals are not randomly distributed, but are clumped. The analysis of the residuals (Figure 6.4) also indicated that individuals could be clumped. In such situations, it is more appropriate to use the quasi-Poisson than the Poisson model. Since the results of a model will depend on the assumptions that the model makes, you should try to ensure that the assumptions are realistic. When the model makes unrealistic assumptions, you will not be able to reach reliable conclusions.

Notice that the significance levels for the effect of precipitation are not quite the same for the t-test for the regression coefficient ($P = 0.07$) and the F-test of the ANOVA table ($P = 0.048$). They are based on different approximations. However, qualitatively they are the same: both suggest evidence for precipitation having an effect. Do not use $P = 0.05$ as a cut-off between significant and non-significant results, but use P as a scale for measuring evidence. In this case both probabilities suggest some but not strong evidence against the null hypothesis of no precipitation effect. However both results depend on the model being appropriate, and a look at observed and predicted values, or the residuals, shows that this may be doubtful.

The graphical representation of the model is provided in Figure 6.8. When you compare this figure with Figure 6.7, then you will see that the only features that are different are the wider

confidence intervals (dashed lines) in Figure 6.8, reflecting the large estimated dispersion parameter, compared with the inappropriately fixed value of 1 in Figure 6.7.

The **negative binomial GLM** is another model that can be used for situations where dispersion is higher than 1, or where individuals are clumped and not randomly distributed. Since organisms often show a clumped distribution, this model will often be suitable for ecological research. Whereas the quasi-Poisson model does not correspond to a known statistical distribution, the negative binomial model is one of several statistical models that model clumping. The negative binomial model is a model with one parameter more than the Poisson model, the parameter theta or k. This parameter models the clumping in the data, ranging from zero to infinity. Values of theta close to zero indicate clumping, whereas larger values indicate distribution that is more random. An infinite value of theta gives a Poisson distribution with dispersion equal to one.

When you fit a negative binomial GLM (with log link) to the same data that we used before, then you will obtain the result shown on the next page.

You can see that the output is similar to the outputs of the Poisson and quasi-Poisson models. The model coefficients are different for the negative binomial GLM, however. As for the Poisson model, there is evidence that precipitation has an effect on abundance since a small significance level is calculated for the coefficient for precipitation and in the ANOVA table ($P < 0.001$). By modelling clumping directly rather than by a second-order assumption as in the quasi-Poisson GLM, we thus obtain a different result.

We can see that a small value was estimated for theta (0.3057), since it can theoretically range from zero to infinity. This provides evidence that individuals are clumped.

Figure 6.9 provides the graphical presentation of the results of the negative binomial model (with log link). You can see that the model predicts very large abundance at lower precipitation levels. Since the observed abundances at the lowest precipitation levels are not the highest, we should be sceptical about these results – remember that the residuals should not show any patterns.

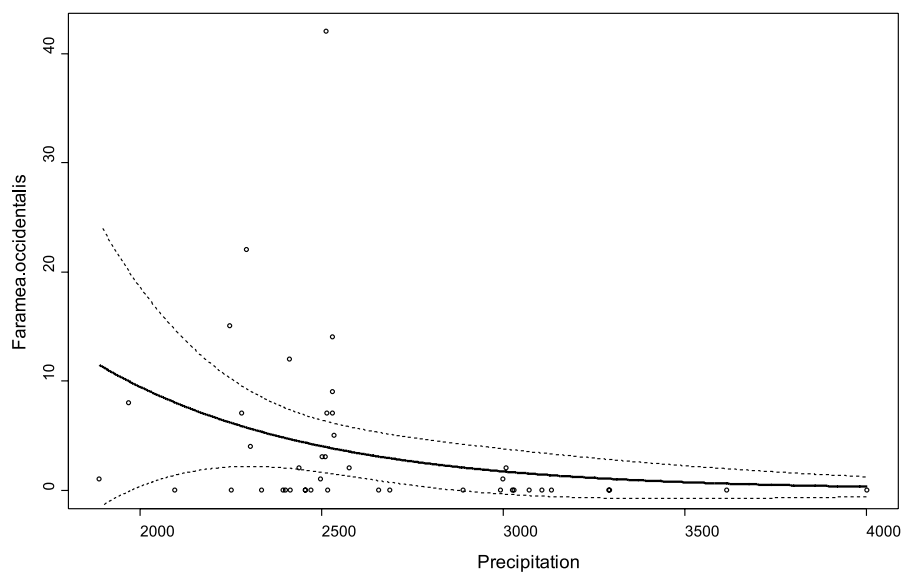


Figure 6.8 Observed values (circles) and predicted values (connected by line) for the quasi-Poisson GLM with log link of the abundance of *Faramea occidentalis* on precipitation.

```
glm.nb(formula = Faramea.occidentalis ~ Precipitation, data = faramea,
    na.action = na.exclude, maxit = 5000, init.theta = 1, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.59806  -1.19031  -0.58758   0.03517   2.12815

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   11.2726451  2.4981876   4.512 6.41e-06 ***
Precipitation -0.0039579  0.0009817  -4.032 5.54e-05 ***

(Dispersion parameter for Negative Binomial(0.3057) family taken to be 1)

    Null deviance: 47.240  on 42  degrees of freedom
Residual deviance: 36.357  on 41  degrees of freedom
AIC: 178.21

              Theta:  0.3057
          Std. Err.:  0.0915


Analysis of Deviance Table

              Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                             42     47.240
Precipitation  1   10.883        41     36.357 10.883 0.0009706 ***

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

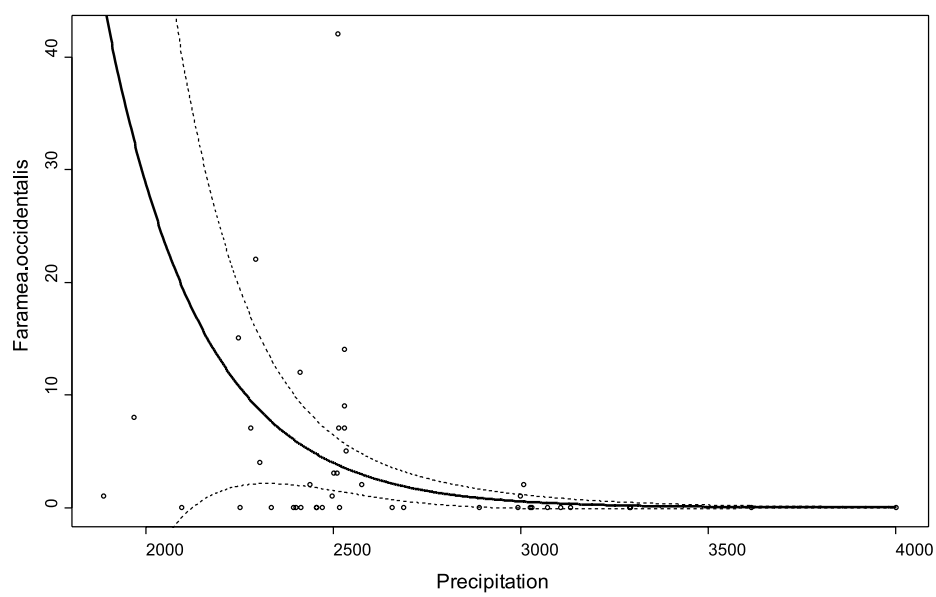


Figure 6.9 Observed values (circles) and predicted values (connected by line) for the negative binomial GLM with log link of the abundance of *Faramea occidentalis* on precipitation.

Using generalized additive models

Generalized additive models (GAM) are more general than GLM. They are based on smoothing – they fit a model that can follow the pattern in the data more closely. Because of this smoothing, the relationships with the explanatory variables are not linear any longer. A smoothing function is a line that flows more freely between the observations than a straight line. Returning to the symbolic description of the assumptions that we used for a GLM, the variance function remains the same, but the explanatory variables part of the model changes into:

Link function: $g(\mu) = a + b_1 \times x_1 + b_2 \times x_2 + s_1(x_3) + s_2(x_4) + \ldots$

The functions s_1 and s_2 are smooth functions of x that are defined in a way which allows a lot of flexibility in the curve.

We fit a negative binomial GAM (with log link) again to the count data of *Faramea occidentalis* using precipitation as an explanatory variable, which produces the result shown below.

You may notice that no coefficient is provided for precipitation. The output provides a significance test for precipitation, however, using another type of test (with $P = 0.002$). This means that precipitation plays a significant role in explaining abundance, and should be left in the model. The `edf` indicates the estimated degrees of freedom of the smoothing function – these degrees of freedom are similar to the order of the polynomial model (see what a polynomial model is in section: using several explanatory variables at the same time).

Figure 6.10 provides the graphical representation of the model. You can see that the model now predicts the highest abundance at low precipitation with a gradual decrease in abundance until a precipitation of around 3250 mm, followed by precipitation levels where no abundance is expected. You can see that the fitted line is no longer straight, but is more flexible in following the data. If such smooth patterns exist in your data, but you can not find a simple mathematical model to describe them, a smoothing curve may be appropriate. The figure hints that abundance could be lower for precipitation levels beyond the lower limit of the precipitation that was recorded, since the optimal abundance is predicted around 2100 mm. Since it is dangerous to extrapolate, the best way for testing this would be to add some sites that were sampled at lower precipitation levels.

```
Family: Negative Binomial(0.4417)

Link function: log

Formula:
Faramea.occidentalis ~ s(Precipitation)

Parametric coefficients:
             Estimate  std. err.    t ratio    Pr(>|t|)
(Intercept)    0.55676     0.3582      1.554    0.12797

Approximate significance of smooth terms:
                       edf       chi.sq     p-value
s(Precipitation)     1.959       14.393     0.0020351

R-sq.(adj) =  0.0461   Deviance explained = 31.1%
GCV score = 1.1264   Scale est. = 1.0489    n = 43
```

Using several explanatory variables at the same time

In the previous models, we only used one explanatory variable for each model. We can use several explanatory variables at the same time, however. Such regression is called a **multiple regression**. You can construct models that use several (or all) of the explanatory variables that you have – an example is provided later in this chapter.

You can also construct models that use new explanatory variables that were derived from existing explanatory variables. The first example of multiple regression is of the second category. The explanatory variables that will be used are the precipitation and precipitation squared. Precipition^2 is easily calculated by squaring each value of precipitation – for site *B0* the value for precipitation^2 = 2530^2 = 2530 × 2530 = 6400900. When you add square or higher order powers of existing variables, then you are fitting a **polynomial model**. A second-order polynomial model includes powers of original variable until the second order, a fourth-order polynomial model includes powers of the original variable until the fourth order (thus variable, variable^2, variable^3 and variable^4). If we are investigating the relationship between precipitation and abundance with a second-order polynomial model, we fit the coefficients a, b and c of the polynomial model: Abundance = $a + b \times$ precipitation + $c \times$ precipitation^2 + deviation.

By constructing a polynomial model, you can fit curved lines. Using polynomial models thus provides an alternative approach to fitting curved relationships than the smoothing approach shown earlier.

A negative binomial GLM (with log link) of the abundance of *Faramea occidentalis* using the second-order polynomial of precipitation gives the result shown on the next page.

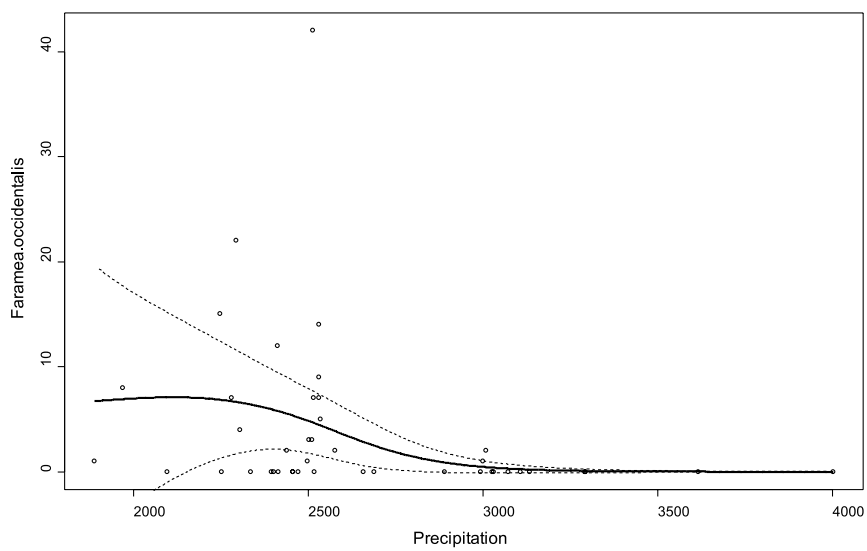


Figure 6.10 Observed values (circles) and predicted values (connected by line) for the negative binomial GAM (with log link) of the abundance of *Faramea occidentalis* on precipitation.

```
glm.nb(formula = Faramea.occidentalis ~ Precipitation + I(Precipitation^2),
    data = faramea, na.action = na.exclude, maxit = 5000, init.theta = 1, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.47623  -1.27733  -0.40550   0.08838   1.82770

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.124e+01  1.708e+01  -1.829   0.0674 .
Precipitation       2.872e-02  1.346e-02   2.133   0.0329 *
I(Precipitation^2) -6.214e-06  2.642e-06  -2.352   0.0187 *

(Dispersion parameter for Negative Binomial(0.364) family taken to be 1)

    Null deviance: 53.418  on 42  degrees of freedom
Residual deviance: 36.043  on 40  degrees of freedom
AIC: 175.51

              Theta:  0.364
          Std. Err.:  0.114

Analysis of Deviance Table

Model 1: Faramea.occidentalis ~ 1
Model 2: Faramea.occidentalis ~ Precipitation + I(Precipitation^2)
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        42     53.418
2        40     36.043  2   17.376 8.6878 0.0001686 ***

Terms added sequentially (first to last)

                   Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                                  42     53.418
Precipitation       1   12.336        41     41.083 12.3359 0.0004443 ***
I(Precipitation^2)  1    5.040        40     36.043  5.0397 0.0247732 *

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

When you compare the output to previous outputs, then you could notice that the results are very similar to those of the other GLMs. The difference is that two coefficients are provided for precipitation and precipitation^2, and also two rows are provided in the ANOVA table. Precipitation^2 is mentioned as `I(Precipitation^2)`. The reason is that otherwise the model would calculate the sum of (precipitation + precipitation^2) and treat this sum as a single explanatory variable – it is just a particularity of the statistical software that we used (I() is a function that isolates the variable). There is evidence that both precipitation and precipitation^2 explain the abundance, since significance level values are low for both variables (for example P = 0.018 for precipitation^2). Therefore both explanatory variables can be left in the model.

The ANOVA table provides similar evidence that both precipitation and precipitation^2 explain the abundance, since the estimated probabilities are small ($P < 0.05$). As there are two variables, the ANOVA table splits the total of the deviance that is explained by the model (explained deviance = null deviance – residual deviance = 53.418 – 36.043 = 17.375) into the deviance that is explained by precipitation (12.336) and the additional deviance that is explained by precipitation^2 (5.040). You could easily check that the sum of both deviances adds up to the total deviance that is explained.

Note that the ANOVA table indicates that the variables were added sequentially. This means

that first the deviance that was explained by precipitation was calculated. After calculating that deviance, the additional deviance that is explained by precipitation^2 is calculated. If you would change the order, then you would obtain different values for deviance. The reason that the sequence will alter the results is that there is correlation between the variables (or the variables are not orthogonal). Because of the correlation, some of the deviance that would be explained by a variable will be explained by a variable that was added into the model earlier – once some deviance was explained, it can not be explained again. For a polynomial model, it makes sense to order variables in order of increasing power as shown in the example, and not to add precipitation^2 before precipitation.

The ANOVA table also listed a comparison between the null model and the model with both variables. This is the GLM alternative to the F-test of a simple linear regression, and it also calculates the significance of the complete model.

Figure 6.11 shows the graphical representation of the model. When you compare this figure with Figure 6.10, then you will notice that a similar shape of curve is obtained for the expected values. In this case, however, a clear optimum in predicted abundance can be seen that occurs around a precipitation of 2300 mm. At lower or higher precipitation levels, a lower abundance is predicted. The reason for this pattern is that a second-order polynomial model in combination with a log scale will fit a unimodal distribution (at a linear scale, second-order polynomial models are rarely useful as they fit a parabola). Many species have unimodal distributions, since there is only a certain range of conditions under which the species occur. Such window where the species occurs can be caused by environmental conditions that are too harsh (too hot, too cold, too dry, too few nutrients, …), or by competition with species that are better adapted to some types of conditions. Since species often have unimodal distributions, a second-order polynomial model will often be the model that will best describe the actual distribution of species abundances. In such situations, the assumption of a unimodel distribution will be appropriate.

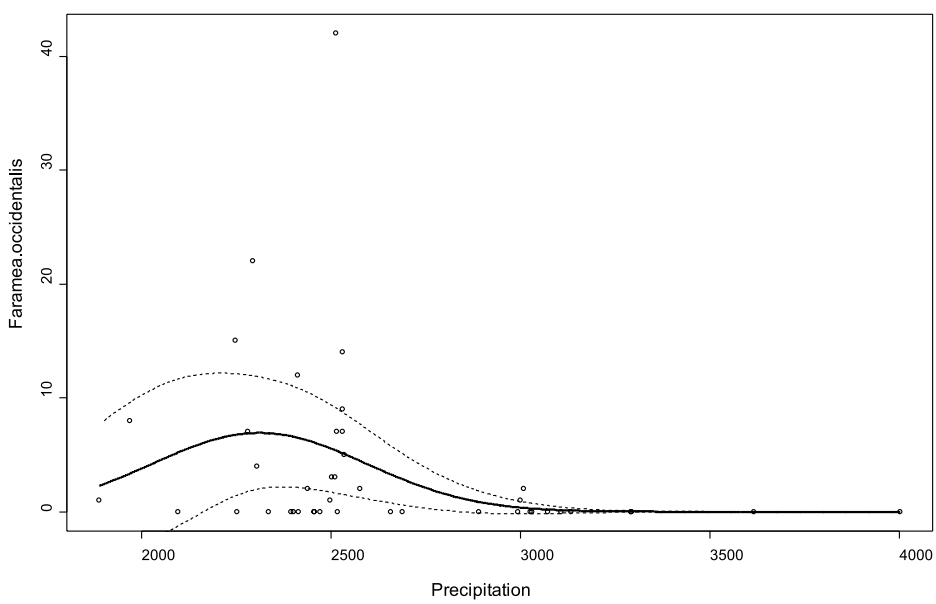


Figure 6.11 Observed values (circles) and predicted values (connected by line) for the negative binomial GLM of the abundance of *Faramea occidentalis* on the second-order polynomial of precipitation.

Note however that the second order polynomial will always give a symmetrical unimodal response. The fact that the fitted response is of that shape is determined by the model type. Whether it fits your data well still needs examining.

In multiple regression models, the explanatory variables that are used do not have to be polynomials of the same variable. The result shown on the next page was obtained by using the categorical variables age and geology, and the second-order polynomials for the quantitative variables precipitation and elevation to explain abundance with a negative binomial GLM with log link.

Note first that although it is technically possible to construct models that include many different explanatory variables, that models should correspond to hypotheses about the influence of an explanatory variable on the response variable. The choice of explanatory variables that are measured in the first place should be based on such hypotheses, which should be realistic or plausible relationships that reflect what we know of ecology.

We analysed age as if it was a categorical variable. Since age is an ordinal variable, we could also have analysed it as if it was a quantitative variable (see Chapter 2). We analysed age as a categorical variable since an analysis with age as a quantitative variable indicated that there was no evidence that age had an effect (see how to interprete ANOVA tables lower in this section and check for yourself with a model where age is a quantitative variable).

As we saw in the outputs of other models, the formula and the distribution of residuals is provided first, followed by the regression coefficients.

Again we can see one regression coefficient for the continuous variables (precipitation, precipitation^2, elevation and elevation^2), and regression coefficients for all but one of the levels of the categorical variables (age and geology). We can see that small significance levels were estimated for the majority of variables.

The negative binomial model calculated a parameter theta that indicates that individuals are clumped since it is not large (4.08).

The model now explains most of the deviance in the data, with an explained deviance of ((210.25-36.28)/210.25) or 82.7% of total deviance. For an ecological model, the explained deviance is very high.

Different to the previous outputs is that two types of ANOVA tables are given. The second one is a type-II ANOVA, which is based on deletions of variables from the model. The type-II ANOVA lists the residual deviance for several models where one variable was deleted. You can verify that the deviance of 41.278 when elevation^2 is removed from the model (as provided by type-II ANOVA) corresponds to the residual deviance of 41.278 after precipitation, precipitation^2, geology, age and elevation were added to the model (as given by type-I ANOVA). The type-II ANOVA investigates whether there is evidence that removing one variable would result in a significantly lower deviance that is explained by the simplified model. For a normal (type-I) ANOVA that we showed earlier, the sequence by which the variables are listed in the model may influence the results of the ANOVA. This will be only the case if the variables are correlated. This is not the case for a type-II ANOVA, where the sequence will not influence the results. We advise to only use a type-II ANOVA when there is no logical order in which the variables should be entered in the model.

Analysing ANOVA tables for models with several explanatory variables may especially be useful when searching for alternative models for the same dataset. When we look at the effect of deleting age from the model, we can see that the more complex model explains more deviance (has less residual deviance: 36.280 – 48.396 = -12.116 or -5.7%), but the simpler model uses one variable less. Which model is better? This will depend partially on what you value more, a higher percentage of deviance that can be explained, or a higher degree of simplicity in your model. There are some statistical criteria that allow choosing between

```
glm.nb(formula = Faramea.occidentalis ~ Precipitation + I(Precipitation^2) +
    Geology + Age.cat + Elevation + I(Elevation^2), data = faramea,
    na.action = na.exclude, maxit = 5000, init.theta = 1, link = log)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.908282  -0.590816  -0.008352   0.136604   2.446988

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -8.800e+01  1.775e+01  -4.957 7.14e-07 ***
Precipitation       7.277e-02  1.438e-02   5.060 4.20e-07 ***
I(Precipitation^2) -1.514e-05  2.931e-06  -5.165 2.40e-07 ***
GeologyTb           2.752e+00  6.374e-01   4.318 1.57e-05 ***
GeologyTbo          3.422e+00  1.754e+00   1.951 0.050999 .
GeologyTc           5.018e+00  9.283e-01   5.405 6.48e-08 ***
GeologyTcm          2.683e+00  7.106e-01   3.776 0.000159 ***
GeologyTgo         -9.910e-02  8.894e-01  -0.111 0.911288
GeologyTl           1.593e+00  7.763e-01   2.052 0.040217 *
Age.catc2          -3.230e+00  9.734e-01  -3.318 0.000906 ***
Age.catc3          -2.162e+00  7.652e-01  -2.825 0.004727 **
Elevation           4.973e-02  2.613e-02   1.903 0.057029 .
I(Elevation^2)     -2.387e-04  1.102e-04  -2.167 0.030248 *

(Dispersion parameter for Negative Binomial(4.0754) family taken to be 1)

    Null deviance: 210.25  on 42  degrees of freedom
Residual deviance:  36.28  on 30  degrees of freedom
AIC: 152.99

              Theta:  4.08
          Std. Err.:  2.39


Analysis of Deviance Table

Model 1: Faramea.occidentalis ~ 1
Model 2: Faramea.occidentalis ~ Precipitation + I(Precipitation^2) + Geology +
    Age.cat + Elevation + I(Elevation^2)

  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        42     210.25
2        30      36.28 12   173.97 14.497 < 2.2e-16 ***

Terms added sequentially (first to last)

                   Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                                  42    210.246
Precipitation       1   41.361        41    168.885 41.3610 1.266e-10 ***
I(Precipitation^2)  1   24.410        40    144.475 24.4098 7.787e-07 ***
Geology             6   80.851        34     63.624 13.4752 2.383e-15 ***
Age.cat             2   17.732        32     45.892  8.8660 0.0001411 ***
Elevation           1    4.614        31     41.278  4.6143 0.0317067 *
I(Elevation^2)      1    4.998        30     36.280  4.9976 0.0253821 *

Single term deletions

                   Df Deviance     AIC F value     Pr(F)
<none>                  36.280 150.986
Precipitation       1   67.045 179.750 25.4393 2.059e-05 ***
I(Precipitation^2)  1   70.773 183.479 28.5223 8.902e-06 ***
Geology             6  106.457 209.163  9.6716 6.123e-06 ***
Age.cat             2   48.396 159.102  5.0095   0.01327 *
Elevation           1   39.930 152.636  3.0186   0.09257 .
I(Elevation^2)      1   41.278 153.983  4.1326   0.05100 .

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

two models. These criteria use different methods of penalizing extra deviance explained with extra explanatory variables. These are very similar to ANOVA in comparing whether the extra deviance explained by model 1 is significantly higher than the deviance explained by model 2. One of such criteria uses the AIC (or Akaike Information Criterion). A model with a lower AIC has a better combination of simplicity and explained deviance – provided that you agree with the way that simplicity and explained deviance are weighted by the AIC. The type-II ANOVA table provides the AIC for the most complex and all models with one deleted variable. We can see that the most complex model has the lowest AIC (150.986) of all the models, which suggests that all variables should be included in the model (although it is a probably worth again to remind you that we assume that only variables were measured for which there was a prior hypothesis that they could explain abundance).

Figure 6.12 plots the predicted abundance against precipitation. We can see in this figure that a complex pattern occurs, since the abundance is now regressed against all the explanatory variables. We therefore do not see the effect of precipitation only, but of the other variables as well. We can see that the observed abundances are predicted better.

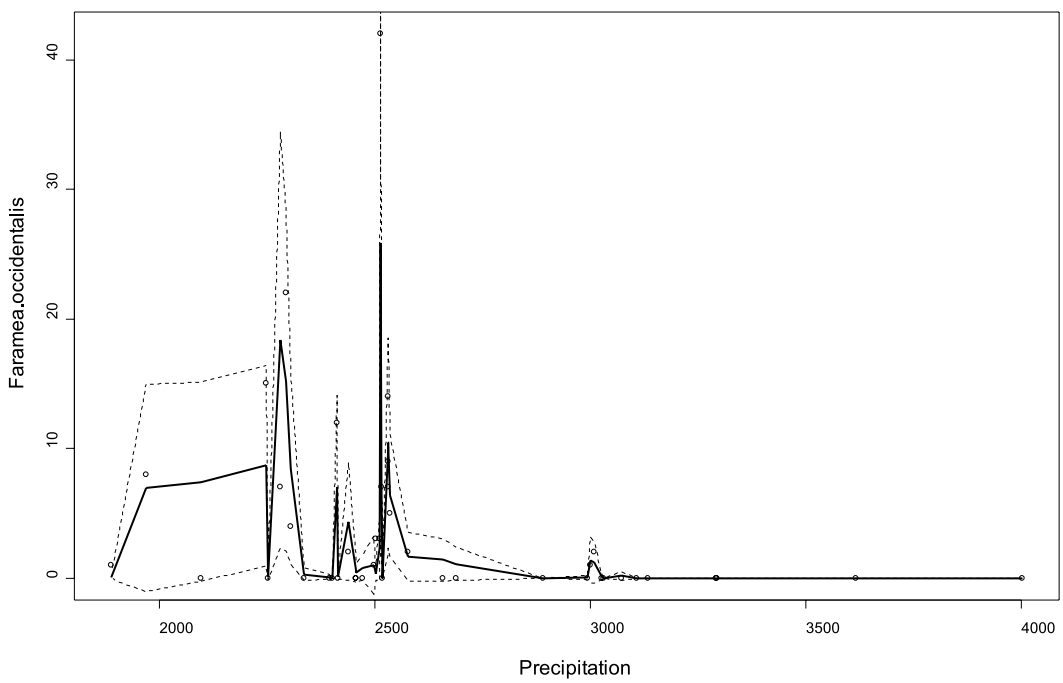


Figure 6.12 Observed values (circles) and predictions (lines) for the negative binomial GLM with log link of the abundance of *Faramea occidentalis* on geology, age category and the second-order polynomials of precipitation and elevation.

We can check the predictions for an explanatory variable in isolation by using a **termplot**. Figure 6.13 provides the termplot for age for the last model that we fitted. The full lines show the abundance that is predicted by the model coefficients. We can see that lower abundance is predicted for age categories 2 and 3, since the confidence intervals do not overlap with the confidence interval for age category 1. We can also see that the deviance that is explained by age is relatively small by comparing the difference between the predicted abundance and the actual abundance. Another feature of the data is that the lowest abundance is predicted for age category 2. Updating our model by using age as a continuous variable shows that there is no evidence that age has an effect – we can not assume that there is a straight line that will fit the effect of age on abundance. Allowing for different predicted abundances for each age category provides a better fit to our data.

Generalized Mixed Models

A lot more could be said about the models that we utilized already. The scope of this manual is too limited, however, to cover these models in more detail. For example, all the models here assume that the random residuals are uncorrelated. This might often be unreasonable, for example if a hierarchical sampling and measurement scheme were used. Mixed models have been developed for such situations (Quinn and Keough 2002).

The original dataset that we analysed here actually also had been sampled and measured in a hierarchical way (with 50 *B* sites), but we avoided the problem by only using the first and last sampled *B* site to remove the dominance of those sites in the dataset. We opted for this approach as this avoided the need for mixed models, and because the original dataset mixed data from different types of sample plots, something discussed in chapter 2. This type of practical approach to overcoming a

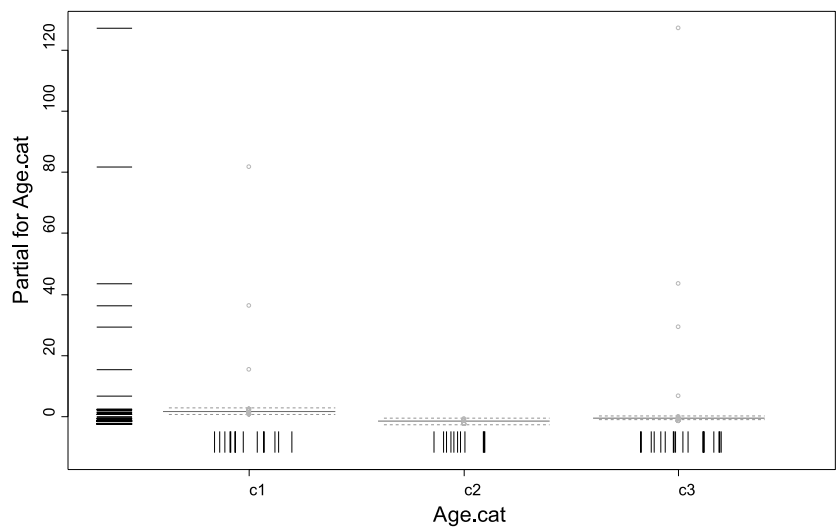


Figure 6.13 Observed values (circles) and predicted values (lines) for the termplot for age category for the negative binomial GLM with log link of the abundance of *Faramea occidentalis* on geology, age category and the second-order polynomials of precipitation and elevation. The rugplots show the distribution of the x and y values by adding a random term for values that are tied (for example, there are only three categories of age but the rugplots show the number of observations within each category).

statistical analysis problem is important. The final result may not be optimal if it does not use all available data, but it is clear and valid.

Choice of the best model

The most important criterion to guide you when you are given a choice between various models is that the assumptions of the models need to be realistic. We showed that the residuals of the models were used to check the reliability of the regression models, which guided us towards generalized linear models.

You may also favour models with a good balance between explanatory power and simplicity. Some tests (such as the AIC) may be used to help you in selecting the model with the best balance.

Analysing diversity

The examples that we provided in this chapter were for the number of trees of a particular species for each site. You can do the same analysis for the total number of species per site, or for the total number of trees per site. The methodology is exactly the same as the response variable is again calculated as a count of the number of objects found in each site, only that it is not the count of the trees of a single species but the count of the number of species or the total number of trees.

You could also perform the same calculations for a measure of diversity (see chapter on diversity) that was calculated for each site.

All these other calculations are possible, in principle. You will need to do diagnostic tests (as with the previous models) to check whether the assumptions of the model were met. In case that this is the case, then you can rely on the results.

References

Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, Foster RB, Nuñez P, Aguilar S, Valencia R, Villa G, Muller-Landau HC, Losos E and Hubbell SP. 2002. Beta-diversity in tropical forest trees. *Science* 295, 666–669. (dataset used as example)

Dalgaard P. 2002. *Introductory Statistics with R.* New York: Springer.

Fewster RM, Buckland ST, Siriwardena GM, Baillie SR and Wilson JD. 2000. Analysis of population trends for farmland birds using generalized additive models. *Ecology* 81: 1970-1984.

Fowler J, Cohen L and Jarvis P. 1998. *Practical statistics for field biology.* Chichester: John Wiley and sons.

Gotelli NJ and Ellison AM. 2004. *A primer of ecological statistics.* Sunderland: Sinauer Associates.

Hayek L-AC and Buzas MA. 1997. *Surveying natural populations.* New York: Columbia University Press.

Jongman RH, ter Braak CJF and Van Tongeren, OFR. 1995. *Data analysis in community and landscape ecology.* Cambridge: Cambridge University Press.

Kent M and Coker P. 1992. *Vegetation description and analysis: a practical approach.* London: Belhaven Press.

Legendre P and Legendre L. 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV.

Pyke CR, Condit R, Aguilar S and Lao S. 2001. Floristic composition across a climatic gradient in a neotropical lowland forest. *Journal of Vegetation Science* 12: 553-566. (dataset used as example)

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press. (recommended as first priority for reading)

Shaw PJA. 2003. *Multivariate statistics for the environmental sciences.* London: Hodder Arnold.

Shaw RG and Mitchell-Olds T. 1993. Anova for unbalanced data: an overview. *Ecology* 74: 1638-1645.

Underwood AJ. 1997. *Experiments in ecology: their logical design and interpretation using analysis of variance.* Cambridge: Cambridge University Press.

White GC and Bennets RE. 1996. Analysis of frequency count data using the negative binomial distribution. *Ecology* 77: 2549-2557.

Doing the analyses with the menu options of Biodiversity.R

Load the datasets Panama species.txt and Panama environmental.txt, and make them the species and environmental datasets, respectively. Give them the names "spec" and "faramea".

Data > Import data > from text file… (Panama species.txt)

→ Enter name for dataset: spec

Data > Import data > from text file… (Panama environmental.txt)

→ Enter name for dataset: faramea

Biodiversity > Community Matrix > Select community dataset…

→ Data set: spec

Biodiversity > Environmental Matrix > Select environmental dataset…

→ Data set: faramea

These are the original datasets, to use the reduced datasets that will be analysed, remove the sites where there is missing information on the variable "Analysed".

Biodiversity > Community matrix > Remove NA from environmental dataset…

→ Select variable: Analysed

As an alternative, load the dataset Faramea.txt, and make it both the species and environmental dataset (as both the species and environmental information is in the same dataset).

Data > Import data > from text file… (Faramea.txt)

→ Enter name for dataset: faramea

Biodiversity > Community Matrix > Select community dataset…

→ Data set: faramea

Biodiversity > Environmental Matrix > Select environmental dataset…

→ Data set: faramea

To calculate a linear regression model:

Biodiversity > Analysis of species as response > Species abundance as response…

→ Model options: linear model

→ Response: Faramea.occidentalis

→ Explanatory: Precipitation

→ print summary

→ print anova

→ Plot options: diagnostic plots

→ Plot variable: Precipitation

→ Plot options: diagnostic plots

→ Plot options: term plot

→ Plot options: effect plot

To calculate a generalized linear regression model (GLM):

Biodiversity > Analysis of species as response > Species abundance as response…
→ Model options: Poisson model
→ Response: Faramea.occidentalis
→ Explanatory: Precipitation
→ print summary
→ print anova
→ Model options: quasi-Poisson model
→ Model options: negative binomial model

To calculate a generalized additive regression model (GAM):

Biodiversity > Analysis of species as response > Species abundance as response…
→ Model options: gam model
→ Response: Faramea.occidentalis
→ Explanatory: s(Precipitation)
→ print summary

To calculate a multiple regression model:

Biodiversity > Analysis of species as response > Species abundance as response…
→ Model options: negative binomial model
→ Response: Faramea.occidentalis
→ Explanatory: Precipitation + I(Precipitation^2)
→ print summary
→ print anova

Doing the analyses with the command options of Biodiversity.R

Load the dataset Faramea.txt and give it the name "faramea".

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

To calculate a linear regression model:

```
Count.model1 <- lm(Faramea.occidentalis ~ Precipitation,
     data=faramea,  na.action=na.exclude)
summary(Count.model1)
fitted(Count.model1)
predict(Count.model1, interval='confidence')
residuals(Count.model1)
shapiro.test(residuals(Count.model1))
ks.test(residuals(Count.model1), pnorm)
anova(Count.model1,test='F')
Count.model2 <- lm(Faramea.occidentalis ~ Age.cat,
     data=faramea, na.action=na.exclude)
levene.test(residuals(Count.model2), na.omit(faramea)$Age.cat)
```

To plot a linear regression model:

```
plot(Count.model1)
termplot(Count.model1, se=T, partial.resid=T, rug=T,
     terms='Precipitation')
plot(effect('Precipitation', Count.model1))
```

To check for the spatial distribution of residuals:

```
surface.1 <- residuals.surface(Count.model1, na.omit(faramea),
     'UTM.EW', 'UTM.NS', gam=F, npol=1, plotit=T, bubble=F,
     fill=F)
surface.2 <- residuals.surface(Count.model1, na.omit(faramea),
     'UTM.EW', 'UTM.NS', gam=F, npol=2, plotit=T, bubble=F,
     fill=F)
surface.2 <- residuals.surface(Count.model1, na.omit(faramea),
     'UTM.EW', 'UTM.NS', gam=F, npol=2, plotit=T, bubble=T,
     fill=F)
surface.gam <- residuals.surface(Count.model1,
     na.omit(faramea), 'UTM.EW', 'UTM.NS', gam=T, npol=2,
     plotit=T, bubble=F, fill=T)
summary(surface.1)
anova(surface.1)
correlogram(surface.1, nint=10)
summary(surface.gam)
```

To calculate a generalized linear regression model (GLM):

```
Count.model3 <- glm(formula = Faramea.occidentalis ~
     Precipitation, family = poisson(),data=faramea,
     na.action=na.exclude)
summary(Count.model3)
anova(Count.model3,test='F')
predict(Count.model3, type='response', se.fit=T)
Count.model4 <- glm(formula = Faramea.occidentalis ~
     Precipitation, family = quasipoisson(), data=faramea,
     na.action=na.exclude)
Count.model5 <- glm.nb(Faramea.occidentalis ~ Precipitation,
     maxit = 5000, init.theta = 1, data=faramea, na.action=na.
     exclude)
```

To calculate a generalized additive regression model (GAM):

```
Count.model6 <- gam(Faramea.occidentalis ~ s(Precipitation),
     family=poisson(), data = na.omit(faramea))
summary(Count.model6)
predict(Count.model6, type='response', se.fit=T)
```

To calculate a multiple regression model:

```
Count.model7 <- glm.nb(Faramea.occidentalis ~ Precipitation
     + I(Precipitation^2), maxit = 5000, init.theta = 1,
     data=faramea, na.action=na.exclude)
summary(Count.model7)
anova(Count.model7, test='F')
Anova(Count.model7, type='II', test='Wald')
vif(lm(Faramea.occidentalis ~ Precipitation +
     I(Precipitation^2), data=faramea, na.action=na.exclude)
```

CHAPTER 7

Analysis of presence or absence of species

Analysis of presence or absence of species

In the previous chapter, we saw how species counts data can be analysed. In this chapter, we describe how data can be analysed that simply indicate whether a species is present in certain sites or absent. As for the analysis of species counts data, the data are analysed for one species at the time.

Analysis of presence or absence by cross-tabulations

As in the previous chapter, we will use a dataset that was collected in Panama, containing information on the abundance of *Faramea occidentalis*. This dataset also has observations for the environmental variables precipitation (quantitative), altitude (quantitative), age (ordinal) and geology (categorical). The dataset is provided in the previous chapter.

Imagine that you had a hypothesis that age had an influence on the chance that species *Faramea occidentalis* was present on a site. We treat age as a categorical variable – since it is an ordinal variable, we can choose whether we treat it as quantitative or categorical variable in subsequent analysis.

In a cross-tabulation analysis, you first need to count the number of sites of each category where the species occurs, and the number of sites where the species does not occur. You obtain these results by doing a cross-tabulation of species presence-absence with the age categories:

```
        c1 c2 c3
FALSE    5  6 12
TRUE     9  5  6
```

In the table, the rows indicate whether the species is absent (FALSE, based on the test whether the abundance > 0) or present (TRUE, based on the same test whether the abundance was > 0). The columns represent the three age categories. This table is a **cross-tabulation** or a **contingency table**. The cells in the table are counts of the number of observations within the specified categories of rows and columns.

We can also present these results graphically as in Figure 7.1.

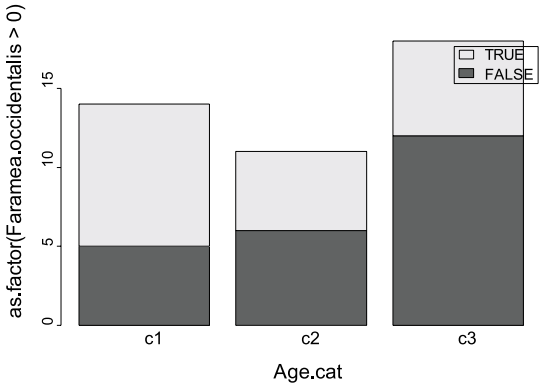


Figure 7.1 Observed frequencies for the presence and absence of *Faramea occidentalis* on three categories of age.

You can see that for sites of age category 1, the species occurs in 9 of 14 (14 = 5 +9) cases. Similarly, the species occurs in 5 of 11 sites of age category 2, and 6 of 18 sites of age category 3. For age category 1, we can use this information to calculate that the species has 9 / 14 × 100% = 64% chance of being present when the site is of age category 1. Using the same method, we can use the information to calculate a chance of 5 / 14 × 100% = 36% of being absent on sites of age category 1. From the table we can therefore calculate the chance that the species is present or absent on sites of a certain category, if sites are selected randomly. When we also calculate the chances for the other categories, we can calculate the following table:

Chances	Age category 1	Age category 2	Age category 3
Chance that the species is present (%)	64.3	45.5	33.3
Chance that the species is absent (%)	35.7	54.5	66.7

To find out whether the proportions are significantly different from each other, you can use a Chi-squared test. The result that you obtain will be:

```
        Pearson's Chi-squared test
data:  cross
X-squared = 3.0393, df = 2, p-value = 0.2188
```

This result shows that there is no evidence that differences in proportions exist between the three age categories. The significance level of P = 0.2188 indicates that there is a large chance that the differences in proportions are an effect from the random sampling of the sites from the survey area.

The Chi-squared test is limited in several ways. First of all, it is a test and is therefore not explicit in providing estimated or predicted values (the chances that were calculated earlier are not generated by the Chi-squared test). Secondly, the Chi-squared test can only be used to analyse the effect of a single categorical variable. Finally, the test is based on some assumptions that may not be reasonable.

The conditions for the Chi-squared test to be reliable are that the **expected frequencies** (expected when there is no relationship between the two variables that generated the crosstab) are not too small. How are the expected frequencies calculated and how do we evaluate whether some expected values are too small? The expected frequencies are calculated by multiplying the chance that the species is present or absent by using frequencies from the entire dataset with the number of sites of each age category. For the entire dataset of 43 sites, species presence was observed in 20 (=9 + 5 + 6) sites or 46.5% (20 / 43 × 100%) of all sites. The number of sites that are expected to contain the species for age category 1, when there is no relationship between age and presence-absence, is therefore 0.465 × 14 (=9 + 5) = 6.51. The expected frequencies can be calculated for each combination of presence-absence and age category as:

```
              c1       c2       c3
 FALSE 7.488372 5.883721 9.627907
 TRUE  6.511628 5.116279 8.372093
```

The condition for the Chi-squared test to provide reliable results is that all the expected frequencies should be larger than 5 (a less strict condition is that more than 20% of expected frequencies should be larger than 5, but none should be smaller than 1). Since all expected frequencies are larger than 5, we can rely on the Chi-squared test to have provided a reliable result. The Chi-squared test actually estimates the probability that the measured and expected frequencies will be the same for the survey area.

Another criticism of the Chi-squared test is that it ignores the ordering of the age categories – the table above showed that there is a trend of decreasing chance of encountering the species when the age of the plot is greater, but the Chi-squared test does not investigate the ordering of the age categories.

Analysis of presence or absence through binomial GLM

We can investigate the same hypothesis that there is an influence of age category on the presence-absence of *Faramea occidentalis* with a binomial GLM with logit link.

As we saw in the previous chapter, a GLM is defined by the variance function and the link function. When we define the variance and link functions, we assume that these functions are realistic descriptions for the dataset that we are investigating.

The logit link is defined as:

$$logit(\mu) = \log(\mu / (1 - \mu)) = a + b_1 \times x_1 + b_2 \times x_2 + b_3 \times x_3 + \ldots$$

The logit link function is one way of guaranteeing that the predicted values will be between 0 and 1, which is appropriate since we want to predict probabilities of presence of *Faramea occidentalis* which also are between 0 and 1. In the previous chapter where we analysed counts, we used a log link that ensured that the predicted values were larger than 0, but not that the predicted values were smaller than 1.

By investigating the hypothesis with a GLM, we overcome some of the shortcomings of the Chi-squared test: the GLM will provide predictions, and several explanatory variables can be analysed – including quantitative variables. Not all shortcomings of the Chi-squared test are overcome, however: the large sample assumptions are still needed for the tests of the GLM to provide realistic results.

The binomial GLM with logit link investigating the influence of age on presence-absence yields the following results:

```
glm(formula = Faramea.occidentalis > 0 ~ Age.cat, family = binomial(link = logit),
    data = faramea, na.action = na.exclude)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5878     0.5578   1.054   0.2920
Age.catc2    -0.7701     0.8233  -0.935   0.3496
Age.catc3    -1.2809     0.7491  -1.710   0.0873 .
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 56.322  on 40  degrees of freedom
AIC: 62.322

Analysis of Deviance Table
Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       42     59.401
Age.cat  2    3.079        40     56.322     0.214
```

The results of the GLM first show the coefficients that were calculated. As for the analysis of counts data, the first category is not included explicitly in the results. The results for the first category correspond to the intercept, however.

As we saw in the previous chapter, it is a bit complicated to directly calculate the expected values from the estimations of the coefficients. The reason is that the inverse link function needs to be calculated to obtain the expected values. In the case of the logit link, the inverse logit is calculated as $y = exp(x)/(1+exp(x))$. However, the program that fits the model should be able to provide the predicted values. Here we obtain following predictions for the three categories: 0.64, 0.45 and 0.33. You could calculate these results yourself by using the inverse logit as in the box below.

You can see that you obtain the same predictions with the GLM as the chances for presence of the species that we calculated earlier. For example, the chance of presence for sites with age category 2 is calculated in both instances to be 45.5%.

The large significance level values estimated for the regression coefficients indicate that there is no evidence for significant differences between the predicted chances, however. The ANOVA table provides similar information by estimating a large significance level ($P = 0.21$). The ANOVA table also provides important information on the deviance that is explained: the model only explains 3.079 or 5.2% of total or null deviance (59.401).

The Chi-squared test of the ANOVA table is a test for the same pattern that was tested at the beginning of this chapter by the Chi-squared test for the contingency table. You could check that the explained deviance (3.079) is very similar to the Chi-squared statistic (3.039) – calculating the deviance is actually another method for analysing a cross-tabulation (called a G-test in some manuals). What is important is that both the ANOVA and the earlier Chi-squared test estimate a similar significance level ($P = 0.214$ and $P = 0.2145$), leading to the same conclusion of no evidence for an influence of age on the presence-absence. You can also see the limitations of the Chi-squared test – it is as if you ignore all other results from the GLM.

The results from the model can also be analysed graphically as in Figure 7.2. Because the observed values are either 0 (absence) or 1 (presence), the figure is not very informative and mainly shows the predicted chances for presence of the species for each age category. The wide overlap between the 95% confidence intervals shows that there is no evidence for an effect of age.

As with all regression models, the binomial GLM makes various assumptions about the data. Whether you can trust the results depends on whether these assumptions are realistic.

The quasi-binomial model was developed for situations where one of the assumptions of the binomial model does not hold – that the dispersion equals 1. As we saw in the previous chapter, for our type of data the dispersion parameter is an indication of how randomly individuals are distributed. A dispersion parameter of 1 means

```
Expected values:

Age category x: inverse logit(intercept + coefficient for age category x)

Age category 1: inverse logit(0.5878+0) =
     exp(0.5878+0) / (1 + exp (0.5878+0) ) = 0.6428602

Age category 2: inverse logit(0.5878-0.7701) =
     exp (0.5878-0.7701) / (1 + exp (0.5878-0.7701) ) = 0.4545508

Age category 3: inverse logit(0.5878-1.2809) =
     exp (0.5878-1.2809) / (1 + exp (0.5878-1.2809) ) = 0.3333438
```

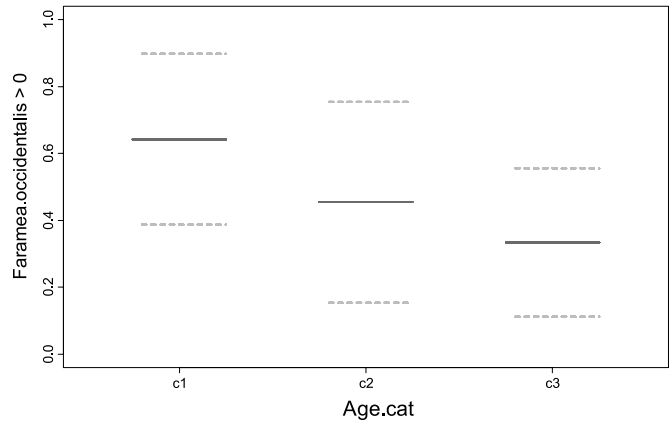


Figure 7.2 Predicted values (horizontal lines) for the binomial GLM with logit link of the presence-absence of *Faramea occidentalis* on age category. Dashed lines are 95% confidence intervals for the mean.

that the individuals are randomly distributed over the sample units. When dispersion is not 1, the GLM will estimate significance levels that are not realistic. A quasi-binomial GLM will estimate the same regression coefficients as the binomial GLM, but will estimate the dispersion parameter and use this dispersion parameter to provide different estimates of standard errors and significance levels.

The quasi-binomial GLM with logit link investigating the influence of age on presence-absence yields the following results:

```
glm(formula = Faramea.occidentalis > 0 ~ Age.cat, family = quasibinomial(link = logit),
    data = faramea, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4350  -1.0008  -0.9005   1.0979   1.4823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5878     0.5783   1.016    0.316
Age.catc2    -0.7701     0.8536  -0.902    0.372
Age.catc3    -1.2809     0.7767  -1.649    0.107

(Dispersion parameter for quasibinomial family taken to be 1.075000)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 56.322  on 40  degrees of freedom
AIC: NA


Analysis of Deviance Table

        Df Deviance Resid. Df Resid. Dev      F Pr(>F)
NULL                     42     59.401
Age.cat  2    3.079        40     56.322 1.4322 0.2508
```

When we compare the results from the quasi-binomial GLM with the binomial GLM, we can see that the quasi-binomial model estimated the dispersion parameter to be 1.075. Since there is a small difference between 1.075 and 1, there will not be any substantive difference in conclusions whether the dispersion is fixed as 1 as in the binomial, or estimated as in the quasibinomial. It is possible to test whether the estimated dispersion is different from 1, and perhaps infer something on the clumpiness of the distribution on the basis of the results. However, such tests have the usual problems: they depend on the validity of the models assumed, and the results depend on both the sample size and the dispersion. Probably a better approach is to first think whether overdispersion is likely given the source of the data, and pay careful attention when it is. Then check whether the results differ in substance between the two models. If interest is really into the extent to which the distribution is clumped, regular or random, then there are better methods which focus on this aspect.

Since the dispersion parameter was estimated to be very close to one, only small differences in the estimated significance levels can be observed (for instance $P = 0.25$ in the ANOVA instead of $P = 0.21$) and both models lead to the same conclusions. Note that it is a good practice to check for the difference between the results of the binomial and the quasi-binomial GLM. A disadvantage of the quasi-binomial GLM implemented here is that it does not calculate the Akaike Information Criterion (AIC), which can be used for model selection (see previous chapter and below: binomial GLM with several explanatory variables). An advantage of the quasi-binomial GLM is that it provides better estimates of significance level when the dataset has a dispersion parameter that is more different from 1.

Binomial and quasi-binomial GLM with continuous variables

We have seen so far that analysis of a binomial or quasi-binomial GLM with logit link provides a more comprehensive analysis of frequencies than the Chi-squared test (which is only a test) and that estimation is explicit in the GLM. Another advantage of the GLM approach is that the explanatory variables can either by continuous or categorical.

To analyse a cross tabulation with continuous explanatory variables with a Chi-squared test, you need to derive a categorical variable from the continuous variable. You need to define several categories, for instance a category for altitude < 200 m and another category for altitude > 200 m. These categories need to be defined by the researcher, and there is no procedure that can tell you how these categories should be defined. With the binomial GLM, you do not need to make this decision. More importantly, when the relationship between the explanatory variable and the response variable is not a stepwise function but a gradual change, it is better to model the gradual change. The binomial GLM will accommodate gradual changes for a continuous variable, which often provides a more realistic model.

We can for example model the presence and absence of *Faramea occidentalis* based on the precipitation of a site with a quasi-binomial GLM with logit link. The results that you will obtain are shown in the box on the next page.

You could notice that we used a quasi-binomial model. The dispersion parameter is estimated to be 0.987, very close to 1. Both the binomial and the quasi-binomial GLM therefore lead to the same conclusions.

As in the regression results that we saw earlier, regression coefficients are provided for the explanatory variable. We can see that there is evidence that precipitation has an effect, since the significance level calculated for the coefficient is low ($P = 0.0172$). We can also infer that there is an effect of precipitation from the low significance level of the ANOVA table ($P = 0.005$).

```
glm(formula = Faramea.occidentalis > 0 ~ Precipitation, family = quasibinomial(link =
logit), data = faramea, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7303  -1.0431  -0.3289   1.1157   1.7268

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.948352   2.828347   2.457   0.0183 *
Precipitation -0.002721   0.001095  -2.484   0.0172 *

(Dispersion parameter for quasibinomial family taken to be 0.9878437)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 50.561  on 41  degrees of freedom
AIC: NA


Analysis of Deviance Table

              Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                              42     59.401
Precipitation  1    8.841        41     50.561 8.9494 0.004682 **

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

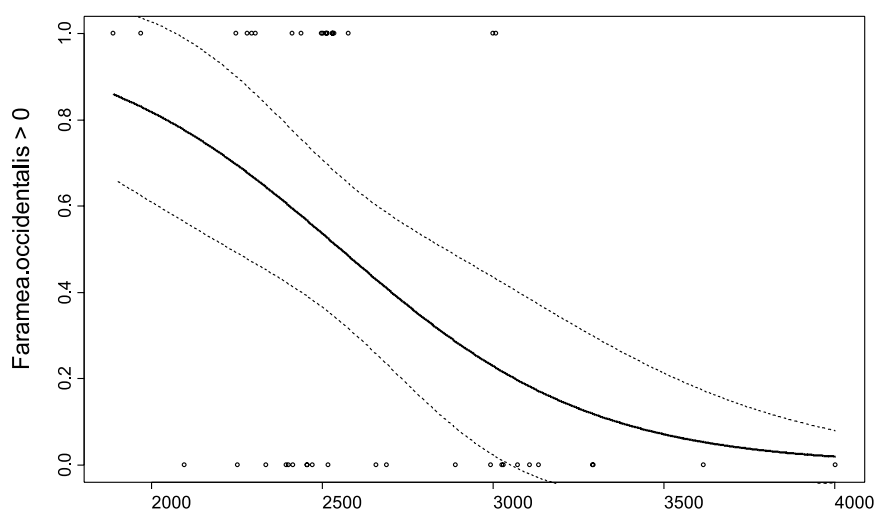


Figure 7.3 Observed values (circles) and predicted values (connected by line) for the quasi-binomial GLM model of the presence-absence of *Faramea occidentalis* on elevation. Dashed lines are 95% confidence intervals for the mean.

The predictions of the model can be shown graphically as well. Figure 7.3 provides the observed and the predicted values.

You can see the merit of the logit link in not predicting values outside of the interval with 0 and 1 as boundaries. By using the logit link, the GLM model does not predict values that we do not expect. The plot of the observations is more informative for a continuous than for a categorical variable: you could observe now that the species was never observed at the highest precipitation levels, and that it was never not observed at the lowest precipitation levels. At intermediate levels, the species was sometimes observed. This pattern is modelled as an S-shaped curve, which provides a reasonable fit to the data.

Another advantage of the binomial and quasi-binomial GLM with logit link is that several explanatory variables can be investigated. An example is provided later in this chapter.

Binomial GAM with several explanatory variables

As for the regression models for count data, you can also fit a Generalized Additive Model (GAM) using the same variance and link functions that are used in a GLM. The GAM allows estimation of a smooth relationship between the response and a quantitative explanatory variable. The smoothing function will generate a curve that can flow more freely in between the data than a straight line.

When we calculate a quasi-binomial GAM with logit link for the presence and absence of *Faramea occidentalis* using smoothing functions of precipitation and elevation, and the categorical variables geology and age category, then we obtain the following results:

```
Family: quasibinomial
Link function: logit

Formula:
Faramea.occidentalis > 0 ~ s(Precipitation) + Geology + Age.cat +
    s(Elevation)

Parametric coefficients:
              Estimate  std. err.    t ratio    Pr(>|t|)
(Intercept)    -382.04      8.204     -46.57    < 2.22e-16
  GeologyTb     30.746      314.9    0.09764    0.92287
 GeologyTbo     14.571  2.499e+04   0.000583    0.99954
  GeologyTc     891.93       20.1      44.37    < 2.22e-16
 GeologyTcm       15.3       3795   0.004032    0.9968
 GeologyTgo     7.3746      1.204      6.126    9.6953e-07
  GeologyTl     137.55  1.114e+04    0.01234    0.99023
  Age.catc2    -103.74       1434   -0.07235    0.9428
  Age.catc3    -88.072      2.189     -40.23    < 2.22e-16

R-sq.(adj) =      1   Deviance explained =  100%
GCV score = 5.0022e-06   Scale est. = 3.4996e-06  n = 43

Analysis of deviance table

Parametric Terms:
               df       chi.sq     p-value
Geology         6       1986.4     < 2.22e-16
Age.cat         2       1618.3     < 2.22e-16

Approximate significance of smooth terms:
                       edf       chi.sq     p-value
s(Precipitation)     2.917       2252.5     < 2.22e-16
    s(Elevation)         1       1499.4     < 2.22e-16
```

The quasi-binomial GAM with logit link indicates that all explanatory variables contribute to explaining the deviance in the presence-absence of the species. The `edf` indicate the estimated degrees of freedom for the smoothing functions. The fact that 3 degrees of freedom are estimated for precipitation shows that a complex pattern was modelled for this explanatory variable.

Since our dataset was quite small, all deviance was explained, and the significance level values that were estimated were ridiculously small, we need to treat the results with caution. We expect that the model overfitted the data, although it is hard to see with presence-absence data and several variables. We therefore opted to analyse the dataset further with a GLM in the next section to find out whether there was further evidence for the complex pattern between presence-absence and the quantitative variables. An alternative approach would have been to fix the degrees of freedom for the smoothing terms in a GAM, forcing the curve to be smoother and not have such a complex shape. The results with 2 degrees of freedom for precipitation and elevation look more reasonable:

```
Family: quasibinomial
Link function: logit

Formula:
Faramea.occidentalis > 0 ~ s(Precipitation, k = 2, fx = T) +
    Geology + Age.cat + s(Elevation, k = 2, fx = T)

Parametric coefficients:
              Estimate  std. err.    t ratio    Pr(>|t|)
(Intercept)    -77.189      39.05     -1.976    0.057364
  GeologyTb     1.4079      1.286      1.094    0.28250
 GeologyTbo     31.498       1023    0.03078    0.97565
  GeologyTc     24.648       9.48        2.6    0.014332
 GeologyTcm     28.945      396.8    0.07294    0.94234
 GeologyTgo    -2.0587       2.03     -1.014    0.3187
  GeologyTl     13.802      634.6    0.02175    0.9828
  Age.catc2      -16.9      88.73    -0.1905    0.85022
  Age.catc3    -4.1095       1.93     -2.129    0.041593

R-sq.(adj) =  0.695   Deviance explained = 75.8%
GCV score = 0.70325   Scale est. = 0.49064   n = 43


Analysis of deviance table

Parametric Terms:
               df       chi.sq     p-value
Geology         6       8.8875     0.21822
Age.cat         2       4.5391     0.12083

Approximate significance of smooth terms:
                       edf       chi.sq     p-value
s(Precipitation)         2       6.2923     0.057491
    s(Elevation)         2       3.8408     0.16414
```

Binomial GLM with several explanatory variables

As already shown for the GAM, it is possible to analyse several explanatory variables together in a GLM – this was also shown in the previous chapter.

Based on the results of the quasi-binomial GAM with logit link that showed curved relationships between elevation, precipitation and the presence-absence of *Faramea occidentalis*, we fitted a second-order polynomial model for the quantitative variables (see previous chapter). The results of the binomial GLM with logit link of the presence-absence of *Faramea occidentalis* on geology, age category and the second-order polynomials of precipitation and elevation are:

```
glm(formula = Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Geology + Age.cat + Elevation + I(Elevation^2), family = binomial(link = logit),
    data = faramea, na.action = na.exclude)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.031e+00  -2.753e-02  -2.107e-08   8.387e-02   1.977e+00

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.137e+02  8.737e+01  -1.302    0.193
Precipitation       1.006e-01  7.594e-02   1.324    0.185
I(Precipitation^2) -2.223e-05  1.644e-05  -1.352    0.176
GeologyTb           1.401e+00  1.718e+00   0.815    0.415
GeologyTbo          2.962e+01  1.075e+04   0.003    0.998
GeologyTc           1.266e+01  8.055e+00   1.572    0.116
GeologyTcm          2.642e+01  4.153e+03   0.006    0.995
GeologyTgo         -1.270e+00  2.709e+00  -0.469    0.639
GeologyTl           1.792e+01  7.274e+03   0.002    0.998
Age.catc2          -9.695e+00  1.024e+01  -0.946    0.344
Age.catc3          -2.983e+00  2.458e+00  -1.214    0.225
Elevation           8.701e-02  1.161e-01   0.749    0.454
I(Elevation^2)     -4.493e-04  5.293e-04  -0.849    0.396

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 17.376  on 30  degrees of freedom
AIC: 43.376

Analysis of Deviance Table

Model 1: Faramea.occidentalis > 0 ~ 1
Model 2: Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Geology + Age.cat + Elevation + I(Elevation^2)
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        42     59.401
2        30     17.376 12   42.025 3.5021 3.298e-05 ***
```

```
Terms added sequentially (first to last)

                   Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                  42     59.401
Precipitation       1    8.841        41     50.561     0.003
I(Precipitation^2)  1    0.722        40     49.838     0.395
Geology             6   21.144        34     28.694     0.002
Age.cat             2    7.605        32     21.089     0.022
Elevation           1    2.837        31     18.252     0.092
I(Elevation^2)      1    0.876        30     17.376     0.349

Single term deletions

                   Df Deviance    AIC    LRT   Pr(Chi)
<none>                  17.376 43.376
Precipitation       1   20.871 44.871  3.495 0.0615481 .
I(Precipitation^2)  1   21.441 45.441  4.065 0.0437749 *
Geology             6   40.674 54.674 23.298 0.0007025 ***
Age.cat             2   24.799 46.799  7.423 0.0244364 *
Elevation           1   18.020 42.020  0.645 0.4220785
I(Elevation^2)      1   18.252 42.252  0.876 0.3492853
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

The main reason for having conducted the analysis is to select variables that meaningfully contribute to explaining the deviance, with special focus on the second-order polynomial terms of precipitation^2 and elevation^2. We saw in the previous chapter that one of the criteria that can be used for selecting variables is the Akaike Information Criterion (AIC). Models with a smaller AIC are preferred over models with larger AIC. The type-II ANOVA provides a column with the AIC. We can see that the AIC for elevation^2 is smaller than the model where this variable is included (42.252 < 43.376). Based on these results, a model where elevation^2 is not included will provide a better combination of simplicity (fewer variables) and explained deviance, provided that the way that the AIC calculates the combination is the best way. The analysis of the AIC indicates what caused our problem with the first GAM results: the GAM sacrificed simplicity to explain all the deviance – this is not necessarily the best model. Because of the smaller AIC, we excluded elevation^2 from further analysis. The results of type-II ANOVA for the quasi-binomial GLM with logit link of the presence-absence of *Faramea occidentalis* on geology, age category, elevation and the second-order polynomial of precipitation now indicated that all variables could be kept in the model (see next page):

```
glm(formula = Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Age.cat + Geology + Elevation, family = quasibinomial(link = logit),
    data = faramea, na.action = na.exclude)


Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.253e+00  -6.919e-02  -2.107e-08   1.639e-01   1.839e+00


Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -8.178e+01  5.183e+01  -1.578   0.1247
Precipitation       7.501e-02  4.577e-02   1.639   0.1114
I(Precipitation^2) -1.655e-05  9.835e-06  -1.683   0.1024
Age.catc2          -8.156e+00  5.234e+00  -1.558   0.1293
Age.catc3          -1.890e+00  1.363e+00  -1.387   0.1753
GeologyTb           1.846e+00  1.324e+00   1.395   0.1730
GeologyTbo          2.551e+01  9.106e+03   0.003   0.9978
GeologyTc           9.729e+00  4.800e+00   2.027   0.0513 .
GeologyTcm          2.531e+01  3.519e+03   0.007   0.9943
GeologyTgo         -3.601e-01  1.696e+00  -0.212   0.8333
GeologyTl           1.839e+01  6.435e+03   0.003   0.9977
Elevation          -1.179e-02  1.330e-02  -0.887   0.3820
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for quasibinomial family taken to be 0.7169455)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 18.252  on 31  degrees of freedom
AIC: NA


Analysis of Deviance Table

Model 1: Faramea.occidentalis > 0 ~ 1
Model 2: Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Age.cat + Geology + Elevation
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        42     59.401
2        31     18.252 11   41.149 5.2178 0.0001320 ***

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                                   42     59.401
Precipitation       1    8.841        41     50.561 12.3310 0.0013893 **
I(Precipitation^2)  1    0.722        40     49.838  1.0076 0.3232568
Age.cat             2    0.991        38     48.847  0.6911 0.5085612
Geology             6   27.758        32     21.089  6.4528 0.0001718 ***
Elevation           1    2.837        31     18.252  3.9577 0.0555415 .
```

```
Single term deletions
                  Df Deviance F value     Pr(F)
<none>                 18.252
Precipitation      1   21.469  5.4648 0.0260342 *
I(Precipitation^2) 1   21.969  6.3138 0.0173944 *
Age.cat            2   28.606  8.7932 0.0009443 ***
Geology            6   40.748  6.3681 0.0001904 ***
Elevation          1   21.089  4.8193 0.0357555 *

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

We showed the results for the quasi-binomial model as this model estimated dispersion to be 0.71. As we saw before, the regression coefficients and deviances of the ANOVA tables are the same for the binomial and quasi-binomial models, but the estimated significance levels will be different.

One important pattern that you can observe in the results is that none of the significance values for the regression coefficients are small. Only the coefficient for the category of geology *Tc* has a significance level (P=0.0513) that is smallish. Despite the fact that there is no evidence from the significance levels of the regression coefficients, the model explains most of the deviance. Moreover, we used the AIC to select those variables that meaningfully contributed to explaining the deviance.

What is going on? The reason for the difference between the regression and ANOVA results is that we have few observations for most categories of geology. This pattern is depicted in Figure 7.4. As we saw at the beginning of this chapter with analyses for cross-tabulations, we will not get reliable results when the number of observations is very small for one category (technically, when the expected frequencies are very small). For category *Tbo* there was only one observation, whereas only two categories had more than 5 observations. Moreover, most categories are dominated either by presence or absence, which made it easier to obtain correct predictions. For example, by predicting for *Tbo*, *Tcm* and *Tl* that the species is present, all predictions will be correct. One of the problems with these data is that the model does not allow that predicted values are exactly 1 or exactly 0. Another problem is that since we only have 1 observation for *Tbo*, 2 observations for *Tl* and 5 observations for *Tcm*, we do not have any information to infer that the same predictions will be valid for the entire survey area – the sample size is simply too small. The sample size is even too small for *pT*, although it is dominated by sites where the species is absent.

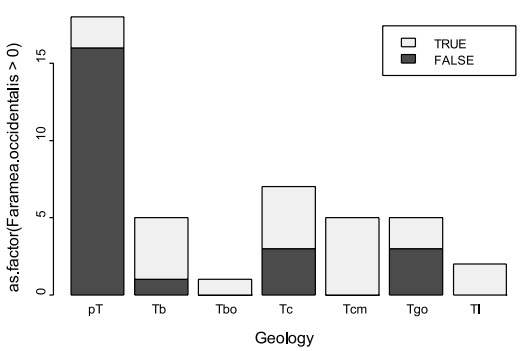


Figure 7.4 Observations of the presence-absence of *Faramea occidentalis* categorized by geology.

The fact that the small sample sizes and the similar observations within the categories of geology lead to a majority of correct predictions (and also a large amount of explained deviance since predicted and observed probabilities for presence were close), the other explanatory variables only needed to explain the odd cases such as the two sites with presence of the species on geology *pT*. This implies that sample size was also small for the other explanatory variables, so that significance levels for the regression coefficients were large.

In conclusion, we saw that the various models that we fitted explained most of the deviance, but that we do not have sufficient information to infer that similar patterns would be observed for the entire survey area. The sample size is simply too small. One possible solution could be to collapse some categories of geology into a smaller number of categories. This can only meaningfully be done if there is a logical method for combining the various categories. To further analyse the influence of geology on presence-absence with the present categories, we need to add new sites to the dataset.

Always think whether your analysis objective is realistic given the data you have. With a small number of presence-absence observations, can you expect to be able to detect and estimate the effects of 7 geology types as well as the complex, curved relationships with precipitation and elevation?

Choice of the best model

As seen at the end of the previous chapter, the most important criterion that you could use to choose between different models is the level to which the assumptions of the models are realistic. We investigated the residuals of the models that were used for count data to check the reliability of the regression models. With presence-absence data, the residuals are more difficult to investigate given that the observations were either 0 or 1. There is actually no standard method for investigating the residuals for presence-absence data.

You may also favour models with a good balance between explanatory power and simplicity. Some tests (such as the AIC) may be used to help you in selecting the model with the best balance.

However, it is not possible to provide tests or a procedure that will always select the one and only best model. The ecologist or biodiversity scientist needs to check (partially on non-statistical grounds) whether a particular model provides the best answer for the research hypothesis. The model with the largest AIC is not always the best in explaining a particular pattern. For example, when results of previous research efforts are also considered, a model with a slightly smaller AIC may be judged to be better for the particular study. All statistical models are approximations and simplifications of ecological patterns, rather than 'correct' descriptions of biological processes. So the purpose of modeling has to be considered along with statistical evidence when choosing between alternatives. For example, if average rate of response to a continuous explanatory variable is needed, a straight line model may be appropriate even if a curvature is statistically significant.

By having analysed the same dataset using the species counts as response variable in the previous chapter and transforming the counts to presence-absence in this chapter, we saw that the analysis lead to different conclusions. Since a different response variable was used, this was not a complete surprise. The different response variable was a transformation of the other response variable, however. The different results therefore showed that the transformation had an important effect, and that a different pattern prevailed. This simply means that count and presence-absence data do not necessarily follow the same pattern. If you are interested in investigating both patterns, then you need to record species counts and not simply presence or absence. Again this should follow from the initial hypotheses that you had.

References

Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, Foster RB, Nuñez P, Aguilar S, Valencia R, Villa G, Muller-Landau HC, Losos E and Hubbell SP. 2002. Beta-diversity in tropical forest trees. *Science* 295: 666–669. (dataset used as example)

Fowler J, Cohen L and Jarvis P. 1998. *Practical statistics for field biology.* Chichester: John Wiley and sons.

Hastie TJ and Pregibon D. 1993. Generalised Linear Models. In: Chambers, JM and Hastie TJ. *Statistical models in S.* London: Chapman and Hall.

Jongman RH, ter Braak CJF and Van Tongeren OFR. 1995. *Data analysis in community and landscape ecology.* Cambridge: Cambridge University Press.

Legendre P and Legendre L. 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV.

Pyke CR, Condit R, Aguilar S and Lao S. 2001. Floristic composition across a climatic gradient in a neotropical lowland forest. *Journal of Vegetation Science* 12: 553-566. (dataset used as example)

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press. (recommended as first priority for reading)

Doing the analyses with the menu options of Biodiversity.R

Load the datasets Panama species.txt and Panama environmental.txt, and make them the species and environmental datasets, respectively. Give them the names "spec" and "faramea".

Data > Import data > from text file… (Panama species.txt)

→Enter name for data set: spec

Data > Import data > from text file… (Panama environmental.txt)

→Enter name for data set: faramea

Biodiversity > Community Matrix > Select community data set…

→Data set: spec

Biodiversity > Environmental Matrix > Select environmental data set…

→Data set: faramea

These are the original datasets, to use the reduced datasets that will be analysed, remove the sites where there is missing information on the variable "Analysed".

Biodiversity > Community matrix > Remove NA from environmental data set…

→Select variable: Analysed

As an alternative, load the dataset Faramea.txt, and make it both the species and environmental dataset (as both the species and environmental information is in the same dataset).

Data > Import data > from text file… (Faramea.txt)

→Enter name for data set: faramea

Biodiversity > Community Matrix > Select community data set…

→Data set: faramea

Biodiversity > Environmental Matrix > Select environmental data set…

→Data set: faramea

To analyse presence or absence by cross-tabs:

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: crosstab

→Response: Faramea.occidentalis

→Explanatory: Age.cat

To calculate a generalized linear regression model (GLM):

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: binomial model

→Response: Faramea.occidentalis

→Explanatory: Age.cat

→print summary

→print anova

→Plot options: diagnostic plots

→Plot variable: Age.cat

→Plot options: term plot

→Plot options: effect plot

→Model options: quasi-binomial model

To calculate a generalized additive regression model (GAM):

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: binomial model

→Response: Faramea.occidentalis

→Explanatory: s(Precipitation) + Geology + Age.cat + s(Elevation)

→print summary

To calculate a GLM with several explanatory variables:

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: binomial model

→Response: Faramea.occidentalis

→Explanatory: Precipitation + I(Precipitation^2) + Geology + Age.cat + Elevation + I(Elevation^2)

→print summary

→print anova

Doing the analyses with the command options of Biodiversity.R

Load the datasets Condit species.txt and Condit environmental.txt. Give them the names "spec" and "faramea". Alternatively, load the dataset Faramea.txt and give it the name "faramea".

```
spec <- read.table(file="D://my files/Condit species.txt")
attach(spec)
faramea <- read.table(file="D://my files/Condit environmental.
  txt")
faramea <- read.table(file="D://my files/Faramea.txt")
attach(faramea)
```

To analyse presence or absence by cross-tabs:

```
faramea$Faramea.occidentalis<<- spec$Faramea.occidentalis
table1 <- table(Faramea.occidentalis>0, Age.cat)
Presabs.1 <- chisq.test(table1)
Presabs.1
Presabs.1$observed
Presabs.1$expected
```

To calculate a generalized linear regression model (GLM):

```
Presabs.model2 <- glm(formula = Faramea.occidentalis>0 ~
  Age.cat, family = binomial(link=logit), data = faramea,
  na.action = na.exclude)
summary(Presabs.model2)
anova(Presabs.model2,test='F')
predict(Presabs.model2, type='response', se.fit=T)
null.model <- glm(formula = Faramea.occidentalis>0 ~ 1,
  family = binomial(link=logit) , data = faramea, na.action =
  na.exclude)
anova(null.model, Presabs.model2, test='Chi')
plot(Presabs.model2)
termplot(Presabs.model2, se=T, partial.resid=T, rug=T,
  terms='Age.cat')
plot(effect('Age.cat', Presabs.model2))
Presabs.model3 <- glm(formula = Faramea.occidentalis>0 ~ Age.
  cat, family = quasibinomial(link=logit) , data = faramea,
  na.action = na.exclude)
Presabs.model4 <- glm(formula = Faramea.occidentalis>0 ~
  Elevation, family = quasibinomial(link=logit) , data =
  faramea, na.action = na.exclude)
```

To calculate a generalized additive regression model (GAM):

```
Presabs.model5 <- gam(formula = Faramea.occidentalis>0 ~
  s(Precipitation) + Geology + Age.cat + s(Elevation), family
  = quasibinomial(link=logit) , data = faramea, na.action =
  na.exclude)
summary(Presabs.model5)
```

To calculate a GLM with several explanatory variables:

```
Presabs.model6 <- glm(formula = Faramea.occidentalis > 0 ~
  Precipitation + I(Precipitation^2) + Geology + Age.cat +
  Elevation + I(Elevation^2), family = binomial(link = logit)
  , data = faramea, na.action = na.exclude)
summary(Presabs.model6)
anova(Presabs.model6,test='Chi')
drop1(Presabs.model6, test='Chi')
```

CHAPTER 8

Analysis of differences in species composition

Analysis of differences in species composition

This chapter describes how the difference in species composition can be investigated by calculating the ecological distance between two sites. The methods can be applied to an entire species matrix by calculating ecological distances between all pairs of sites.

The results can be presented as a distance or dissimilarity matrix. In later chapters on clustering and ordination, some methods are shown for analysing such distance matrices.

How can differences in species composition be investigated?

This chapter describes some methods for measuring the difference in species composition between two sites. Rather than stating the difference in abundance between each and every species, a single statistic is calculated that expresses the difference in species composition. Consider Figure 8.1 as an example. You could report that site A and B share the same species S1, S2 and S3 with the same abundance, whereas species S4 and S5 only occur on site B. An ecological distance will summarize these differences in a single distance statistic. In the case of the Bray-Curtis distance (there are many different methods of calculating a distance), you would calculate that the difference between site A and B is 0.25.

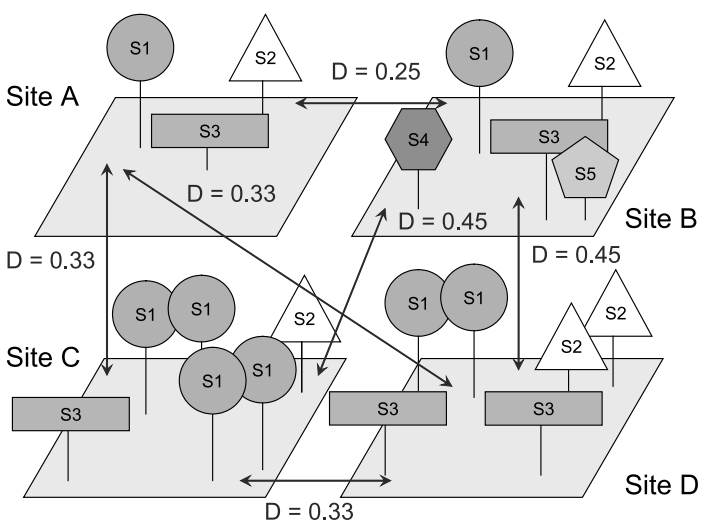


Figure 8.1 Four sites with different species composition. The differences in species composition between each subset of two sites can be expressed by a single ecological distance such as the Bray-Curtis distance. The Bray-Curtis distance between A and B is 0.25, and between A and C it is 0.33.

The advantage of using an ecological distance is that differences in species composition can be summarized with a single statistic. The disadvantage of using an ecological distance is that information on the identities of the species is not available any longer. For some research objectives, loosing information on species identities does not pose problems, whereas other equally valid objectives require that information is available on the species identities. We will see in chapter 10 that species identities can be added to an ordination diagram, although many ordination methods use distance matrices. If you are interested in fully exploring how abundance or presence-absence of particular species changes in between sites, then you need to conduct a separate analysis for each species (chapter 6 and 7).

Ecological distance

A good ecological distance describes the difference in species composition. For sites that share most of their species, the ecological distance should be small. When sites have few species in common, the ecological distance should be large. There is no single way to define ecological distance. The literature lists a large number of distances. We only present a subset of the possible distances. These are among the most common distances that are used for analysing differences in species composition.

Distance matrices

Distance matrices provide information on the ecological distance between all pairs of sites within your data. Table 8.1 provides one of the possible distance matrices that can be calculated from the dune meadow dataset.

The cells in the distance matrix contain the distance between the sites indicated by the column and row names. As the species composition is exactly the same when you compare a site with itself, its ecological distance is zero. Also the order in which you make the comparison does not matter: the distance between X1 and X2 is the same as that between X2 and X1.

Euclidean distance

The Euclidean distance is calculated by using each species as a different axis to plot each site and then measuring the distance between the sites (see Figure 8.2). Formulae for calculating the Euclidean distance and other distances are provided in Box 8.1 (page 129).

The Euclidean distance is not a good ecological distance if it is used on raw species matrices – matrices that contain the abundance of each species on each site. When the species matrix is modified by a particular transformation (see below: standardizations of the species data before calculating the distance matrix), then the Euclidean distance becomes better at expressing ecological distance.

The following example illustrates that the Euclidean distance on the raw species matrix will not always describe ecological distance well. Imagine that you recorded the following abundances for 3 species for 3 sites:

Site	Species 1	Species 2	Species 3
A	1	1	0
B	5	5	0
C	0	0	1

You can see that sites A and B have the same species, whereas site C has a different species.
When we calculate the Euclidean distance, then we obtain the following distance matrix:

```
          A        B        C
A         0 5.656854 1.732051
B 5.656854        0 7.141428
C 1.732051 7.141428        0
```

Table 8.1 Distance matrix for the dune meadow dataset using the Bray-Curtis distance

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
X1	0.000	0.467	0.448	0.524	0.639	0.636	0.552	0.655	0.600	0.574	0.560	0.925	0.843	1.000	1.000	0.922	0.879	0.778	1.000	1.000
X2	0.467	0.000	0.341	0.356	0.412	0.511	0.439	0.537	0.476	0.294	0.541	0.714	0.600	0.788	0.908	0.893	0.825	0.594	0.808	0.945
X3	0.448	0.341	0.000	0.271	0.470	0.568	0.475	0.325	0.341	0.470	0.556	0.440	0.425	0.750	0.714	0.671	0.891	0.612	0.831	0.775
X4	0.524	0.356	0.271	0.000	0.500	0.634	0.506	0.412	0.379	0.477	0.584	0.525	0.513	0.797	0.735	0.667	0.900	0.667	0.789	0.763
X5	0.639	0.412	0.470	0.500	0.000	0.297	0.229	0.639	0.506	0.349	0.627	0.692	0.684	0.881	0.848	0.895	0.621	0.543	0.703	0.892
X6	0.636	0.511	0.568	0.634	0.297	0.000	0.227	0.591	0.600	0.319	0.450	0.639	0.753	0.806	0.803	0.852	0.683	0.493	0.722	0.848
X7	0.552	0.439	0.475	0.506	0.229	0.227	0.000	0.525	0.488	0.277	0.444	0.627	0.644	0.875	0.841	0.890	0.673	0.552	0.746	0.887
X8	0.655	0.537	0.325	0.412	0.639	0.591	0.525	0.000	0.317	0.542	0.528	0.440	0.370	0.562	0.429	0.425	0.891	0.642	0.746	0.493
X9	0.600	0.476	0.341	0.379	0.506	0.600	0.488	0.317	0.000	0.600	0.595	0.351	0.413	0.758	0.662	0.653	0.895	0.681	0.781	0.699
X10	0.574	0.294	0.470	0.477	0.349	0.319	0.277	0.542	0.600	0.000	0.413	0.718	0.737	0.761	0.848	0.895	0.621	0.486	0.703	0.892
X11	0.560	0.541	0.556	0.584	0.627	0.450	0.444	0.528	0.595	0.413	0.000	0.672	0.754	0.821	0.745	0.877	0.702	0.322	0.556	0.810
X12	0.925	0.714	0.440	0.525	0.692	0.639	0.627	0.440	0.351	0.718	0.672	0.000	0.353	0.695	0.621	0.588	0.920	0.742	0.697	0.697
X13	0.843	0.600	0.425	0.513	0.684	0.753	0.644	0.370	0.413	0.737	0.754	0.353	0.000	0.649	0.679	0.606	0.875	0.800	0.812	0.719
X14	1.000	0.788	0.750	0.797	0.881	0.806	0.875	0.562	0.758	0.761	0.821	0.695	0.649	0.000	0.362	0.544	0.897	0.843	0.855	0.455
X15	1.000	0.908	0.714	0.735	0.848	0.803	0.841	0.429	0.662	0.848	0.745	0.621	0.679	0.362	0.000	0.357	0.895	0.720	0.778	0.296
X16	0.922	0.893	0.671	0.667	0.895	0.852	0.890	0.425	0.653	0.895	0.877	0.588	0.606	0.544	0.357	0.000	1.000	0.867	0.906	0.344
X17	0.879	0.825	0.891	0.900	0.621	0.683	0.673	0.891	0.895	0.621	0.702	0.920	0.875	0.897	0.895	1.000	0.000	0.762	0.565	0.913
X18	0.778	0.594	0.612	0.667	0.543	0.493	0.552	0.642	0.681	0.486	0.322	0.742	0.800	0.843	0.720	0.867	0.762	0.000	0.552	0.690
X19	1.000	0.808	0.831	0.789	0.703	0.722	0.746	0.746	0.781	0.703	0.556	0.697	0.812	0.855	0.778	0.906	0.565	0.552	0.000	0.742
X20	1.000	0.945	0.775	0.763	0.892	0.848	0.887	0.493	0.699	0.892	0.810	0.697	0.719	0.455	0.296	0.344	0.913	0.690	0.742	0.000

One feature of our distance matrix is that information about the original species is not longer present – the distance matrix does not mention Species 1, Species 2 or Species 3 anywhere. Some textbooks mention in this case that the analysis is in Q-mode rather than in R-mode. This simply means that differences in sites are being investigated (differences between the rows of the species matrix) rather than differences in species (differences between the columns of the species matrix) – but remember that the differences were actually derived from differences in species composition. Since ecological datasets often contain more species than sites, the distance matrix will often be of smaller size than the species matrix.

Because of the properties of distance matrices that the distance between the same sites is zero and that the order of calculating the distance does not matter, a distance matrix can be summarized with no loss of information as:

```
          A         B
B 5.656854
C 1.732051 7.141428
```

You can see in the distance matrix that the distance between A and C is about 1.7, whereas the distance between A and B is roughly 5.7. But remember now that A and C do not share any species, whereas A and B have the same species. The Euclidean distance depends greatly on the abundances of each species, not just which species are shared. A and B are far apart using Euclidean distance because A has 2 plants and B has 10. However in most applications we would like to give more emphasis to the extent to which species are shared and give more weight to differences in composition than abundances of the same species.

Although the Euclidean distance is not a very good distance for investigating how species are shared between sites, it is still used in some ordination and clustering techniques as these only allow for this distance.

The fact that the Euclidean distance is not a good distance for all situations also constrains one graphical method of representing differences in species composition. Each species can be represented by one axis. Sites can then be plotted by using the abundance of each site as coordinates. For our example, we could graphically represent our sites as in Figure 8.2.

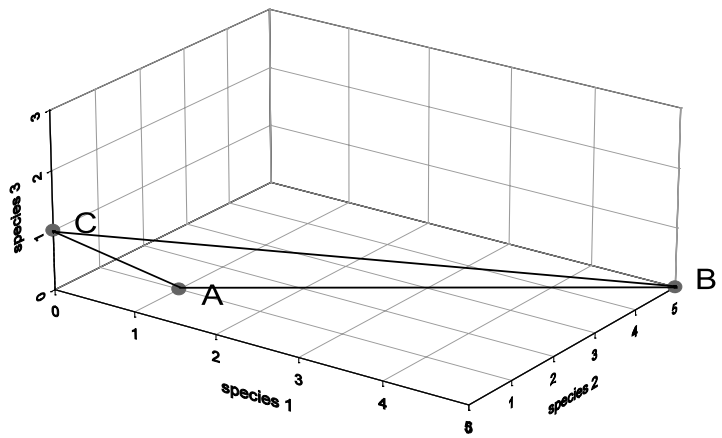


Figure 8.2 By using each species as an axis, you can position sites on a plot that reveals their ecological distance.

When you measure the straight line distance between the sites in Figure 8.2, you will obtain the Euclidean distance. You can see again that site A is closer to C than to B.

One of the solutions to give greater weight to differences in composition and still to use the Euclidean distance is to calculate species proportions first and then calculate the Euclidean distance (see below: standardizations of the species data before calculating the distance matrix). Another method would be to use presence-absence, transforming all non-zero abundance values to 1. The Euclidean distance is then the same as a count of the number of species that occur in one site but not both. The solution that is more commonly used is to adopt another method of calculating distance as shown in the next section.

Other distances

Distances can be classified in various ways. One classification method can be by the range of output values that can be expected.

Some distances are restricted to be within the range of zero to one. When the distance is zero, two sites are completely similar for every species. When the distance equals 1 they are completely dissimilar, which means that they do not share any species. The **Bray-Curtis distance** (or Odum distance, or the one-complement of the Steinhaus similarity) and **Kulczynski distance** fall within this category.

When we calculate the Bray-Curtis distance for the dataset that we described earlier, we obtain:

```
          A         B
B 0.6666667
C 1.0000000 1.0000000
```

Similarly, for the Kulczynski distance, we obtain:

```
    A   B
B 0.4
C 1.0 1.0
```

You can see now that the distance of site C from the other sites is 1. This indicates that site C does not share any species with the other sites. If site C shared any species with the other sites, then the distance would be smaller than 1. The distance between A and B is smaller, which is what we wanted since we know that these sites share some species.

Another thing that you can observe is that you obtain different values for the distance between A and B when using the Bray-Curtis distance and when using the Kulczynski distance. You are thus faced with having to make a choice of a particular distance as different results are obtained. Although there are some methods for comparing distances (see below: choice of a distance), you will in general need to make a prior choice of the distance that you want to use.

The Bray-Curtis and Kulczynski distances are calculated from differences in abundance of each species. Because of this calculation method, the final distance will be influenced more by species with largest differences in abundances. When some species are dominant in your dataset (for example that one species has 90% of differences in abundance among sites), the Bray-Curtis and Kulczynski distances will mainly reflect differences for those species only. For this reason, some researchers prefer to transform the species matrix by a square-root, double square-root (fourth-root) or logarithmic transformation (see chapter 2), so that dominant species will influence the analysis less.

Some other distances will not be constrained to a maximum value of one, but can have larger distances. The Euclidean distance is one of those measures. Other distances of this category that are better in representing differences in species composition include the **Hellinger** and the **Chi-square** distance. Both depend on differences in proportions of species between the two sites. The Chi-square distance is actually not that good for species data, but it is the distance that is used in some common ordination methods (see chapter 10). The Hellinger distance will

normally perform better (be a better reflection of ecological distance) than the Chi-square distance. Both the Chi-square and Hellinger distance will be influenced differently by the species with smaller abundances than by the species with larger abundances. Some researchers find this a feature that is not desirable, since species with smaller abundances are usually not well sampled and species with smaller abundances often contribute more to the Chi-square and Hellinger distances. We would expect that species with larger abundances are better sampled, and express differences among sites better. This could be a reason to opt for the Bray-Curtis or Kulczynski distance.

The results for the example that we used earlier are for the Hellinger distance:

```
          A        B
B 0.000000
C 1.414214 1.414214
```

For the Chi-square distance, we obtain:

```
              A        B
B 1.570092e-16
C     3.752777 3.752777
```

One thing that we can observe immediately is that the distance between A and B is 0 (with a small calculation difference for the Chi-square distance). This means that they share exactly the same species, and that the proportions of each species are the same (0.5 in this example). Distances with C are also exactly the same for sites A and B. The Chi-square and Hellinger distances calculate different values, so this means again that you will need to choose which measure to use (but see below: choice of a distance measure).

Box 8.1 How to calculate ecological distance?

We will calculate the ecological distance between sites A and C of Figure 8.1. Site A has three species (S1, S2 and S3) each with 1 tree, whereas site C has 4 trees of species S1, 1 tree of S2 and 1 tree of S3. In the formulae, the abundances of the species of site A are indicated by a_1, a_2 and a_3, and the abundances of the species of site C are indicated by c_1, c_2 and c_3. If a species does not occur on a specific site, it should be listed for the site with abundance = 0.

The formulae for calculating the ecological distances are:

Bray-Curtis: $$D = 1 - 2\frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S}(a_i + c_i)}$$

Kulczynski: $$D = 1 - \frac{1}{2}\left(\frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S}a_i} + \frac{\sum_{i=1}^{S}\min(a_i, c_i)}{\sum_{i=1}^{S}c_i}\right)$$

Euclidean: $$D = \sqrt{\sum_{i=1}^{S}(a_i - c_i)^2}$$

Chi-square: $$D = \sqrt{\sum_{i=1}^{S}\frac{(a_+ + c_+)}{(a_i + c_i)}\left(\frac{a_i}{a_+} - \frac{c_i}{c_+}\right)^2} \text{ with } a_+ = \sum_{i=1}^{S}a_i$$

Hellinger: $$D = \sqrt{\sum_{i=1}^{S}\left(\sqrt{\frac{a_i}{a_+}} - \sqrt{\frac{c_i}{c_+}}\right)^2} \text{ with } a_+ = \sum_{i=1}^{S}a_i$$

For the Bray-Curtis distance, inserting the abundances for each species in the formula gives:

$$D = 1 - 2\frac{\min(1,4) + \min(1,1) + \min(1,1)}{(1+4) + (1+1) + (1+1)} = 1 - 2\frac{1+1+1}{5+2+2} = 1 - 2\frac{3}{9} = 1 - \frac{6}{9} = \frac{3}{9}$$

Note that use of the functions within Biodiversity.R will directly calculate the distance.

Some distances only look at differences in presence or absence of a species. Some of these distances measures are the Jaccard and Sorenson distance. These also fall into the category of the indices that are constrained within the 0-1 interval. The Sorenson distance is actually the Bray-Curtis when the species matrix has been transformed to presence-absence or 1/0. These distances are not influenced by differences in species abundance between samples, so that species that have larger abundances will not carry a larger weight in the analysis. Often however, this is not a characteristic that you want in your analysis – observing just one individual may not mean very much, whereas observing many individuals on one site and none on another site may be much less likely to be by random chance only. As calculation time with modern computers can be short, you could actually test whether abundance and presence-absence data result in the same patterns when you analyse your data.

Standardizations of the species data before calculating the distance matrix

Except for presence-absence data, differences in abundances among species will influence the calculation of the distances. This is a desirable characteristic of some distances, but for some datasets that are strongly dominated by a few species this may cause a constraint. When most of the abundance is taken by a few species, the distance may only express the differences of sites for this small subset of species. In such cases, you could diminish the influence of strongly dominant species by first taking logarithms, taking square-root, whereas some scientists have even advocated taking 4^{th} roots of all the values in your data. As the typical species matrix contains many zeroes, you would need to take a $\log(n+1)$ as described in chapter 2.

You can have a similar effect of a strong dominance in your dataset by sites that contain a larger number of individuals. An approach that standardizes each site to the same abundance is to divide cell abundance by the total abundance for each site (so that species sum up to 1 for each site). When you use this method, you will compare differences in **species proportions** for each site. For example, for our original dataset, the species matrix with the proportions will be:

Site	Species 1	Species 2	Species 3	Total
A	0.5	0.5	0	1
B	0.5	0.5	0	1
C	0	0	1	1

You can now see that site A and B have exactly the same values as differences in total abundance are not longer there. The data of the species proportions for each site have been described as **species profiles** in some texts. By investigating species profiles, you are sure that differences in total abundance among sites will not influence the results. When looking at the results shown in Figure 8.1, you could have noticed that an ecological distance of 0.33 was calculated between sites A and D by the Bray-Curtis distance, although both sites contain the same species and the same proportions of each species. If the species matrix would have been standardized, then the ecological distance between sites A and D would have become 0.

It is your choice to determine whether ecological distance should reflect the raw abundances (and thus also differences in site totals) or species proportions. When you opt to standardize the species matrix, you could investigate differences in total abundance among sites by the regression techniques described in an earlier chapter as an additional analysis. Dividing by the site total has the added advantage that some distances will now provide you with similar outcomes. For example,

the results for the Bray-Curtis and Kulczynski distances will be the same for species profiles.

Some standardizations of the species data have the feature that when the Euclidean distance is calculated from the standardized data, this distance will be a more suitable ecological distance **for the original dataset**. Such standardization approaches have been described for the distance between species profiles, the Hellinger distance, the Chi-square distance and the chord distance (Legendre and Gallagher 2001). For example, if we calculate the Euclidean distance for the matrix with proportions (species profiles), then we obtain the following distance matrix:

```
          A         B
B 0.000000
C 1.224745 1.224745
```

You can verify that site A is closer to B than to C. Remember that the Euclidean distance matrix for the original data shown in the beginning of this chapter indicated that site A was closer to C than to B, which was not a preferred method of indicating differences in species composition.

Choice of a distance measure

One desirable characteristic of an ecological distance could be that sites that do not share any species are all given the same maximum distance. The distances that we described here that have this property are the Bray-Curtis and the Kulczynski distances. They should thus be preferred for analysing differences in species composition according to this criterion.

The behaviour of a distance can be analysed with **artificial datasets**. Imagine for instance that you have 10 sites with the following species abundances:

	Species 1	Species 2	Species 3	Species 4	Species 5	Site index
Site 1	7	1	0	0	0	1
Site 2	4	2	0	1	0	2
Site 3	2	4	0	1	0	3
Site 4	1	7	0	0	0	4
Site 5	0	8	0	0	0	5
Site 6	0	7	1	0	0	6
Site 7	0	4	2	0	2	7
Site 8	0	2	4	0	1	8
Site 9	0	1	7	0	0	9
Site 10	0	0	8	0	0	10

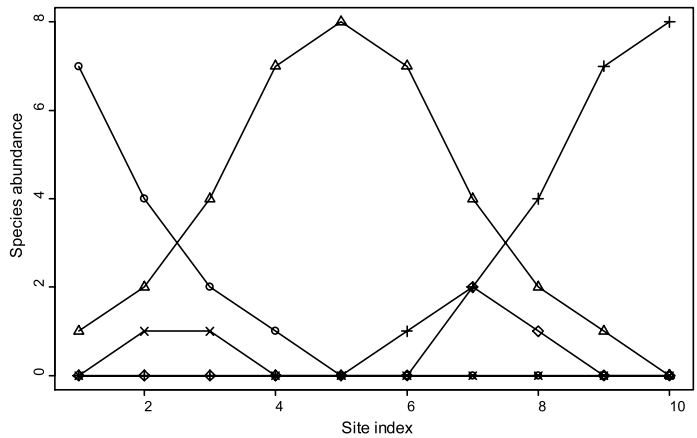


Figure 8.3 Artificial dataset with the abundance of 5 species for 10 sites.

For this artificial dataset, we arranged sites in the sequence of 1-10 (site index) based on ecological similarities between the sites, an ordering that reflects the 'smooth' changes in species abundance from site 1 as displayed in Figure 8.3. These changes are related to unimodal patterns in species abundances that are often observed in surveys. You can see that the abundance of one species (O) decreases compared to site 1, whereas abundance increases and then diminishes for another species (Δ). The particular sequence of the sites that is shown (the site index) is the only sequence that provides such smooth changes. The resulting smooth patterns show that the site index is an intuitively appealing ordering - every ecological distance provides a particular rank order that reflects certain assumptions. When we calculate the ecological distance from site 1 to the other sites, then we obtain the results that are shown in Figure 8.4. The horizontal axis shows

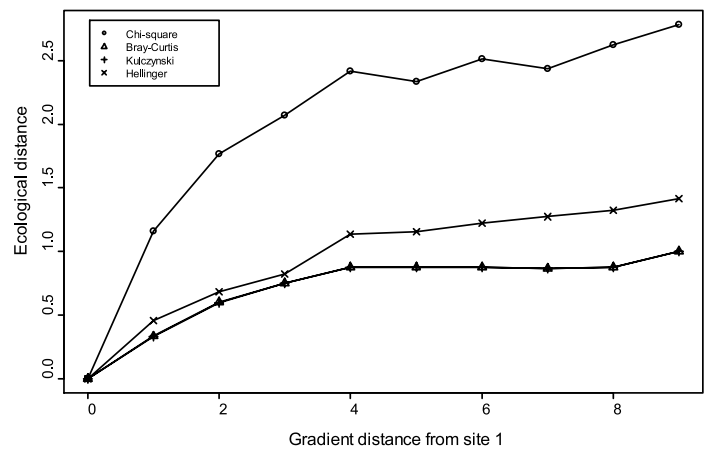


Figure 8.4 Ecological distances from the first site with the other sites of Figure 8.3.

the differences in position on the horizontal axis of Figure 8.3. This distance is the distance on a hypothetical gradient that is best related to changes in species composition.

You can see that the increments are monotonic (never decreasing) for the Bray-Curtis, Kulczynski and Hellinger distances. For the chi-square distance, the pattern fluctuates from the fourth site onwards. Although our impression of distance on the horizontal axis keeps increasing, the ecological distance does not always increase. This is the reason why we do not prefer the chi-square distance since we prefer distances that respect the order in which we arranged the sites. In this example, we prefer the other distances as they show monotonic relationship with the preferred arrangement of sites.

Another way in which the artificial dataset can be analysed is by comparing the ecological distance matrix with a **gradient distance matrix**. The gradient distance matrix is calculated from all the differences between the sites on the x axis (the gradient axis). In this matrix, the difference between site 1 and 5 equals 4, and the difference between site 3 and 5 equals 2. A summary statistic that can be calculated is the **correlation** between the values of the ecological distance matrix and the values of the gradient distance matrix. The significance level for the test that the real correlation could equal zero is calculated with a Mantel test. When you conduct a **Mantel test** for the Bray-Curtis distance for the artificial dataset, then you will obtain the following result:

```
Mantel statistic based on Pearson's product-moment correlation

Mantel statistic r: 0.833
      Significance: < 0.001

Empirical upper confidence limits of r:
  90%   95% 97.5%   99%
0.219 0.281 0.338 0.406

Based on 1000 permutations
```

The results show that correlation is high (r = 0.833) and that the significance level for the test of zero correlation is low ($P < 0.001$). This means that we have evidence for large correlation among the values of the gradient and ecological distance matrices, reflecting what we see in Figure 8.3.

Although analysis of artificial datasets (such as the one provided above) have shown that the Bray-Curtis and Kulczynski distances are good at reflecting the intuitive ordering of sites, these distances suffer from a problem that they are not metric, which can cause some problems in subsequent analyses. Being metric means that the distance between sites A and C is smaller or equal to the sum of the distance between A and B and the distance between B and C. When distances are metric, you could construct a triangle that represents the distances between A, B and C. Being metric is one requirement for the use of some ordination methods. Sometimes a transformation could help – you could for instance calculate the square-root of the Bray-Curtis distance if the Bray-Curtis itself are not metric for your data. Often the square-root of distances will be metric. Another choice that you have is to use a metric distance that also performs well for artificial datasets, such as the Hellinger distance.

Comparing a distance matrix with differences for an environmental variable

The relationship between a distance matrix and a quantitative environmental variable can be analysed with a Mantel test or by a graph as shown above for the artificial dataset. The ecological distance matrix can be based on any distance measure discussed. Then also calculate an **environmental distance matrix**, which defines the distance between pairs of sites based on the environmental variables measured. If these are quantitative the Euclidean distance may be appropriate. As in many other situations, the graph gives much insight into the relationships. The Mantel test serves to ensure that we are not mislead by patterns arising by chance. Remember it is based on correlation, only describing linear relationships between the ecological and environmental distances.

For example, if we were interested in the relationship between species composition as expressed by the Bray-Curtis distance and the depth of the A1 horizon for the dune meadow dataset, then we would obtain the following result for the Mantel test:

```
Mantel statistic based on Pearson's
product-moment correlation

Mantel statistic r: 0.2379
      Significance: 0.045

Empirical upper confidence limits of r:
  90%   95% 97.5%   99%
0.182 0.229 0.267 0.310

Based on 1000 permutations
```

Simultaneously with the Mantel test, we can plot the ecological distance against the environmental distance as in Figure 8.5.

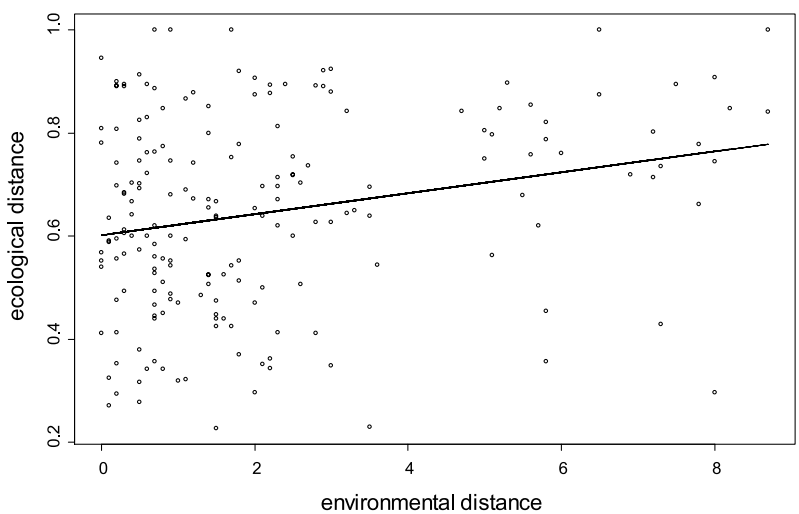


Figure 8.5 Bray-Curtis distances in relationship with differences in depth of the A1 horizon for the dune meadow dataset. The line indicates the fitted relationship between the distances by GAM.

The results show that although the significance level of the correlation is quite small (P =0.045), there is a large scatter of observations and correlation is low. This means that the correspondence between the differences in depth of the A1 horizon and the ecological distance is not very good.

A similar method to the Mantel test is to use the ANOSIM (Analysis of Similarity) test. When sites are classified by a categorical environmental variable, the method examines whether sites within categories are more similar than sites in different categories. A significance level for a test of no difference between categories is calculated. The R statistic that is calculated can be interpreted as a correlation coefficient: values close to 0 indicate little correlation with the groups, and values close to 1 or close to -1 indicate strong correlation. This statistic is unlikely to be smaller than 0 since that would indicate that similarities within categories are systematically lower than similarities among categories.

For example, the ANOSIM for the distance matrix of the dune meadow dataset based on the Kulczynski distance and for management gives the following result:

```
anosim(dis = ecology.distance, grouping
= Management)
Dissimilarity: kulczynski

ANOSIM statistic R: 0.2397
      Significance: 0.016

Based on  1000  permutations
```

Also similar to the procedure for a quantitative variable, we can plot the ecological distance against the environmental distance as in Figure 8.6. To express environmental distance, we calculate a distance of 0 if both sites have the same type of management (for example if both sites are of category hobby farming), and a distance of 1 if both sites have a different type of management (for example, one with hobby farming and one with standard farming). This method of expressing distance for categorical variables is the **Gower distance**. When the environmental variable is a categorical variable, you can also conduct the Mantel test by using the Gower distance to calculate the environmental distance matrix.

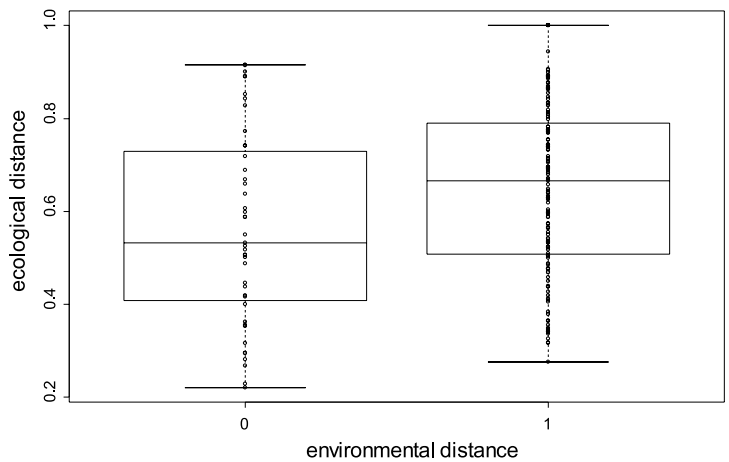


Figure 8.6 Kulczynski distances in relationship with differences in management for the dune meadow dataset.

The ANOSIM test and Figure 8.6 show that there is evidence for a relationship between the ecological distance and the type of management, but that the relationship is not very strong. We can observe several cases that have the same type of management but have a large ecological distance. We can also observe several cases that have a different type of management but have a small ecological distance. This pattern is summarized by the low ANOSIM statistic of 0.24.

In many situations it will be better to use a constrained ordination technique (see Chapter 10) to investigate the influence of environmental variables on species composition since these techniques provide a more comprehensive result.

References

Gotelli NJ and Ellison AM. 2004. *A primer of ecological statistics.* Sunderland: Sinauer Associates.

Jongman RH, ter Braak CJF and Van Tongeren OFR. 1995. *Data analysis in community and landscape ecology.* Cambridge: Cambridge University Press.

Kent M and Coker P. 1992. *Vegetation description and analysis: a practical approach.* London: Belhaven Press.

Krebs CJ. 1994. *Ecological methodology.* Second edition. Menlo Park: Benjamin Kummings.

Legendre P and Gallagher ED. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271-280.

Legendre P and Legendre L 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV. (recommended as first priority for reading)

Magurran AE. 1988. *Ecological diversity and its measurement.* Princeton, N.J.: Princeton University Press.

McGarigal K, Cushman S and Stafford S. 2000. *Multivariate statistics for wildlife and ecology research.* New York: Springer-Verlag.

Pielou E C. 1969. *An introduction to mathematical ecology.* New York: Wiley - Interscience.

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press,

Shaw PJA. 2003. *Multivariate statistics for the environmental sciences.* London: Hodder Arnold.

Doing the analyses with the menu options of Biodiversity.R

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

Calculating distance matrices:

Biodiversity > Analysis of ecological distance > Calculate distance matrix…

→Distance: bray

Transformations of the species data:

Biodiversity > Community matrix > Transform community matrix…

→Method: Hellinger

Biodiversity > Analysis of ecological distance > Calculate distance matrix…

→Distance: Euclidean

Calculating the rank-correlation with the mantel test

Biodiversity > Analysis of ecological distance > Compare distance matrices…

→Type of test: mantel

→Community distance: kulczynski

→Environmental distance: euclidean

→Environmental variable: A1

→Correlation: kendall

Calculating the ANOSIM test

Biodiversity > Analysis of ecological distance > Compare distance matrices…

→Type of test: anosim

→Community distance: kulczynski

→Environmental variable: Management

Doing the analyses with the command options of Biodiversity.R

Calculating distance matrices

```
euclidean.distance <- vegdist(dune,method="euclidean")
euclidean.distance
bray.distance <- vegdist(dune,method="bray")
bray.distance
```

Transformations of the species data

```
community.hel <- disttransform(dune, method='Hellinger')
hellinger.distance <- vegdist(community.hel,
  method="euclidean")
```

Calculating the rank-correlation with the mantel test

```
envir.distance <- vegdist(dune.env$A1, method="euclidean")
ecology.distance <- vegdist(dune, method="kul")
mantel(envir.distance, ecology.distance, "kendall")
plot(envir.distance, ecology.distance)
```

Calculating an ANOSIM test

```
ecology.distance <- vegdist(dune, method="kul")
anosim(ecology.distance, dune.env$Management)
```

CHAPTER 9

Analysis of ecological distance by clustering

Analysis of ecological distance by clustering

This chapter describes how information from distance matrices – expressing differences in species composition among sites – can be used to group sites. The aim is to put sites which have similar species compositions into the same group. The groups are known as clusters and the methods for finding them are known as clustering methods. The results of a cluster analysis might be used to describe the different vegetation types within your study area, for example prior to mapping them or investigating management options for each type.

Many different clustering methods and algorithms exist. There is no single way to decide which is 'best'. Hence results are somewhat subjective – they may depend on the choice of methods used, with little to guide that choice. The results should therefore be seen as 'ways of viewing the data', rather than definitively showing real biological structure. One way you can reduce the subjective nature of results is to try a range of different methods and see if the results are similar for each.

What is cluster analysis?

A cluster analysis arranges the sites into groups. Clusters are formed of sites that are similar in species composition, as measured by a chosen ecological distance. Cluster analysis provides a summary of the similarity in species composition of various sites. Sites that are grouped into the same cluster are more similar in species composition than sites that are grouped into different clusters (Figure 9.1).

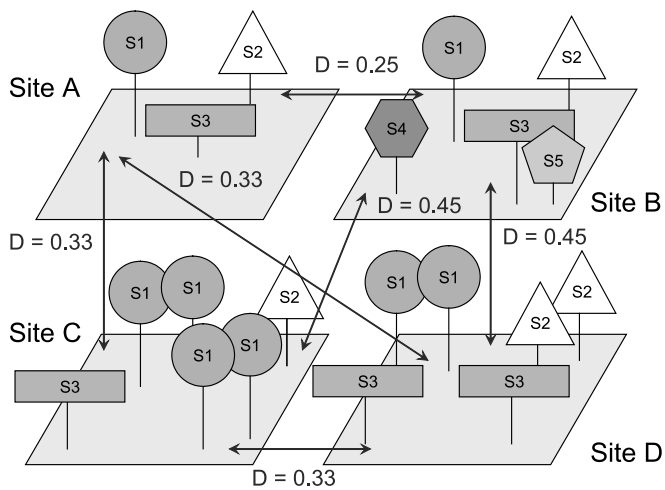


Figure 9.1 (a) A cluster analysis will group the sites into several clusters based on the ecological distance between them. The same four sites and ecological distance (Bray-Curtis) were used as in Figure 8.1.

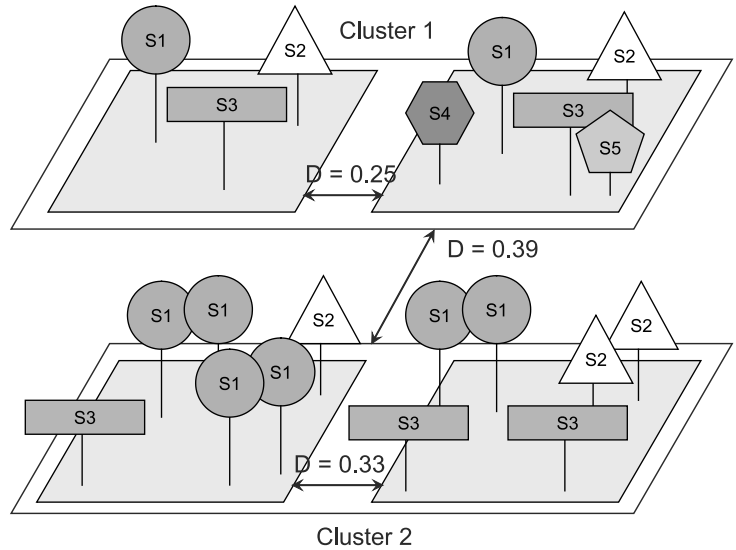


Figure 9.1 (b) The cluster algorithm (average linkage hierarchical clustering) grouped the two sites at the top in one cluster and the two sites at the bottom in a second cluster. The information that sites A and B (top) belong to cluster 1 and that sites C and D (bottom) belong to cluster 2 is a summary of the ecological distance among the sites. Distances in the figure are the distance within clusters and the average of the pairwise distances between clusters.

Hierarchical clustering methods

Hierarchical clustering methods do not only cluster sites, but also cluster the various clusters that were formed earlier in the clustering process. Hierarchical clustering methods thus provide a hierarchy for the similarity of sites. For example, the clustering method may tell you that cluster A and cluster B are more similar than a cluster A and C or a cluster B and C. Cluster A and B will then be joined into a cluster D. At a later stage in the clustering process, cluster D is then merged with cluster C.

Agglomerative clustering algorithms start by treating each site as a cluster of 1. The closest two clusters are joined to form a new cluster. Now we have a new set of clusters, one of size 2 and the rest size 1. The closest pair of this new set is merged in the next step of the clustering process. The clustering process continues until one cluster has been formed that contains all the sites.

The process can be imagined as proceeding over a range of distances from 0 to some maximum value (see chapter 8 on ecological distance). Gradually during the process, the distance considered is increased. At the beginning of the process, the method looks for the smallest distance in the distance matrix. At this step, the two sites that have this distance are placed in the same cluster. Next in the process, larger distances are considered – these could either be distances between clusters that were formed earlier, or distances with sites that were not included in any cluster yet.

The following example could make this process clearer. We generated a hierarchical clustering for the dune meadow dataset, using a single linkage algorithm. The distance matrix was calculated by using the Bray-Curtis distance (only part of the distance matrix is listed, the entire distance matrix is provided in Chapter 8):

```
       X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11   X12   X13
X2  0.467
X3  0.448 0.341
X4  0.524 0.356 0.271
X5  0.639 0.412 0.470 0.500
X6  0.636 0.511 0.568 0.634 0.297
X7  0.552 0.439 0.475 0.506 0.229 0.227
X8  0.655 0.537 0.325 0.412 0.639 0.591 0.525
X9  0.600 0.476 0.341 0.379 0.506 0.600 0.488 0.317
X10 0.574 0.294 0.470 0.477 0.349 0.319 0.277 0.542 0.600
X11 0.560 0.541 0.556 0.584 0.627 0.450 0.444 0.528 0.595 0.413
X12 0.925 0.714 0.440 0.525 0.692 0.639 0.627 0.440 0.351 0.718 0.672
X13 0.843 0.600 0.425 0.513 0.684 0.753 0.644 0.370 0.413 0.737 0.754 0.353
X14 1.000 0.788 0.750 0.797 0.881 0.806 0.875 0.562 0.758 0.761 0.821 0.695 0.649
```

The clustering algorithm reported following sequence of additions:

```
Merge:
      [,1] [,2]
 [1,]   -6   -7
 [2,]   -5    1
 [3,]   -3   -4
 [4,]   -2  -10
 [5,]  -15  -20
 [6,]   -8   -9
 [7,]  -11  -18
 [8,]    5  -16
 [9,]  -12  -13
[10,]    3    6
[11,]    4    2
[12,]   10    9
[13,]  -14    8
[14,]   11    7
[15,]  -17  -19
[16,]   14   12
[17,]   -1   16
[18,]   17   15
[19,]   18   13
```

This output shows how sites or clusters were merged in a hierarchical sequence of 19 steps, indicated between the square brackets. At the first step, the sixth (X6) and the seventh (X7) site were joined. These are labelled -6 and -7 in the output. When you check the distance matrix, then you can see that these two sites have a distance of 0.227, the smallest value of all pairwise distances (indicated in bold in the distance matrix). In the second step, the first cluster (labelled 1) was merged with the fifth site (X5, labelled -5) at a distance of 0.229 (distance of X5-X7). In the third step, a new cluster is formed by joining X3 and X4 at a distance of 0.271. The clustering process continues until finally all clusters are joined. In case a number has no "-" sign, then this indicates that clusters are joined that were formed at earlier steps with the actual numbers indicating the step. For instance, in the last step (step 19), clusters formed in steps 18 and 13 are merged.

There are many ways by which the decision can be made to join clusters. The evaluation is done at a particular level of ecological distance. These ways include:

- Joining clusters when 2 members of different clusters have the ecological distance level. This is called nearest neighbour or single linkage clustering.
- Joining clusters when all members of different clusters have the ecological distance level or a smaller value. This is called farthest neighbour or complete linkage clustering.
- Joining clusters based on the average distance between all the members. This is called average linkage clustering. An alternative name is UPWGA, which stands for Unweighed Pair-Wise Group Average. There is a variant by which a larger cluster is downweighted.
- Joining clusters based on the distance of the centroids (central positions) of clusters. It is also possible to use a weighted method, by which the size of the clusters is considered, and the algorithm favours merging of larger clusters.
- Joining clusters based on maximizing the differences among clusters and minimizing the differences within clusters. This method is called the Ward's minimum distance algorithm. This method can only use a matrix based on the Euclidean distance. With a proper transformation (see chapter 8), we could however investigate other distances than the Euclidean distance such as the Hellinger or Chi-square distance.

The various options usually result in different allocation levels of sites to clusters. It can also happen that the hierarchy of clustering changes. An example of this phenomenon is provided below (Figure 9.2).

The clustering results can be portrayed by a **dendrogram** or clustering tree. This dendrogram shows the level where clusters were joined together, and the sites within each cluster.

The graphs in Figure 9.2 show the dendrograms for average, single and complete linkage for the dune meadow dataset. The graph for the average linkage corresponds to the written information given earlier on the clustering process.

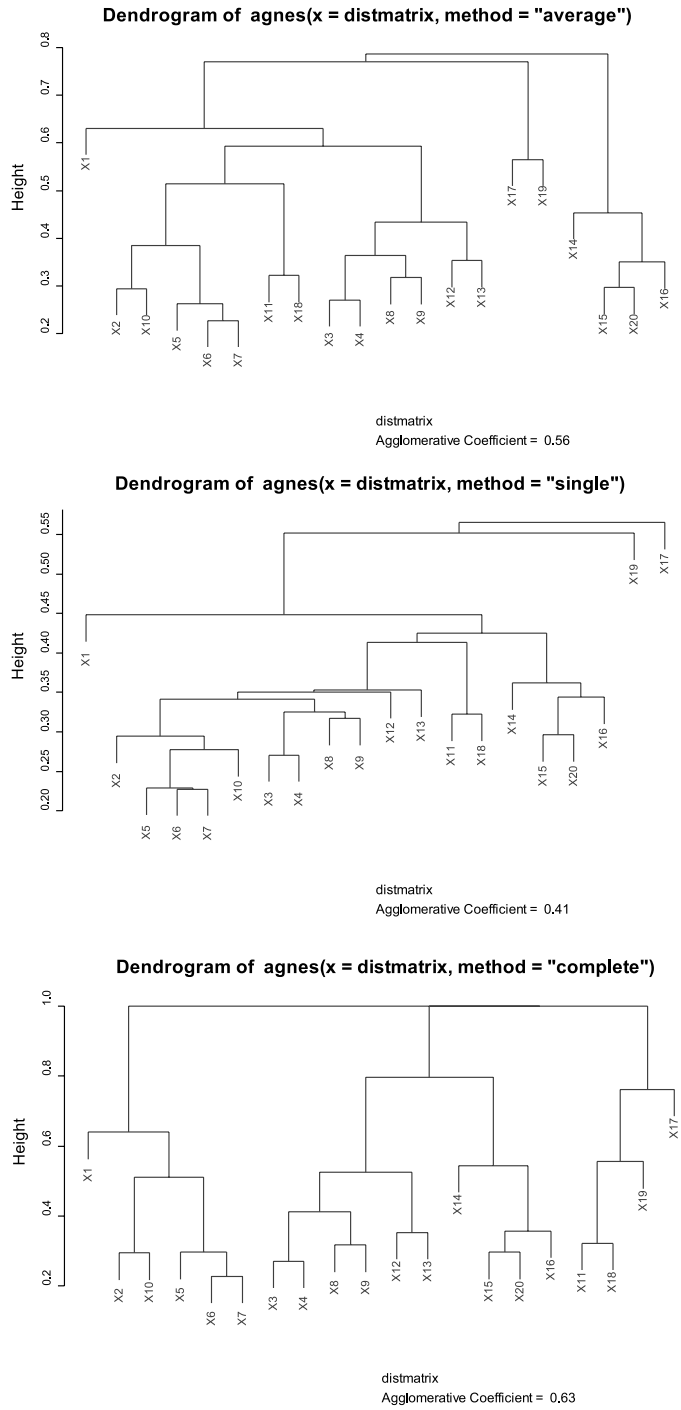


Figure 9.2 Dendrograms for different clustering methods.

From the dendrograms, you can observe some differences in the ways by which clusters are added. For example, sites X2 and X10 at the left-hand side of the histogram are added together in one cluster first in complete linkage. For single linkage, X10 is first added to the cluster that already contains X5, X6 and X7, and later X2 is added to this cluster. The dendrogram suggests that the similarity between X10 and X5 is about 0.5 for complete linkage, whereas it is less than 0.3 for single linkage.

When you study the differences between single, average and complete linkage in greater detail, you will notice that single linkage provides step-by-step addition of clusters so that you obtain a kind of stairway pattern. Complete linkage on the other hand resulted more in a pattern of fixed levels at which various clusters were joined, followed by distance levels without any merging of clusters.

It is important that you realize that the dendrograms only show at which ecological distances clusters are formed ('height' in the figures). Dendrograms do not show relationships between members of the various clusters, but only the relationship between the clusters as a whole. In Figure 9.2 the results for the complete linkage do not show that site X7 is closer to X3 (put directly right) than to X17 (put at the right-hand side of the dendrogram). The dendrogram could equally well have been drawn with X3 and X17 exchanged. You could see dendrograms as children's mobiles that can be flipped on each vertical branch, the right and left hand sides interchanged. Thus we could have drawn the same diagram with X1 somewhere in the middle, rather than at the left hand side. A very large number of dendrograms can be produced that provide the same clustering information.

Divisive hierarchical clustering methods start at the top of the clustering dendrogram by first splitting a cluster containing all sites in several clusters. The process continues at smaller distances until the last cluster of 2 sites is split.

Figure 9.3 shows a divisive clustering of the dune meadow dataset using the Bray-Curtis distance.

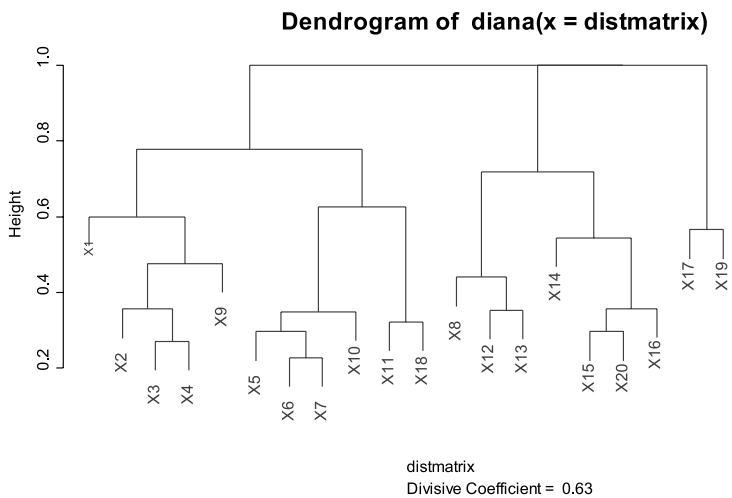


Figure 9.3 Dendrogram for a divisive clustering of the dune meadow dataset based on the Bray-Curtis distance.

In Figure 9.3, you see that X2 and X10 are only joined at a distance of about 0.8, whereas in Figure 9.1 they joined at about 0.3. Divisive clustering can thus yield quite different results from agglomerative methods.

These differences in results from different methods limit the value of clustering. However, when there are clear actual clusters in your dataset, the various methods will result in the same patterns. When your dataset contains more gradual changes in ecological distance among sites (as seems to be the case for the dune meadow dataset), the different methods will yield different results. Often we would expect species variation to follow a number of gradients rather than distinct classes. Ordination methods are then more appropriate than clustering methods. Note that clustering methods are not appropriate for describing differences or data structure that you know, or expect, to exist. For example, if the sites come from three different land uses, clustering methods should not be used to find the differences in species composition between them.

Another reason that clustering may not be used is that the dendrogram becomes too dense when you have a large number of sites. However, in such cases, also ordination methods (see chapter 10) may yield graphs of high density.

Cophenetic correlation

We mentioned above that clustering methods group sites in clusters based on their ecological distance, but that the clustering results show the ecological distance among clusters rather than among all the sites. You can evaluate how well the ecological distance among the sites is portrayed by the clustering results by calculating the **cophenetic correlation**. The cophenetic correlation is the correlation between the distances of the original distance matrix with the **cophenetic distances**.

The cophenetic distance is the distance at which clusters are combined in a dendrogram. In the example that we gave at the beginning of this chapter, a first cluster was formed by joining sites X6 and X7 at a distance of 0.227. In the second step, the {X6, X7} cluster was joined with site X5 at a distance of 0.229. The cophenetic distance between X5 and X6 and between X5 and X7 is therefore 0.229. This cophenetic distance is compared with the actual distances of 0.297 (X5-X6) and 0.229 (X5-X7).

If the cophenetic correlation is large, then the distance portrayed in the dendrogram is a good representation of distances between individual sites. The lower the cophenetic correlation, the less representative the clustering results will be of the pairwise differences. The cophenetic correlation for the average, single and complete linkage clustering of the dune meadow dataset discussed earlier are 0.82, 0.66 and 0.67, respectively. This means that in this example average clustering provides the best representation of pairwise differences.

As we mentioned for the Mantel and ANOSIM tests in Chapter 8, it is good data analysis practice to expand the investigation of the correlation to a graph that plots the relationship between the two distances by plotting all the observations. Figure 9.4 shows the correspondence between the original distance and the cophenetic distance for the average linkage of the dune meadow dataset. The horizontal pattern occurred because the cophenetic distance is calculated where clusters are joined together. You can see that the average linkage provides a good representation of the original distance – many points are close to the diagonal. Other points are farther apart from the diagonal line, so distance in the dendrogram differs more from the original distance.

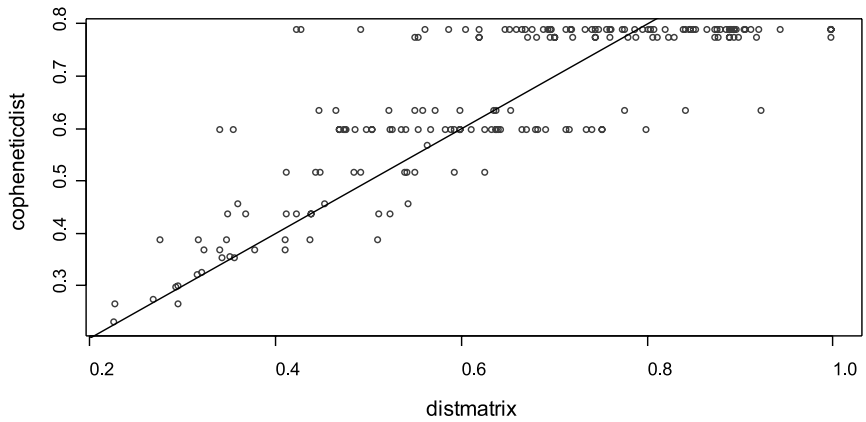


Figure 9.4 Correspondence between the original distances and the cophenetic distances for average linkage for the dune meadow dataset using the Bray-Curtis distance.

Selecting cluster memberships from hierarchical clustering

The results of a hierarchical clustering can be used to classify sites into groups. The classification is defined by choosing a certain distance level, or by choosing a fixed number of clusters. For instance, Figure 9.5 shows the grouping obtained if you decide that you want to classify all the sites into 4 groups. Note that one cluster contains only one site (X1).

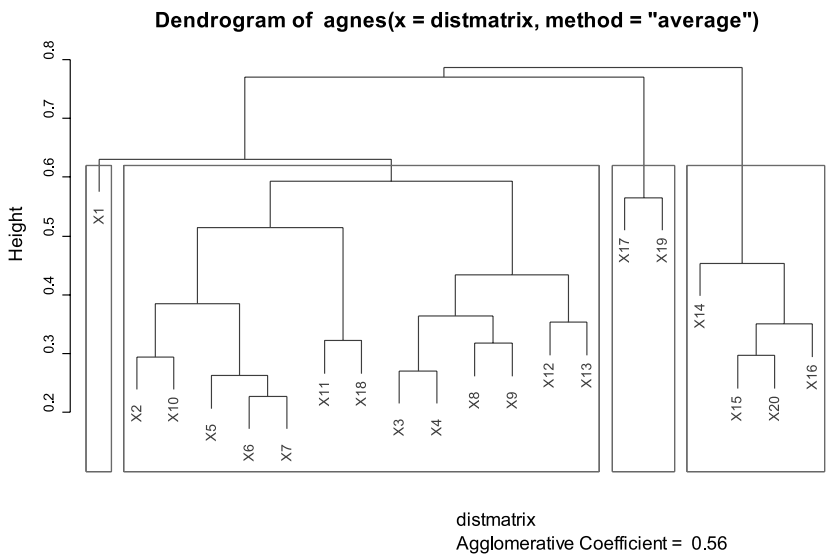


Figure 9.5 Grouping sites into 4 clusters for a dendrogram. The dendrogram is the average linkage clustering of Figure 9.2.

Table 9.1 Partitioning of sites into four clusters for two methods of non-hierarchical clustering (K-means and pam) and one hierarchical clustering method (agnes) for the dune meadow dataset.

Site	Agnes	K-means	Pam
X1	3	3	1
X2	1	3	1
X3	1	3	1
X4	1	3	1
X5	1	1	3
X6	1	1	3
X7	1	1	3
X8	1	2	1
X9	1	2	1
X10	1	1	3
X11	1	1	4
X12	1	2	1
X13	1	2	1
X14	2	4	2
X15	2	4	2
X16	2	4	2
X17	4	1	3
X18	1	1	4
X19	4	1	4
X20	2	4	2

Non-hierarchical clustering

Non-hierarchical methods divide the sample of sites into a pre-determined number of separate groups, rather than a hierarchical group structure.

There are various methods, including K-means, clara, pam and fanny. Each of these methods will put each sample in a cluster, but the algorithms differ. The K-means method can only be used with the Euclidean distance, whereas the other methods are more flexible. Table 9.1 shows the results of three different clustering algorithms used to construct 4 clusters. The agnes results (a hierarchical method) correspond to Figure 9.5. You can see from this table that most sites are put in different clusters by each method. The only cluster that is the same for the three methods is the cluster with sites X14, X15, X16 and X20. This cluster was joined at a smaller distance than the other clusters by agnes (Figure 9.5), indicating a smaller average distance between its members. The key pattern of Table 9.1 is that different methods give different results, however.

Interpretation of clustering results

The clustering results can be interpreted by analysing the relationship between the cluster membership and the data of the environmental matrix. You could, for example, draw a boxplot of the values of an environmental variable within each cluster, and look for differences in the boxplots. However, since we saw earlier that different clustering methods and different clustering options may produce different results, using clustering methods to investigate the influence of environmental variables on species composition has limited value. Clustering methods should definitely not be used in attempts to recover known structure in your data: since the clustering method is not analysing whether such structure exists, you can not expect that the method will provide such analysis.

For analyses of the influence of environmental characteristics on differences in ecological distance, ordination methods probably offer a better solution, particularly when you have hypotheses about the relationship between environmental characteristics and species composition.

Results from clustering can **complement ordination**, however. As we will see in the next chapter, the results of ordination are usually analysed by 2-dimensional graphs. A problem with these graphs could be that the distances in the 2-dimensional graph are poorly related to the overall distances. It could even be that 2 sites that are close together in the 2-dimensional graph are actually separated by a large distance, and if you plotted a 3-dimensional graph that that might become apparent. To check whether this situation occurs, you can plot the results of a nearest neighbour clustering on top of the ordination, which is also the 'minimal spanning tree'. An alternative would be to connect each site with its nearest neighbour for the entire distance matrix.

In brief, the clustering results provide a **summary** of the data. When you want to explore differences in greater detail, other techniques are better.

References

Everitt BS, Landau S and Leese M. 2001. *Cluster analysis.* Fourth edition. London: Arnold.

Gotelli NJ and Ellison AM. 2004. *A primer of ecological statistics.* Sunderland: Sinauer Associates.

Kent M and Coker P. 1992. *Vegetation description and analysis: a practical approach.* London: Belhaven Press.

Krebs C J. 1994. *Ecological methodology.* Second edition. Menlo Park. Benjamin Kummings.

Legendre P and Legendre L. 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV. (recommended as first priority for reading)

Magurran AE. 1988. *Ecological diversity and its measurement.* Princeton, N.J.: Princeton University Press.

McGarigal K, Cushman S and Stafford S. 2000. *Multivariate statistics for wildlife and ecology research.* New York: Springer-Verlag .

Pielou E C. 1969. *An introduction to mathematical ecology.* New York: Wiley - Interscience.

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press.

Shaw PJA. 2003. *Multivariate statistics for the environmental sciences.* London: Hodder Arnold.

Doing the analyses with the menu options of Biodiversity.R

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

Calculate and plot agglomerative clustering:

Biodiversity > Analysis of ecological distance > Clustering…

→Cluster method: agnes

→Distance: bray

→Cluster options: average

Calculate and plot divisive clustering:

Biodiversity > Analysis of ecological distance > Clustering…

→Cluster method: diana

→Distance: bray

Calculating cophenetic correlation:

Biodiversity > Analysis of ecological distance > Clustering…

→Cluster method: diana

→Distance: bray

→cophenetic correlation

→Plot options: cophenetic

Selecting cluster membership from a hierarchical clustering:

Biodiversity > Analysis of ecological distance > Clustering…

→Cluster method: diana

→Distance: bray

→clusters: 5

→save cluster membership (you can look at the environmental matrix later)

→Plot options: rectangles (first plot dendrogram1 or dendrogram2)

Calculating non-hierarchical clusters:

Biodiversity > Analysis of ecological distance > Clustering…

→Cluster method: pam

→Distance: bray

→clusters: 5

Doing the analyses with the command options of Biodiversity.R

Calculate and plot agglomerative clustering:

```
distmatrix <- vegdist(dune, method='bray')
distmatrix
Cluster.1 <- agnes(distmatrix, method='single')
summary(Cluster.1)
plot(Cluster.1, which.plots=2)
plot(Cluster.1, which.plots=2, hang=-1)
Cluster.2 <- agnes(distmatrix, method='single')
summary(Cluster.2)
Cluster.3 <- agnes(distmatrix, method='complete')
summary(Cluster.1)
```

Calculate and plot divisive clustering:

```
distmatrix <- vegdist(dune, method='bray')
Cluster.4 <- diana(distmatrix)
summary(Cluster.4)
plot(Cluster.4, which.plots=2)
plot(Cluster.4, which.plots=2, hang=-1)
rect.hclust(Cluster.1, k=5, border='blue')
```

Calculating cophenetic correlation (for hierarchical clusters):

```
copheneticdist <- cophenetic(Cluster.1)
copheneticdist
mantel(distmatrix,copheneticdist,permutations=1000)
plot(distmatrix, copheneticdist)
abline(0,1)
```

Selecting cluster membership from a hierarchical clustering:

```
cutree(Cluster.1,k=4)
rect.hclust(Cluster.1, k=4)
```

Calculating non-hierarchical clusters:

```
distmatrix <- vegdist(dune,method='bray')
Cluster.5 <- kmeans(dune, centers=5, iter.max=100)
Cluster.5
Cluster.6 <- pam(distmatrix, k=5)
summary(Cluster.6)
Cluster.7 <- clara(distmatrix, k=5, sampsize=6)
summary(Cluster.7)
Cluster.8 <- fanny(distmatrix, k=5)
summary(Cluster.8)
```

CHAPTER 10

Analysis of ecological distance by ordination

Analysis of ecological distance by ordination

This chapter presents some alternative methods for analysing ecological distance (differences in species composition among sites) to those introduced in the previous chapter on clustering.

Ordination methods geometrically arrange sites so that distances between them in the graph represent their ecological distances. The results of ordination are typically viewed as 2-dimensional graphs. In these graphs, each site is presented. Sites that are close together in the graph are interpreted as being similar in species composition, whereas sites that are far apart in the graph are interpreted as containing different species. Various numerical summaries support interpretation of the graphs.

In this chapter we describe both unconstrained ordination methods, which only use ecological distance, and constrained methods, which use environmental variables to guide the ordination. Some guidelines are provided at the end on choice of methods to use.

What is ordination?

In an ordination graph, sites are plotted so that distances between them in the graph reflect the ecological differences between them. In Figure 10.1, site A and site B are placed closest together. This reflects the smaller distance between these two sites.

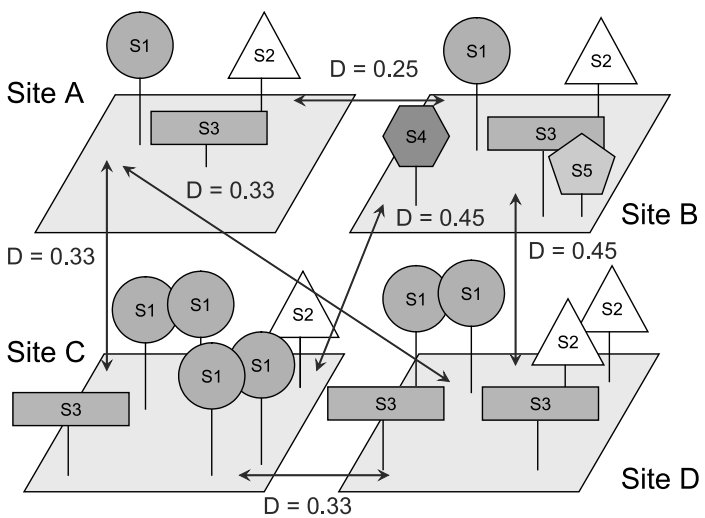


Figure 10.1(a) An ordination analysis will produce a graph that will reflect the ecological distances between sites. The same four sites and ecological distance (Bray-Curtis) were used as in Figure 8.1.

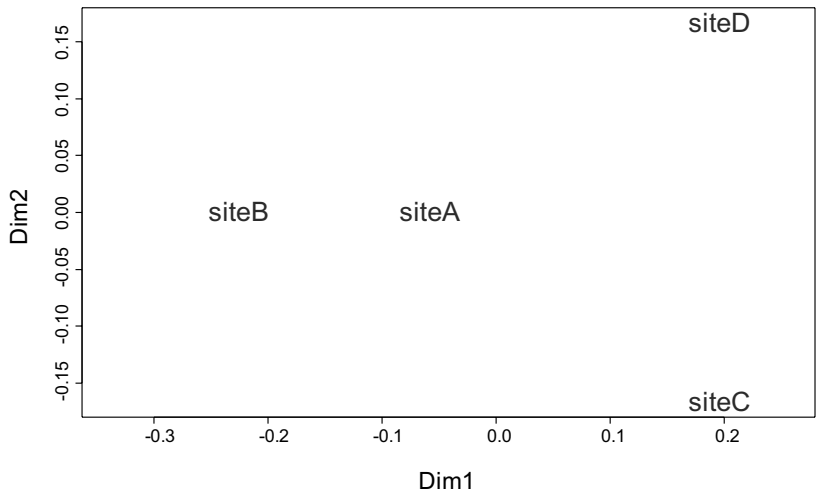


Figure 10.1(b) Sites that are closer together in the graph are more similar in species composition.

Constrained and unconstrained ordination techniques

The ordination techniques can be divided into two groups: unconstrained and constrained ordination techniques. Unconstrained ordination techniques are only based on the species matrix. Constrained ordination techniques use information from both the species and the environmental matrices. The constrained ordination techniques attempt to explain differences in species composition between sites by differences in environmental variables. You can thus examine the relationship between environmental variables and species composition.

Principal components analysis

Principal components analysis (PCA) is one of the oldest ordination techniques. It provides graphs that show the Euclidean distance between sites. No other ecological distances can be investigated with PCA. This ordination method is not ideal for analysis of information on species abundances because of the limitations of the Euclidean distance for describing community differences (see chapter 8 on ecological distance). However, when the original species matrix is suitably transformed, then the output of a PCA can become more meaningful (see below: principal components analysis on transformed species matrices).

If you calculate the PCA for the dune meadow dataset, then you will obtain the following result:

```
Partitioning of variance:

Total         84.12
Unconstrained 84.12

Eigenvalues, and their contribution to the variance

              PC1     PC2    PC3    PC4    PC5    PC6    PC7    PC8   PC9  PC10
lambda    24.7953 18.1466 7.6291 7.1528 5.6950 4.3333 3.1994 2.7819 2.482 1.854
accounted  0.2947  0.5105 0.6012 0.6862 0.7539 0.8054 0.8434 0.8765 0.906 0.928
             PC11   PC12   PC13   PC14   PC15   PC16   PC17   PC18   PC19
lambda    1.7471 1.3136 0.9905 0.6378 0.5508 0.3506 0.1996 0.1488 0.1158
accounted 0.9488 0.9644 0.9762 0.9838 0.9903 0.9945 0.9969 0.9986 1.0000

Scaling 1 for species and site scores
-- Sites are scaled proportional to eigenvalues
-- Species are unscaled: weighted dispersion equal on all dimensions


Species scores

             PC1       PC2       PC3       PC4       PC5        PC6
Achmil -1.112134  0.26681  0.02811  0.54725  1.57082  0.563344
Agrsto  2.530733 -2.07559  0.55423  0.91383 -0.33687  0.208704
…
Trirep -1.071640 -0.04554 -0.55554  0.67400  0.71926  4.092156
Viclat -0.196552  0.24913  0.30824 -0.19065 -0.59315  0.571611


Site scores (weighted sums of species scores)

          PC1       PC2       PC3       PC4       PC5        PC6
X1  -0.465152 -0.08008  0.78537 -0.32939  0.11728 -0.565387
X2  -0.892957 -0.57121  0.26702 -0.28747  0.52939  0.410929
…
X19  0.152525  1.01721 -0.59824 -0.94304 -0.11906 -0.005422
X20  1.270780  0.60335  0.27190  0.20930 -0.01970 -0.219946
```

The interpretation of the various parts of the output is as follows.

The total variance is the total variance of the species between sites. It is the sum of the individual variances of each species (column) of the species matrix. Since PCA is an unconstrained ordination method, the unconstrained variance equals the total variance. When you calculate the variance for separate species, then you will get the following results:

```
Achmil 1.5368421
Agrsto 7.2000000
Airpra 0.6184211
Alogen 6.9052632
Antodo 2.8921053
Belper 1.0815789
Brarut 3.6289474
Brohor 1.9868421
Calcus 1.5263158
Chealb 0.0500000
Cirarv 0.2000000
Elepal 5.5657895
Elyrep 4.3263158
Empnig 0.2000000
Hyprad 1.5236842
Junart 2.6210526
Junbuf 1.9236842
Leoaut 2.4315789
Lolper 7.9894737
Plalan 3.8000000
Poapra 3.4105263
Poatri 7.9236842
Potpal 0.3789474
Ranfla 1.3789474
Rumace 3.2526316
Sagpro 2.4210526
Salrep 1.9447368
Tripra 1.5236842
Trirep 3.6078947
Viclat 0.2736842
```

You can check that the sum of these variances equals the total variance in the PCA output.

Next, the eigenvalues are given. PCA is a technique that will create a new matrix from the species matrix. This new matrix has the same total variance as the original species matrix. The rows of this new matrix still correspond to the sites of the species matrix. The columns are new: these are the principal components. The eigenvalues show how much variance is found in each of the principal components.

We could have analysed differences in species composition by using each species as a separate axis in a graph. For example, Figure 10.2 shows a graph for the dune meadow dataset with axes for the two first species, using the values of the species matrix as coordinates to plot each site (see also Figure 8.2). You can notice for instance that site X4 has a value of 8 for *Agrostis stolonifera* and a value of 5 for *Alopecurus geniculatus*. Other sites are plotted in different positions, with some sites having the same position at the origin as these sites do not have either of the two species. This is an ordination graph, since sites that are close together are more similar in their composition of these two species. However, we only see information for two species and not for the remaining 28 species. Unfortunately, we can not produce graphs to show 30 axes at the same time. We are limited to graphs of 2 dimensions (3-dimensional graphs can be produced, but they are often difficult to read).

You could see PCA as a technique that creates new axes (or a matrix with new columns). The distances between the sites will remain the same in the new matrix (and thus the total variance remains the same). The advantage of the creation of new principal component axes is that more variance will be shown for the first two new axes, than if we plotted two original species axis. We can thus see a larger fraction of the total distance between sites.

The eigenvalues are listed in decreasing size. The first eigenvalue accounts for nearly 30% of the total variance. The first two eigenvalues account for more than 51% of all variance (the accounted cumulative variance of 0.5105 is provided below the second eigenvalue), and the first three eigenvalues account for more than 60% of variance. Thus, if we plot the positions on the first two new column variables (similar to the creation of Figure 10.3), we will see 51% of the variance in distances between sites. The maximum percentage of variance that could be shown by choosing original species is only 16% (Lolper + Poatri). Thus, the advantage of PCA is that axes are created that allow more variance to be shown in an ordination graph with a few axes. It will rarely be possible to show 100% of the variance in two or three axes. If too little of the variance can be shown, it is not very useful to provide an ordination graph.

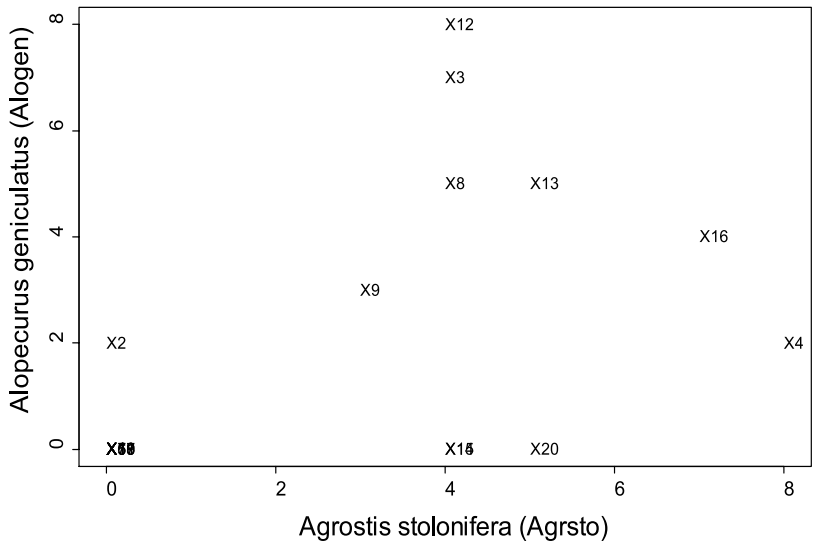


Figure 10.2 An ordination for the dune meadow dataset by using two species.

Next in the output, we get information on scaling of species and site scores. In PCA, different types of scaling of sites and species can be made. In graphs with scaling method 1, the distances between sites on the graph are approximations of the Euclidean distance between sites. For this reason, a synonym for scaling method 1 is distance scaling. If you use scaling method 2, then the distances between sites are not represented as accurately as with scaling method 1. Scaling method 2 is used to show the correlations among species better (see Figure 10.5).

The species scores provided next are coordinates that allow the plotting of species on a graph. The site scores provided next are the plotting coordinates for the sites. You can see that both scores are listed for columns named PC1 through PC6. There are actually 19 PC scores, one for each eigenvalue – the same as the number of species in the original matrix, unless there is something unusual in the data. The output was limited to 6 columns to save some space.

Plotting the site scores gives an ordination graph. Adding the species scores helps in its interpretation, as explained below (Figure 10.4).

Figure 10.3 shows the ordination graph for axes PC1 and PC2. You can interpret this ordination graph as follows. Since scaling method 1 was used, the distances between the sites reflect the Euclidean distance among them. Sites that are closer together have a small Euclidean distance in species composition – remembering, however, that the 2 axes do not show all the variance in distance. If most of the variance is captured in 2 dimensions, the correspondence will be very good, and anyway it is 'as good as it can be', as the 2 dimensions show as much of the variance as possible.

Various methods exist of investigating the goodness-of-fit of the first two axis of the PCA ordination. One method is to calculate the proportion of total variance that is explained for each site in the ordination graph, as shown in the following results:

	PC1	PC2
X1	0.13	0.14
X2	0.38	0.54
X3	0.03	0.69
X4	0.00	0.42
X5	0.40	0.41
X6	0.41	0.50
X7	0.65	0.67
X8	0.20	0.43
X9	0.00	0.38
X10	0.55	0.58
X11	0.20	0.39
X12	0.13	0.34
X13	0.07	0.55
X14	0.31	0.41
X15	0.62	0.78
X16	0.77	0.78
X17	0.00	0.39
X18	0.01	0.44
X19	0.01	0.40
X20	0.61	0.75

These results show that sites X15 and X16 are the best represented in Figure 10.3 (78% of variance), whereas site X1 is the least well represented (14%).

It is distances on the ordination graphs that are important. These are not changed if the signs (+ or -) of the scores are changed. Some software will yield scores that are the negative of the scores that you obtain with your software. The interpretation of the ordination graphs remains the same. This is true for any type of ordination graph.

The species scores show the direction from the origin (the point with coordinates (0,0) shown in the middle of Figure 10.3) where sites occur that have a larger than average value for the particular species. For example, sites X2 and X3 are expected to have larger than average values for *Poa trivialis* since this species and the two sites occur in the same direction (lower-left) from the center.

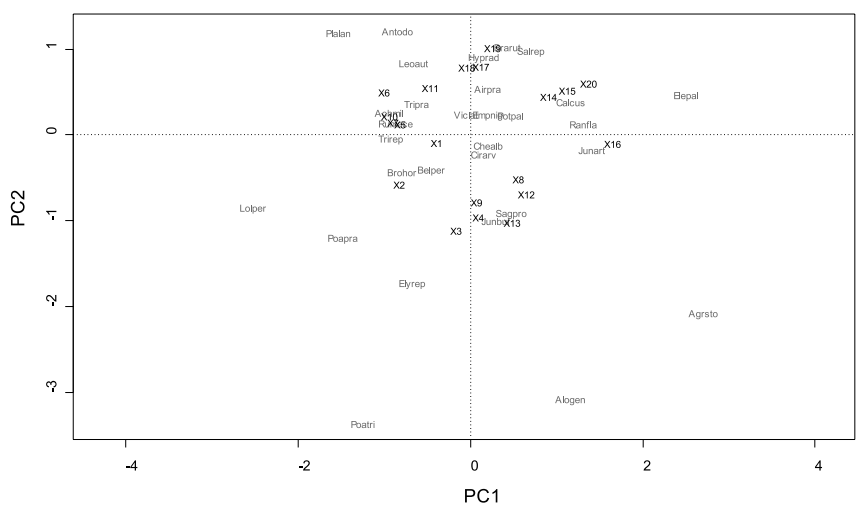


Figure 10.3 PCA ordination graph for the first two axes for the dune meadow dataset using scaling method 1.

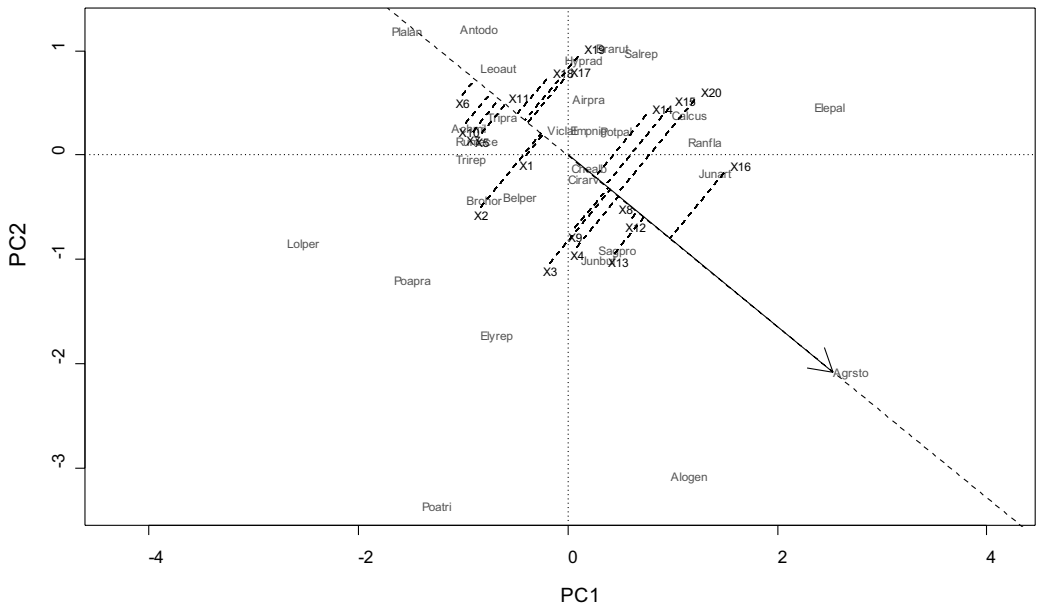


Figure 10.4 Interpretation of species scores for an ordination graph. The vector is drawn for species *Agrostis stolonifera*. The projections for the sites indicate a ranking of sites from low (lowest X6) to high (highest X16) abundance for the species.

Figure 10.4 shows a more formal method of investigating the species scores. The arrow is drawn for species *Agrostis stolonifera*. This arrow shows the direction from the origin for which sites have larger abundances for this species. By constructing perpendicular lines for each site, showing their projection onto this arrow, we get an indication of differences in abundance between sites. Sites X13 and X16 are projected farthest from the origin in the direction of the species vector. We could expect that these sites have larger abundances for the particular species than other sites. Sites X6 and X10 are projected at the opposite side of the species vector. We can expect lower than average abundances for these sites. When we check the original species matrix (chapter 2), we can see that this interpretation is a good approximation of the actual situation, with X3 and X16 having large abundance and X6 and X10 having low abundance. However, the original species matrix shows that site X4 has the largest abundance for *Agrostis stolonifera* whereas it does not have the largest position on the species vector. Since we do not see all variance in a graph, the positions will not completely reflect the exact rankings. The value of the ordination lies more in providing a good summary of distances among sites and their overall relationship with species. If you are interested in the information about a particular species, such as *Agrostis stolonifera*, then you would be better looking at that species alone, perhaps with regression analysis (chapter 6).

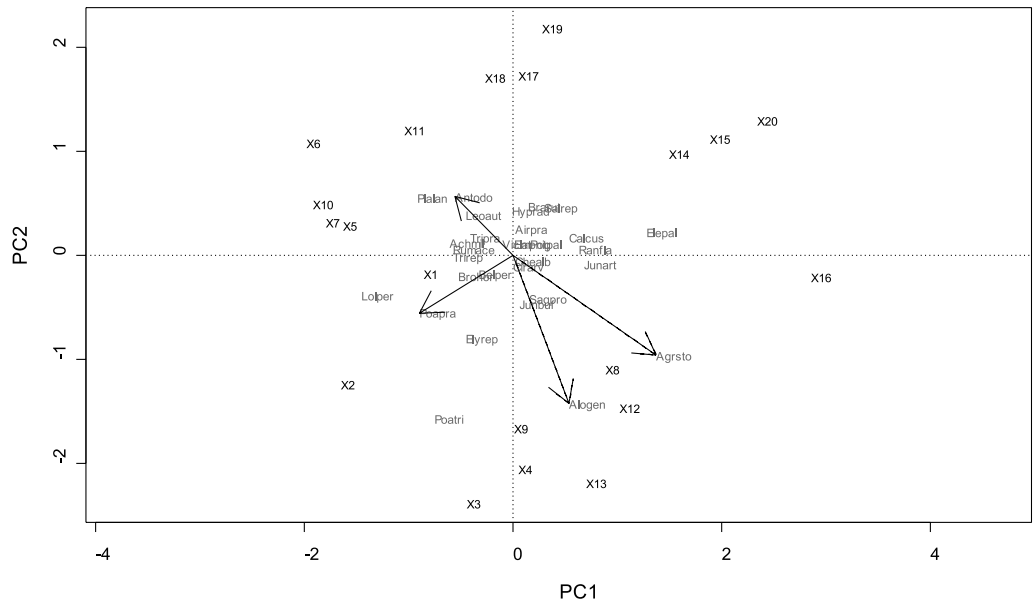


Figure 10.5 Interpretation of species scores for an ordination graph with scaling method 2. The angles between the species vectors indicate correlation among species.

As we saw earlier, scaling method 2 is used to reflect the correlations among species better. Species that have a small angle between their vectors are expected to be strongly positively correlated. Species with angles between vectors at 90 or 270 degrees are expected to not to be correlated and species with angles of 180 degrees are expected to be strongly negatively correlated. We can thus see from Figure 10.5 that *Agrostis stolonifera* is positively correlated with *Alopecurus genitalicus*, not correlated with *Poa pratensis*, and negatively correlated with *Anthoxanthum odoratum*. Correlation between species refers to similarity in abundances for the sites of two species. The respective correlations are 0.54, 0.07 and -0.45.

Note that different software packages will provide different coordinates (scores) for species and sites, even if the same scoring type was used. These differences are caused by different approaches to multiplying species and site scores in attempts to provide better graphs of sites and species. The interpretation of the graphs will remain the same, despite the differences in scales.

Some research has been done on how many principal component axes should be analysed to get a good ecological picture of the total variance of a dataset. One of the better methods is based on the broken-stick distribution. The name comes from the idea that if you chopped up the total variance into pieces randomly, in much the same way you could randomly break a stick into pieces, some of the parts would be larger than others. We are interested in parts of the variance which reflect real structure, not just random 'breaking'. So a test for the importance of principal components can be based on comparison of what you would get with random breaking and what you actually see. If you use the test based on the broken-stick distribution for the dune meadow dataset, then you obtain the following result:

	1	2	3	4
eigenvalue	24.79532	18.14662	7.629135	7.152772
percentage of variance	29.47484	21.57136	9.068950	8.502685
cumulative percentage of variance	29.47484	51.04620	60.115145	68.617831
broken-stick percentage	18.67231	13.40916	10.777577	9.023191
broken-stick cumulative %	18.67231	32.08147	42.859047	51.882238
% > bs%	1.00000	1.00000	0.000000	0.000000
cum% > bs cum%	1.00000	1.00000	1.000000	1.000000

Two criteria are available to select the number of significant axes. The first criterion is to select axes for which the percentage of variance is larger than the corresponding percentage of variance of the broken-stick distribution. For the PCA results reported above, only two axes are significant: axis 1 with 29.5% (>18.7% of the broken-stick distribution) and axis 2 with 21.6% (>13.4%). Another criterion is to select the axes for which the cumulative percentage of variance is larger than the corresponding cumulative percentage of variance of the broken-stick distribution. For the dune meadow dataset, this criterion would result in all axes to be significant.

If you used scaling method 1 (distance scaling), then there is a technique to select those species that significantly contributed to the axes shown in an ordination graph. This technique is based on an equilibrium circle. Species that significantly contribute to the ordination will have vectors outside of the equilibrium circle (Legendre and Legendre 1998). Figure 10.6 shows the vectors for the significant species for axis 1 and 2 of the PCA of the dune meadow dataset.

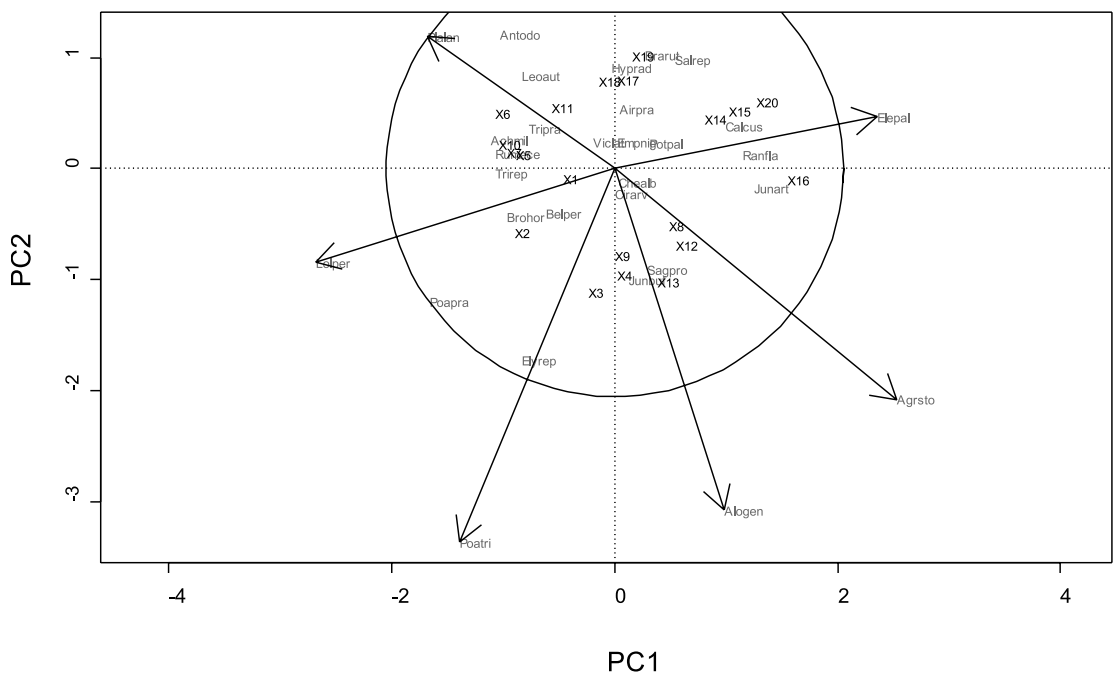


Figure 10.6 Equilibrium circle for the first two PCA axes for the dune meadow dataset. Vectors are drawn for the species that significantly contributed to the ordination graph.

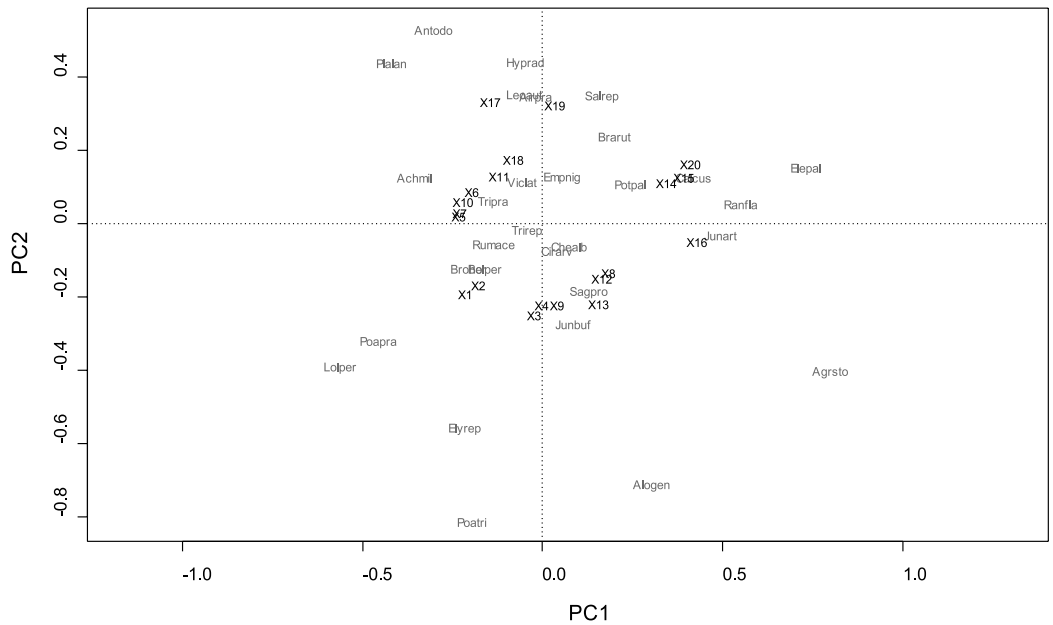


Figure 10.7 The first two PCA axes for the dune meadow dataset, after the species matrix was transformed so that distances among sites reflect their Hellinger distance.

Principal components analysis on transformed species matrices

A principal components analysis (PCA) will produce ordination graphs that portray the Euclidean distance among sites. This ordination technique is therefore limited to investigating the Euclidean distance among sites. There is one method of using PCA to investigate some other ecological distances among sites, however. This method is based on transforming the species matrix first, followed by PCA on the transformed matrix. By proper transformations, investigations are thus possible of the chi-square distance, Hellinger distance and distances between species profiles, which are better ecological distances than the Euclidean distance (see Chapter 8; Legendre and Gallagher 2001).

Figure 10.7 gives an example of the dune meadow dataset after a transformation that leaves distances representing the Hellinger distance.

Principal coordinates analysis

Principal coordinates analysis (PCoA) is an ordination technique that is similar to PCA. The technique has the advantage over PCA that any ecological distance can be investigated. In fact, when you calculate a PCoA with the Euclidean distance, then you will obtain the same result as with a PCA. A synonym for principal coordinates analysis is metric multidimensional scaling. The idea of PCoA is to start with a distance matrix. Then try to find an arrangement of sites such that that the distances between the sites in the graph match as closely as possible those in the distance matrix.

A PCoA for the dune meadow dataset based on the Bray-Curtis distance, gives the following result:

```
$points
             Dim1         Dim2
X1  -0.237339546 -0.17173114
X2  -0.197122801 -0.12450970
…
X19 -0.003609141  0.31387032
X20  0.340688509  0.10539824

$eig
[1]  9.032980e-02  5.381042e-02

$ac
[1] 0

$GOF
[1] 0.5714973 0.5961463

$eigen.total
[1] 4.990226

$R.constant
[1] 1.494617

$Rscale
[1] TRUE

$scaling
[1] 1

$cproj
               Dim1          Dim2
Achmil  -4.01695503   0.88822361
Agrsto   9.92187299  -7.98344244
…
Trirep  -2.19199995  -0.53783022
Viclat  -0.73957850   0.84489420
```

The results are similar to those for a PCA. First, the scores of the sites are given. These scores correspond to the different axes calculated by the PCoA, analogous to the PCA. Figure 10.8 shows a graph of the first two axes of the PCoA. You can see, for example, that site X1 is expected to have a small Bray-Curtis distance from site X2, and a large distance from site X20.

Next in the results, the eigenvalues are given. The eigenvalues can be interpreted in the same way as the eigenvalues of PCA: they express how much variance is shown on each axis.

Lower in the output, some parameters are provided that can be used to interpret the goodness-of-fit of the PCoA – that is, the extent to which the ordination graph reflects the distance matrix accurately. The GOF provides two ways of assessing the fit with the extracted axes. The sum of the eigenvalues for the 2 extracted axes in our results may be expressed as a ratio over the

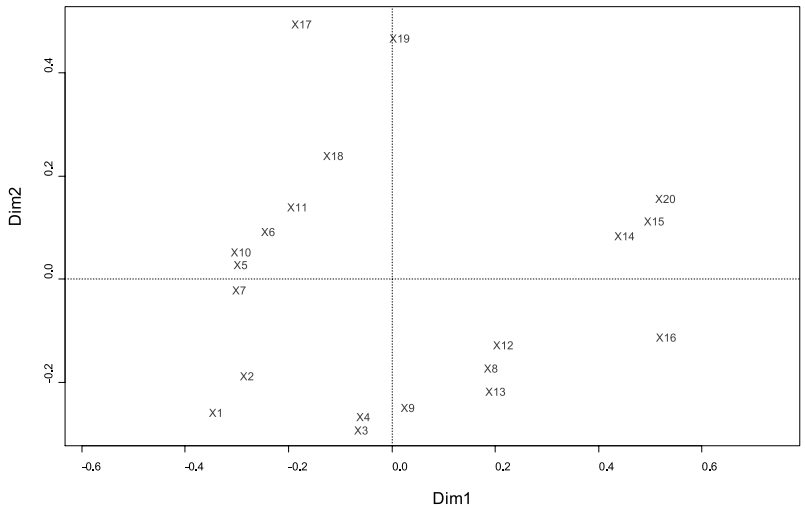


Figure 10.8 Ordination graph for the first two axes of a PCoA of the dune meadow dataset based on the Bray-Curtis distance.

total of absolute values of all eigenvalues, or over the total of all positive eigenvalues. This method of calculating the GOF may seem strange in comparison to PCA, but results from the fact that some eigenvalues may be negative in PCoA.

If all 19 axes are calculated for the PCoA, then you obtain the eigenvalues that are shown below.

You can see that eigenvalues are provided in a sequence of largest to the smallest value. This means, as for PCA, that most variance will be shown by making an ordination graph with the first two axes. When you take a closer look, you will see that the last 5 eigenvalues are negative.

The GOF gives the percentage of the first two axes when the absolute value is taken for the negative eigenvalues or when the negative eigenvalues are not considered. You can see that calculated by these methods, the first two axes show 57% and 60% of variance.

Negative eigenvalues are a result of the fact that it is impossible to calculate site scores so that the distance among the sites will equal those in the distance matrix. This will occur when distances are not metric (see chapter 8). When the distances are not metric, some eigenvalues will be negative. It is not possible to calculate site scores for negative eigenvalues. It is therefore not possible to obtain an ordination of sites that will exactly reproduce the distances of the distance matrix, even if all 19 axes are used.

The Bray-Curtis and Kulczynski distances are known to produce negative eigenvalues for some datasets. One solution (called the Caillez correction method) that will not produce negative eigenvalues is to add some constant to the elements of the distance matrix. This solution will now allow calculation of site scores for each eigenvalue. However, the distance matrix is modified and total variance was artificially increased in a rather arbitrary way. If your main interest is investigating

```
$eig
 [1]  9.032980e-02  5.381042e-02  2.428758e-02  2.011838e-02  1.469129e-02
 [6]  1.245426e-02  8.901072e-03  5.065536e-03  3.920619e-03  3.248000e-03
[11]  2.891603e-03  1.009173e-03  8.483670e-04  2.105743e-04 -5.112869e-18
[16] -1.390796e-03 -2.255631e-03 -2.880746e-03 -3.901214e-03
```

ecological distance, then you may opt not to correct for negative eigenvalues and only to analyse the first axes. Another solution would be to analyse the square-roots of the distance (the square root of the Bray-Curtis distance is expected to be metric).

A standard PCoA does not calculate scores for species. There is one way of calculating the same species scores as for PCA when the Euclidean distance is investigated. This approach is based on the correlations between species and axes, and weighting these scores by the percentage of variance expressed on each axis. The site and species scores provided in the PCoA output will be exactly the same as for PCA when you investigate the Euclidean distance. These species scores are labelled as `cproj` in the output. There is no fixed technique to calculate species scores for PCoA, however. Another technique is to calculate weighted average scores (see below: non-metric multidimensional scaling).

The species scores can be plotted in an ordination graph. Figure 10.9 shows the graph for the dune meadow dataset. The species scores should be interpreted similarly as the species scores of a PCA (see also Figure 10.4): sites that are plotted at the same direction from the origin as species are expected to have higher abundance of the species. We expect therefore that site X1 has higher abundance of *Lolium perenne* and *Poa pratensis*.

Non-metric multidimensional scaling

Non-metric multidimensional scaling (NMS or NMDS) is an ordination technique that is related to PCoA. The calculations are also done on a distance matrix. However, the positions of sites in the ordination are chosen so that rank order only of intersite distances is represented. If sites 1 and 2 are the closest in the distance matrix, they will be in the ordination graph. If site 1 is further

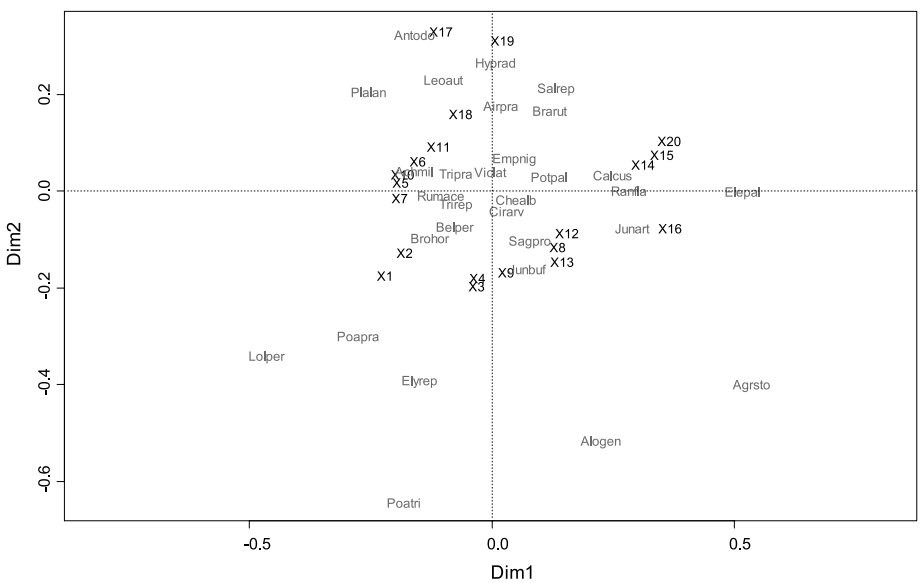


Figure 10.9 Ordination graph for the first two axes of a PCoA of the dune meadow dataset based on the Bray-Curtis distance. The species coordinates indicate the correlation of site scores with each species.

from 3 than 2, then this will also be true in the graph. However, the graph could look the same for distance matrices A and B:

Distance matrix A

	Site 1	Site 2	Site 3
Site 1	0.0	0.1	0.2
Site 2	0.1	0.0	0.3
Site 3	0.2	0.3	0.0

Distance matrix B

	Site 1	Site 2	Site 3
Site 1	0.0	0.1	0.8
Site 2	0.1	0.0	0.9
Site 3	0.8	0.9	0.0

Unless you have no faith in the quantitative nature of the distances calculated in your distance matrix, this does not seem a desirable feature.

Another difference between PCoA and NMS is that the final result is obtained through some random process. When you repeat the analysis, then you may obtain different results. You can repeat the analysis several times and then represent the best result. The best result is the one that reflects the rank-order of distances in the original distance matrix the best. The statistic that reflects how well the configuration represents the distances is called 'stress'. A smaller stress means that a better NMS ordination was calculated. Final stress values should ideally be smaller than 10% and not larger than 30% to represent species abundance data accurately.

When you calculate a NMS ordination for the dune meadow dataset based on the Bray-Curtis distance, and show the best ordination out of 100 for two axes, then you will obtain a similar result to:

```
$points
           [,1]       [,2]
X1    3.3829673 -4.7738116
X2    2.0565435 -2.6775501
X3    2.6890856 -0.4458917
X4    3.1468039 -0.7250620
X5    0.1917841 -2.9803663
X6   -0.7784914 -2.2831423
X7    0.1469503 -2.4612942
X8    1.6818537  1.8604024
X9    2.9114147  0.5966568
X10 -0.2924911 -2.2601592
X11 -1.7602236 -0.9072029
X12  2.9598205  2.4678441
X13  4.0871950  2.0987138
X14 -1.0003700  6.0644557
X15  0.3415692  5.6201219
X16  2.6892989  6.1292587
X17 -5.8442737 -2.9432844
X18 -2.5547698  0.1139317
X19 -5.1139348  1.3438575
X20  0.2423127  6.4558772

$stress
[1] 11.91173
```

You can see that this output provides the scores for the sites, and an indication of the stress. Figure 10.10 provides the ordination graph with the calculated scores.

The NMS does not provide species scores. As in PCoA, you could add the correlation or weighted average scores to the ordination graph. The interpretation of weighted average scores is as follows. The sites that are closest to a species are expected to have the highest abundance, whereas sites that are farther away are expected to have lower abundance. Figure 10.13 gives a more explicit example for the interpretation of weighted average scores. Note this is not the same as the interpretation in PCA, illustrated in Figure 10.4.

Figure 10.11 shows a graph where the weighted average scores of species were added for the NMS ordination of Figure 10.10. The species scores suggest that the abundance of *Vicia lathyroides* is highest for site X11, for example.

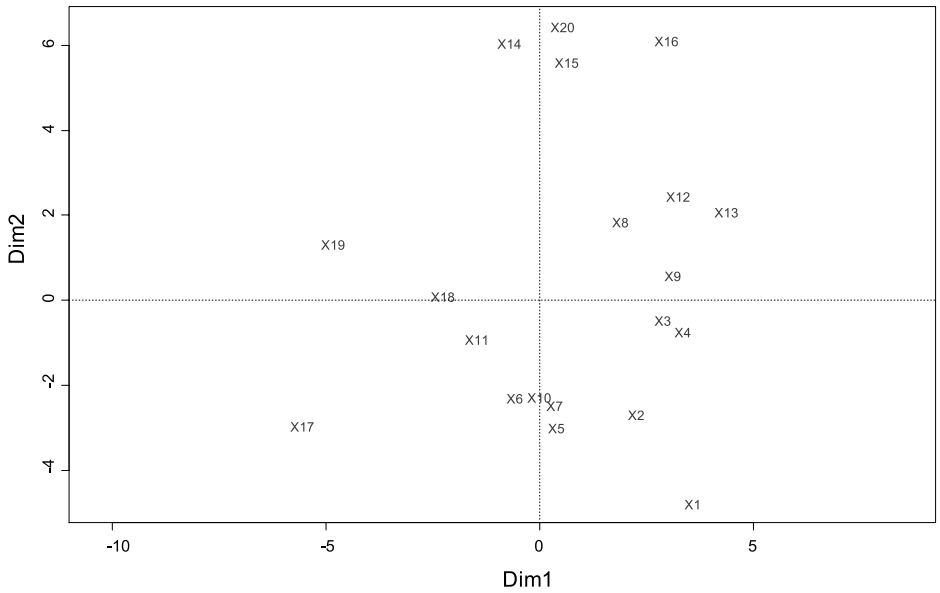


Figure 10.10 Ordination graph for a two-dimensional NMS of the dune meadow dataset based on the Bray-Curtis distance. The best configuration out of 100 is shown.

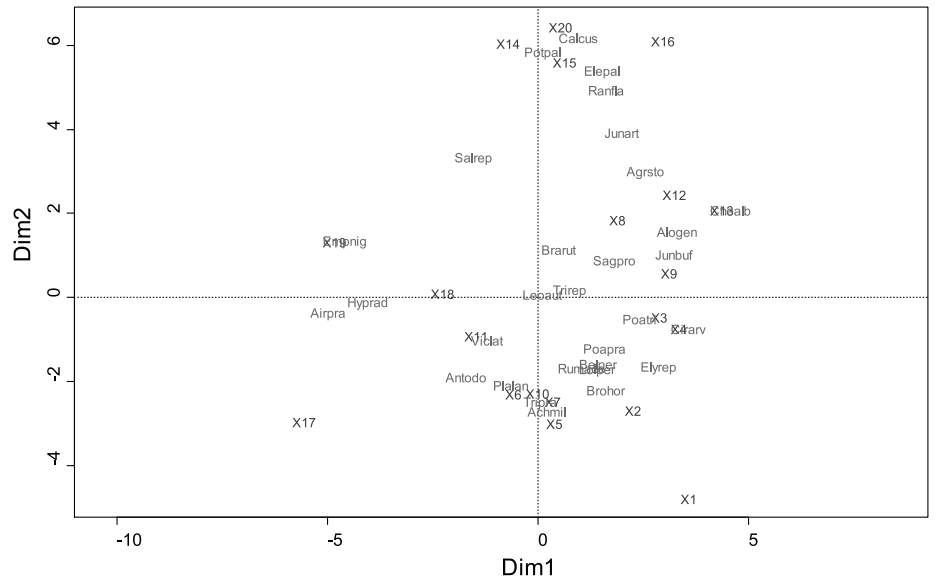


Figure 10.11 Ordination graph for a two-dimensional NMS of the dune meadow dataset based on the Bray-Curtis distance, with weighted-average scores for species.

Correspondence analysis

Correspondence analysis (CA) or reciprocal averaging (RA) is an ordination technique that shows the chi-square distance among sites. This technique was developed independently by several authors that gave it different names. One of the algorithms calculates site and species scores by consecutive steps of calculating site scores as the weighted average of the species scores, followed by calculating species scores as the weighted average of the site scores, until the results converge. This technique has an advantage over PCA because the chi-square distance is a better ecological distance than the Euclidean distance that is shown in PCA. The PCA on transformed species matrices is an alternative method that allows using better ecological distances than the chi-square distance (see above: Principal component analysis on transformed species matrices).

A correspondence analysis of the dune meadow dataset gives the result shown on the next page.

The interpretation of the different parts of the output is quite similar to the output of the PCA provided above. The interpretation is as follows.

The partitioning of the mean squared contingency coefficient reflects how much of the variance in chi-square distances was calculated. As CA is an unconstrained ordination technique, all variance is calculated.

Next the eigenvalues are given. As for PCA and PCoA, the eigenvalues are listed from highest to lowest, and their sum equals the total variance. Again as for PCA and PCoA, by plotting the first axes, more variance will be shown. When plotting the first two axes, 44% of variance will be shown (the accounted values indicate the cumulated proportion of variance).

The scaling method 1 is the same as for PCA. This scaling method means that the distances between sites reflect the chi-square distance. Scaling method 2 means that the distances between the species reflect the chi-square distance among species.

The species and site scores are the coordinates that are used in ordination graphs of the CA. Figure 10.12 shows the ordination graph for the first two axes of the CA for the dune meadow dataset.

```
Partitioning of mean squared contingency coefficient:

Total         2.115
Unconstrained 2.115

Eigenvalues, and their contribution to the mean squared contingency coefficient

             CA1    CA2    CA3    CA4    CA5    CA6     CA7     CA8     CA9   CA10
lambda    0.5360 0.4001 0.2598 0.1760 0.1448 0.1079 0.09247 0.08091 0.07332 0.0563
accounted 0.2534 0.4426 0.5654 0.6486 0.7170 0.7680 0.81175 0.85000 0.88467 0.9113
             CA11    CA12    CA13    CA14    CA15     CA16     CA17     CA18
lambda    0.04826 0.04125 0.03523 0.02053 0.01491 0.009074 0.007938 0.007002
accounted 0.93410 0.95360 0.97025 0.97995 0.98700 0.991293 0.995046 0.998356
             CA19
lambda    0.003477
accounted 1.000000

Scaling 1 for species and site scores
-- Sites are scaled proportional to eigenvalues
-- Species are unscaled: weighted dispersion equal on all dimensions


Species scores

             CA1      CA2      CA3       CA4      CA5      CA6
Achmil -1.241039  0.13375 -1.15041 -0.021261 -1.73513 -0.57465
Agrsto  1.275444 -0.32647  0.55258  0.057909 -0.36618 -0.06869
…
Trirep -0.104710 -0.03213 -0.40404  0.063081 -0.49082  1.64252
Viclat -0.845393  0.58712 -0.90362 -2.384693  3.05575  4.41315


Site scores (weighted averages of species scores)

         CA1        CA2      CA3      CA4      CA5      CA6
X1  -0.59425 -0.6848644 -0.07380 -0.88373 -0.14948 -0.60267
X2  -0.46320 -0.4401641 -0.04948 -0.49792 -0.37167  0.02160
…
X19 -0.50536  2.0648336  0.99756 -0.07422 -0.02797 -0.05283
X20  1.42353  0.6761410 -0.33944 -0.23205  0.60727 -0.55941
```

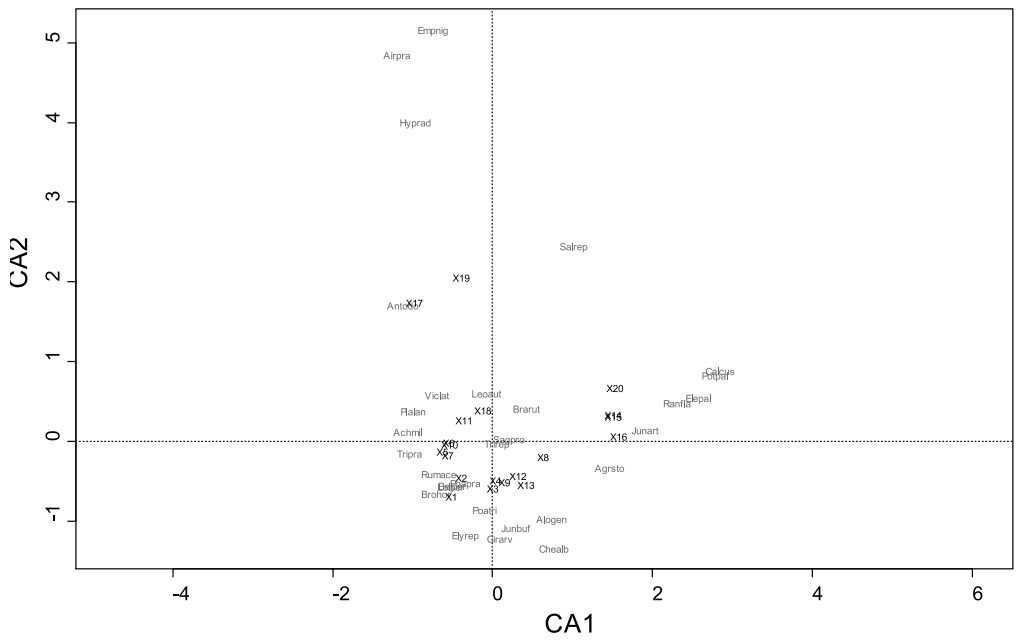


Figure 10.12 Ordination graph for the first two axes of correspondence analysis of the dune meadow dataset, using scaling method 1.

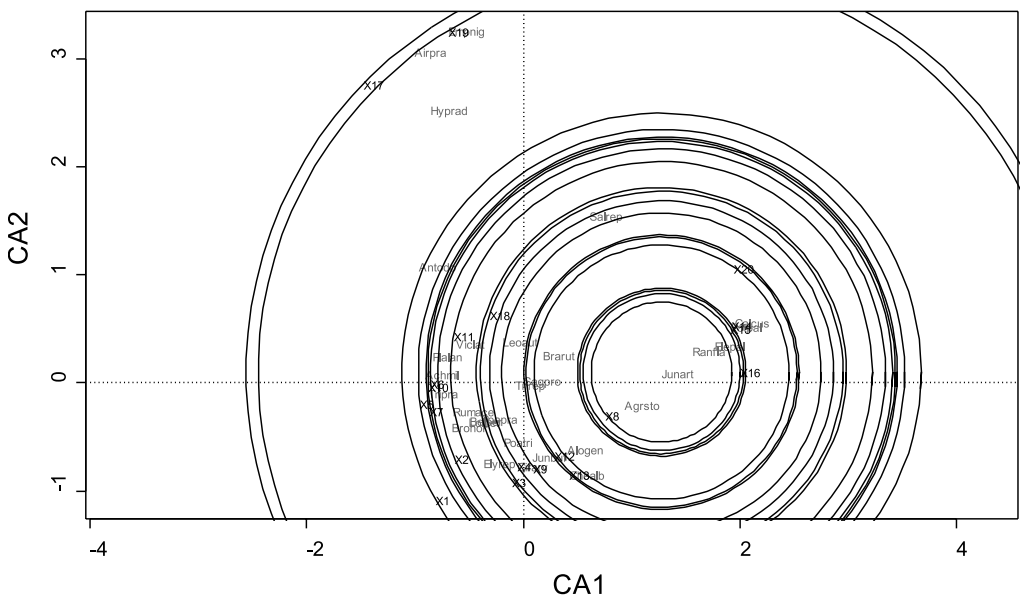


Figure 10.13 Interpretation of species scores for a correspondence analysis. Larger circles indicate a smaller expectation for the abundance of *Juncus articulatus* for the correspondence analysis of the dune meadow dataset (scaling method 2).

Species scores should be interpreted as weighted average scores. Sites that have positions in the graph close to the species are expected to have high abundance for the species, whereas sites that are farther are expected to have low abundance. Figure 10.13 shows how the position of *Juncus articulatus* should be interpreted. A smaller circle – or smaller distance between the species and site – means that we expect higher abundance for a site. We thus expect high abundance for sites X8 and X16, and low abundance for site X17. Because we do not show all variance in a two dimensional graph, the actual abundance may differ. Species that are drawn at intermediate distance from the origin are often best portrayed, not species at the edges or close to the origin.

The site scores can be thought of as a measure of 'suitability' of the site. CA assumes a unimodal distribution of species on this suitability gradient – either too high or too low results in low abundance of a species. PCA, on the other hand, assumes a linear relationship with 'suitability'. This difference in the underlying assumptions may be another advantage of CA over PCA, because the unimodal distribution is more common in nature than the linear distribution.

Simulation studies have shown that the ordination provided by a CA will often not reconstruct the known structure of the data. Based on this information, some modifications were made in CA resulting in the ordination technique of detrended correspondence analysis (DCA). However, some other simulations have shown that DCA is often not successful either in reconstructing the known structure of data. Although DCA has often been used in the past, we prefer to use other methods for ordination.

Redundancy analysis

Redundancy analyis (RDA) is the first of the constrained ordination techniques that will be discussed in this chapter. RDA is an ordination technique that is related to PCA. Both techniques show the Euclidean distance between sites in the ordination graphs. The difference is that in RDA the ordination is constrained by the environmental variables, shown in the environmental matrix. The new axes for displaying the species matrix are constrained to be linear combinations of the columns of the environment matrix, similar to the fitted values you would get by regressing PCA scores on the environmental variables.

A RDA of the dune meadow dataset, using the type of management as a constraining variables gives the following result:

```
Partitioning of variance:

Total         84.12
Constrained   29.23
Unconstrained 54.89

Eigenvalues, and their contribution to the variance

             RDA1    RDA2   RDA3     PC1    PC2    PC3    PC4    PC5    PC6    PC7
lambda    14.8654 10.6904 3.6750 15.2700 8.4275 6.8989 5.6749 3.9884 3.1212 2.5875
accounted  0.1767  0.3038 0.3475  0.1815 0.2817 0.3637 0.4312 0.4786 0.5157 0.5464
             PC8    PC9   PC10   PC11   PC12   PC13   PC14   PC15   PC16
lambda    2.3802 1.8181 1.3762 0.9951 0.7853 0.6610 0.4666 0.2827 0.1594
accounted 0.5747 0.5963 0.6127 0.6245 0.6339 0.6417 0.6473 0.6506 0.6525
```

```
Scaling 1 for species and site scores
-- Sites are scaled proportional to eigenvalues
-- Species are unscaled: weighted dispersion equal on all dimensions

Species scores

           RDA1     RDA2     RDA3      PC1      PC2      PC3
Achmil  0.71492  1.19801  0.61663 -0.79733  0.03573  0.02448
Agrsto -0.29018 -3.22454  0.08531  2.85395 -0.50083 -0.52967
…
Trirep  0.87258  1.85598  1.07994 -0.13477 -0.84708  1.12297
Viclat  0.09410  0.44974  0.85491 -0.09270  0.38100  0.10988

Site scores (weighted sums of species scores)

       RDA1     RDA2      RDA3      PC1     PC2      PC3
X1   0.3169  0.10573  0.312124 -0.88490  0.4845 -0.83687
X2   0.9030  0.39920  0.667488  0.10346 -0.7506 -0.26163
…
X19 -0.9860  0.53097  0.103590 -0.60726 -0.2858  0.50996
X20 -1.2599 -0.33664 -0.270509  0.81481  0.2488 -0.22649

Site constraints (linear combinations of constraining variables)

       RDA1    RDA2     RDA3
X1   0.2150 -0.7191  0.12068
X2   0.5179  0.7110  0.50471
…
X19 -0.8890  0.1518 -0.02099
X20 -0.8890  0.1518 -0.02099

Biplot scores for constraining variables

                RDA1    RDA2     RDA3
ManagementHF  0.4838  0.2911 -0.82535
ManagementNM -0.9793  0.1972 -0.04651
ManagementSF  0.2369 -0.9340  0.26735

Centroids for factor constraints

                RDA1    RDA2     RDA3
ManagementBF  0.5179  0.7110  0.50471
ManagementHF  0.4980  0.2541 -0.42245
ManagementNM -0.8890  0.1518 -0.02099
ManagementSF  0.2150 -0.7191  0.12068

Test for significance of all constrained eigenvalues
Pseudo-F:         2.840018
Significance:     < 0.001
Based on 1000 permutations under reduced model.
```

When you analyse the results, you could notice that the format is similar to the output of a PCA and CA. Note that there are 4 management practices, so 3 degrees of freedom or dimensions for this variable.

First the total, constrained and unconstrained variances are provided. The RDA provides an ordination that is constrained by some environmental variables. The RDA shows 29.2 from the total 84.1 variance, or 34.7% of variance is shown. This means that not all differences in Euclidean distance among sites are shown, but only those differences that can be related to differences in environmental variables.

Next the eigenvalues are provided. As for methods that were discussed earlier, the eigenvalues show how much variance is expressed on each axis. You can see two types of axes. The RDA axes are calculated for the RDA ordination. The PCA axes are calculated by a PCA on the variance that was not explained by the RDA. You can see that the RDA axes show the 34.7% of the total variance that can be explained by constrained axes (the accounted values indicate the cumulated proportion of variance). Again eigenvalues are listed from highest to lowest, thus making a graph of the first axes will result in most variance being shown.

The output follows with information on the scaling method used. As for PCA, scaling method 1 means that the distances among sites in the ordination graph corresponds to Euclidean distance among sites. Scaling method 2 would produce a correlation scaling as for PCA.

Next the species scores and site scores are provided. These scores are used in an ordination graph. These scores depend on the scaling method, and on differences in scaling of axes between software packages. The species scores should be interpreted as vectors indicating the direction of larger abundance for the species (see Figure 10.4). The biplot scores for constraining variables tell you how to interpret the RDA axes. RDA1 is roughly NM versus the rest, and RDA2 is roughly SF versus the rest (Figure 10.14).

RDA is a regression-type model that predicts where sites will occur in an ordination graph based on the environmental variables. The site constraints show the predicted value for each site. Note that the same value is predicted for each site with the same type of management. This is what we expect, as the predicted position will depend only on the environmental variables included, in this case management.

The biplot scores and the centroid scores are two methods of plotting an environmental variable in the ordination graph. The biplot scores are scores for vectors of each of the environmental variables used in the analysis. The direction from the origin to the biplot score shows sites with higher values for the environmental variable. The interpretation is thus similar to the species scores (see above and Figure 10.4). The centroid scores show the average position that is predicted for sites of the same category. The biplot and centroid scores are further discussed below in this section.

The output finishes with a significance test. The value of $P < 0.001$ shows that it is not very likely that the pattern that we observed was just random. We can thus be confident that we described an actual pattern of our data. The significance value was calculated from a permutation test. Recommended numbers of permutations are 1000 for a significance level of 0.05 and 5000 for a significance level of 0.01.

When we plot the site, species and centroid scores for the first RDA axes, then we obtain the ordination graph shown in Figure 10.14 (next page).

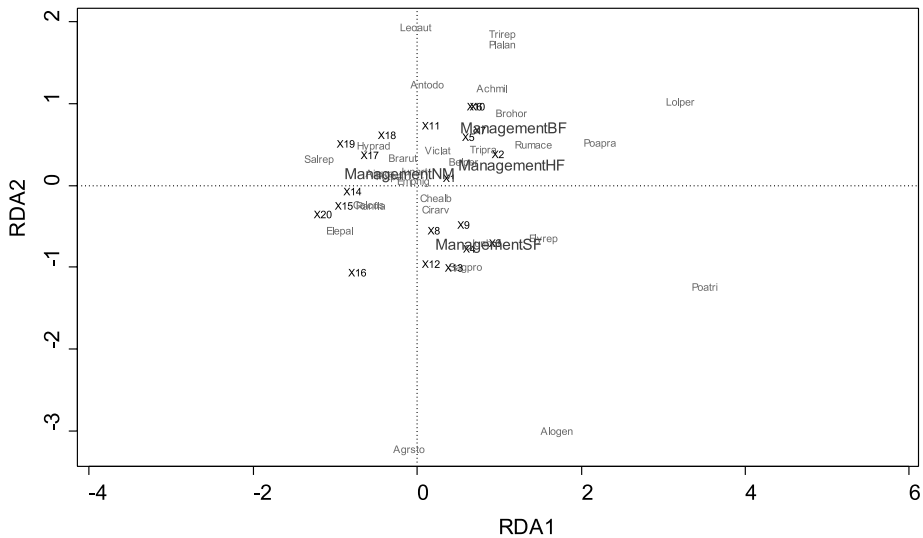


Figure 10.14 RDA ordination graph for the first two axes for the dune meadow dataset using scaling method 1 and the type of management as a constraining variable.

As discussed earlier, the species scores show the direction of higher abundance of a particular species. We thus expect higher abundance of *Alopecurus genitalicus* for sites X13 and X12, and lower abundance for X17, X18 and X19. The centroid scores show where sites of the same category of management are expected in the graph. These scores can be interpreted in a similar way as a site score. For instance, we expect sites with standard farming to contain more *Poa trivialis and Alopecurus genitalicus.*

As for regression analysis, it is possible to include several environmental variables. Figure 10.15 shows the ordination graph for the first two RDA axes when all environmental variables were used as explanatory variables. You can see that the ordination graph has become very busy with all the site, species, biplot and centroid scores. A method of producing graphs that can be interpreted more easily is to produce separate graphs for different sets of scores, for instance one graph with site and centroid scores, and one graph with species scores.

A slightly more complicated modification of RDA is partial redundancy analysis. This method is based on first removing the variance that can be explained by one subset of data, and then only analysing the residual variance. By this technique, it is for instance possible to remove the influence of spatial variables in the data first, and then analyse the influence by other variables. It is most useful for removing known environmental effects so that possible further effects can be explored (Borcard et al. 1992).

Similar to PCA, it is possible to transform the species matrix before the RDA, so that ecological distances other than the Euclidean distance are shown, such as the Hellinger distance or the distance between species profiles (see above: Principal components analysis on transformed species matrices). This transformation approach can substantially increase the usefulness of RDA.

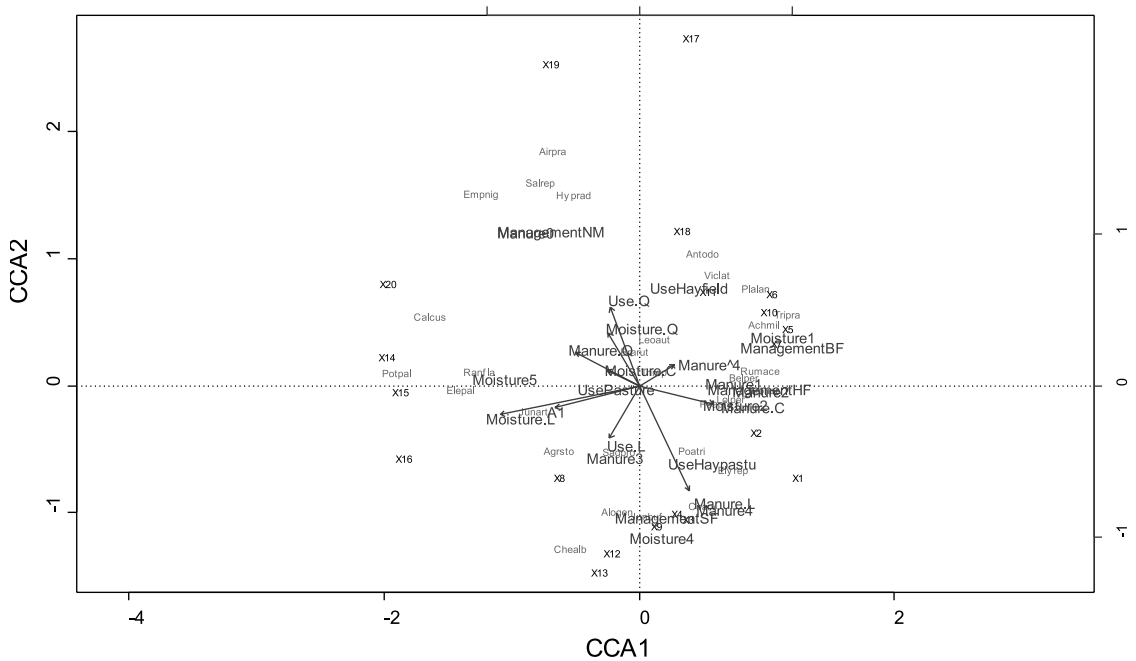


Figure 10.15 RDA ordination graph for the first two axes for the dune meadow dataset using scaling method 1 and all the environmental variables as constraining variables.

Canonical correspondence analysis

Canonical correspondence analyis (CCA) is an ordination technique that is related to CA, as suggested by its name. Both techniques show the chi-square distance among sites in the ordination graphs. As discussed in chapter 8, this distance is not the best ecological distance measure, although it is better than the Euclidean distance. In CCA the ordination is constrained by the environmental variables, shown in the environmental matrix. The approach of CCA is similar to RDA, with the CCA axes constrained to be linear combinations of environmental variables.

A CCA of the dune meadow dataset, using the type of management as a constraining variable gives the following result (next page):

```
Partitioning of mean squared contingency coefficient:

Total         2.1153
Constrained   0.6038
Unconstrained 1.5114

Eigenvalues, and their contribution to the mean squared contingency coefficient

            CCA1   CCA2   CCA3    CA1    CA2    CA3    CA4    CA5     CA6     CA7
lambda    0.3186 0.1825 0.1027 0.4474 0.2030 0.1630 0.1346 0.1294 0.09494 0.07904
accounted 0.1506 0.2369 0.2855 0.2115 0.3075 0.3845 0.4482 0.5093 0.55421 0.59158
              CA8     CA9    CA10    CA11    CA12    CA13     CA14     CA15
lambda    0.06526 0.05004 0.04321 0.03870 0.02385 0.01773 0.009172 0.007959
accounted 0.62243 0.64609 0.66651 0.68481 0.69609 0.70447 0.708805 0.712568
              CA16
lambda    0.004157
accounted 0.714533

Scaling 2 for species and site scores
-- Species are scaled proportional to eigenvalues
-- Sites are unscaled: weighted dispersion equal on all dimensions

Species scores

           CCA1     CCA2      CCA3       CA1       CA2      CA3
Achmil  0.19825  0.72622  0.300280  0.536978 -0.435913  0.64609
Agrsto -0.10058 -0.73046  0.007614 -0.681142  0.217728  0.03744
…
Trirep -0.04953  0.33644  0.186947 -0.090075 -0.056136  0.12455
Viclat -0.14617  0.98179  1.261557  0.241368 -0.211654 -0.51063


Site scores (weighted averages of species scores)

        CCA1    CCA2     CCA3       CA1      CA2      CA3
X1   1.35561  0.4747  1.06564  0.823027 -2.26986 -0.78064
X2   0.86419  0.8142  1.87245 -0.198170 -0.21094 -0.45453
…
X19 -2.61902  0.6190  0.42849  2.309910  2.21260 -0.08448
X20 -2.32115 -1.3309 -1.00044 -1.476716 -0.09132 -0.90420


Site constraints (linear combinations of constraining variables)

       CCA1     CCA2     CCA3
X1   0.5601 -1.38599  0.35086
X2   0.4313  1.32739  1.70492
…
X19 -1.8785 -0.05503 -0.06852
X20 -1.8785 -0.05503 -0.06852
```

```
Biplot scores for constraining variables

                CCA1     CCA2     CCA3
ManagementHF  0.3716  0.42702 -0.82287
ManagementNM -0.9989 -0.02624 -0.03819
ManagementSF  0.3665 -0.90307  0.22847

Centroids for factor constraints

                CCA1     CCA2     CCA3
ManagementBF  0.4313  1.32739  1.70492
ManagementHF  0.5583  0.63731 -1.22396
ManagementNM -1.8785 -0.05503 -0.06852
ManagementSF  0.5601 -1.38599  0.35086

Test for significance of all constrained eigenvalues
Pseudo-F:         2.130750
Significance:     0.001
Based on 1000 permutations under reduced model.
```

The output provided above is similar to the output for CA and for RDA.

First the total and constrained variance is given, followed by the variance shown on each axes by the eigenvalues. We can see that 28.5% of variance is accounted in the CCA, with 23.7% on the first two axes (the accounted values indicate the cumulated proportion).

The output is followed by an indication of the scaling method that was used. The scaling method is the same as for CA, scaling method 1 meaning that the chi-square distances among sites are shown.

The species, site, biplot, and centroid scores are the coordinates for species, sites and environmental variables in an ordination graph. The interpretation of these scores is similar to the interpretation for RDA. Differences with RDA are that the species scores are weighted average scores (as for CA, see Figure 10.13), and that the distance among sites reflects the chi-square distance.

Figure 10.16 (next page) shows the ordination graph for the dune meadow dataset, using the scores provided above.

The significance test is also based on permutation, as for RDA. The $P < 0.001$ indicates that the observed relationship between environmental variables and ecological distance is not due to chance.

As for RDA, there is a partial CCA method. In this method, it is possible to remove the effect of one subset of variables first, and then analyse the effects of other variables.

Distance-based redundancy analysis and canonical analysis of principal coordinates

Distance-based redundancy analyis (db-RDA) and canonical analysis of principal coordinates (CAP) are constrained ordination techniques that analyse results of PCoA further. The advantage of these techniques is that any type of ecological distance can be analysed, including distances that are better for investigating differences in species composition as seen in chapter 8. Since db-RDA and CAP are constrained ordination techniques, the influence of environmental variables on differences in the selected good ecological distance can be investigated, and they therefore offer more advanced ways of constrained analysis than RDA and CCA. Although both db-RDA and CAP analyse the results of PCoA, the methods that they

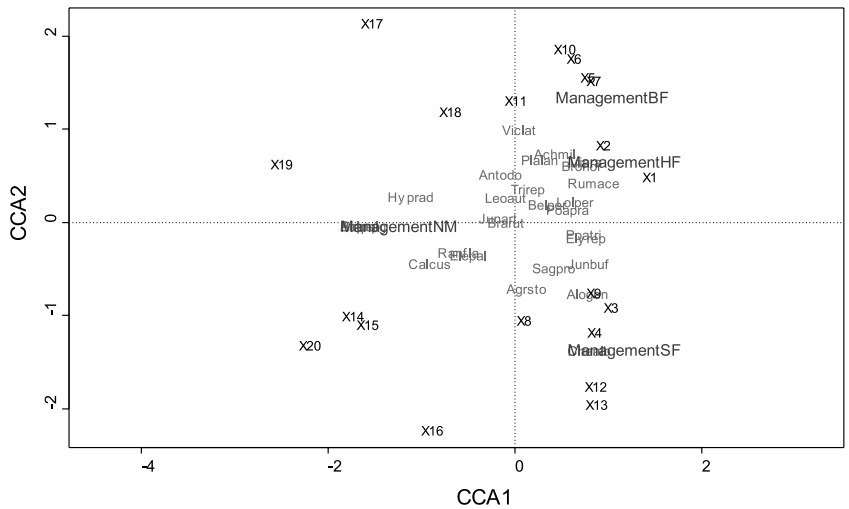


Figure 10.16 CCA ordination graph for the first two axes for the dune meadow dataset using scaling method 1 and management as constraining variable.

use are different: db-RDA uses RDA, whereas CAP uses linear discriminant analysis or canonical correlation analysis. A detailed discussion of the differences between both methods is beyond the scope of this chapter (see Legendre and Anderson 1999; Anderson and Willis 2003).

A db-RDA of the dune meadow dataset, using the type of management as a constraining variable, and the Bray-Curtis distance to express ecological distance gives the following result:

```
Partitioning of squared Bray distance:

Total         87.28
Constrained   28.50
Unconstrained 58.78

Eigenvalues, and their contribution to the squared Bray distance

             CAP1    CAP2    CAP3      PC1     PC2     PC3    PC4     PC5     PC6     PC7
lambda    17.0957 8.6126 2.7943 24.1862 9.2599 7.1894 5.891 3.9894 2.8640 1.6239
accounted  0.1959 0.2945 0.3265  0.2771 0.3832 0.4656 0.533 0.5787 0.6116 0.6302
             PC8     PC9    PC10    PC11    PC12     PC13     PC14
lambda    1.385 1.1412 0.6116 0.3267 0.1858 0.08386 0.04491
accounted 0.646 0.6591 0.6661 0.6698 0.6720 0.67294 0.67345
```

```
Scaling 1 for species and site scores
-- Sites are scaled proportional to eigenvalues
-- Species are unscaled: weighted dispersion equal on all dimensions

Species scores

           CAP1       CAP2      CAP3      PC1      PC2      PC3
Achmil  0.48210 -1.5325659 -0.459148 -0.64885 -0.19187 -1.38213
Agrsto  0.23882  3.6158066 -0.582905  2.26057 -0.05501 -0.12210
…
Trirep  0.53227 -2.3206841 -0.882848  0.27147  0.43253 -0.78951
Viclat  0.03078 -0.5633329 -0.911828 -0.07164  0.03791  0.53027

Site scores (weighted averages of species scores)

        CAP1    CAP2     CAP3      PC1      PC2      PC3
X1   1.10024 -0.3275 -0.50105 -1.16792 -1.16579  0.02871
X2   0.85167 -0.5197 -0.51581  0.05581 -0.23602 -0.26113
…
X19 -1.07753 -0.5148 -0.12207 -0.74980  0.89546  0.40619
X20 -1.29531  0.8673  0.17204  1.00331 -0.28929  0.22866

Site constraints (linear combinations of constraining variables)

       CAP1       CAP2      CAP3
X1   0.4100  0.5796560 -0.124936
X2   0.4034 -0.7544929 -0.397613
…
X19 -0.9646 -0.0001609  0.004098
X20 -0.9646 -0.0001609  0.004098

Biplot scores for constraining variables

                CAP1       CAP2     CAP3
ManagementHF  0.3872 -0.3126081  0.86739
ManagementNM -0.9999 -0.0002350  0.01051
ManagementSF  0.4250  0.8465954 -0.32035

Centroids for factor constraints

                CAP1       CAP2      CAP3
ManagementBF  0.4034 -0.7544929 -0.397613
ManagementHF  0.4235 -0.2426984  0.383574
ManagementNM -0.9646 -0.0001609  0.004098
ManagementSF  0.4100  0.5796560 -0.124936

Test for significance of all constrained eigenvalues
Pseudo-F:         2.26279
Significance:     0.006
Based on 1000 permutations under reduced model.
```

The output is very similar to the output of a RDA. The site, species, biplot and centroid scores should be analysed in a similar way. We can see that db-RDA expresses 32.6% of total squared distance. The first two axes express 29.4% of squared distance (the accounted values indicate the cumulated proportion).

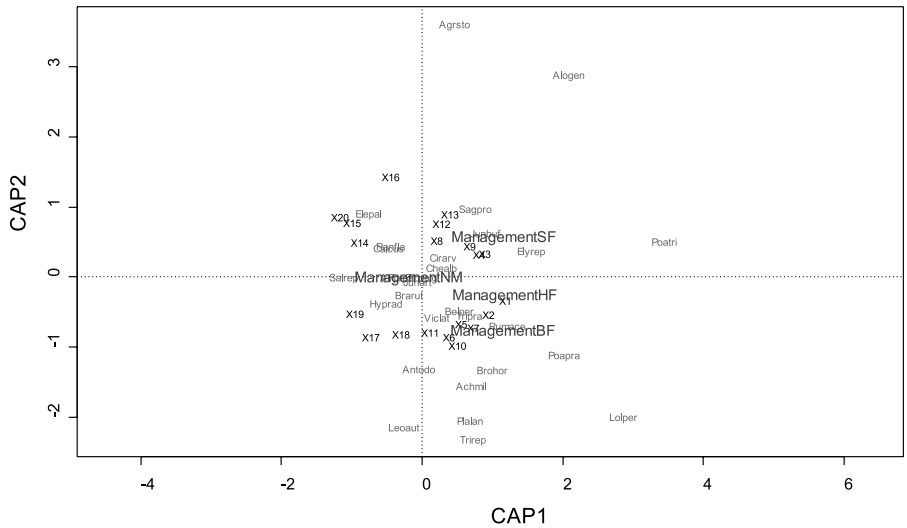


Figure 10.17 Db-RDA ordination graph for the first two axes for the dune meadow dataset using scaling method 1, the Bray-Curtis distance and the type of management as a constraining variable.

When we plot the site, species and centroid scores for the first db-RDA axes, then we obtain the ordination graph shown in Figure 10.17

As discussed for RDA, the species scores show the direction of higher abundance of a particular species. We thus expect higher abundance of *Alopecurus genitalicus* for sites X13 and X12, and lower abundance for X17, X18 and X19. The centroid scores show where sites of the same category of management are expected in the graph. These scores can be interpreted in a similar way as a site score. For instance, we expect that sites with standard farming will contain more *Agrostis stolonifera* and *Alopecurus genitalicus*.

Whereas we investigated the influence of a categorical variable on the differences in species composition above, you can also include continuous variables as explanatory variables. A db-RDA of the dune meadow dataset using the depth of the A1 horizon as the constraining variable and the Bray-Curtis distance to express ecological distance gives the result shown on the next page.

The results are very similar to those for management shown earlier. Because only one continuous explanatory variable was used, only one constrained ordination axis was obtained accounting for 15.8% of total squared Bray-Curtis distance.

Because only one constrained axis was obtained, an ordination graph was constructed that also used the first residual axis as shown in Figure 10.18. The vector for the depth of the A1 horizon indicates the direction in the graph where sites are expected that have deeper A1 horizon: these sites occur on the right-hand side of the graph. We therefore infer from the graph that sites that have a deeper A1 horizon are expected to contain more *Agrostis stolonifera* and *Eleocharis palustris*, whereas sites that have a more shallow A1 horizon would have more *Lolium perenne* and *Poa pratensis*.

```
Call:
capscale(formula = dune ~ A1, data = dune.env, distance = "bray")

Partitioning of squared Bray distance:

Total         87.28
Constrained   13.80
Unconstrained 73.49

             CAP1     PC1     PC2    PC3    PC4    PC5   PC6    PC7    PC8
lambda    13.7975 21.6515 18.8817 8.7678 6.8894 4.5809 4.083 3.2127 1.8245
accounted  0.1581  0.2481  0.4644 0.5648 0.6438 0.6962 0.743 0.7798 0.8007
            PC9   PC10   PC11   PC12   PC13    PC14
lambda    1.414 1.1725 0.5201 0.3357 0.1033 0.05102
accounted 0.817 0.8304 0.8363 0.8402 0.8413 0.84193

Scaling 1 for species and site scores
-- Sites are scaled proportional to eigenvalues
-- Species are unscaled: weighted dispersion equal on all dimensions

Species scores

           CAP1      PC1      PC2      PC3      PC4      PC5
Achmil -0.67633 -1.04282  0.27838 -0.28142  0.95411 -0.65458
Agrsto  1.64726  2.89575  0.65578  0.51099  0.68881  0.23853
…
Trirep  0.21369 -0.86349  0.61722  0.31855 -0.82554  0.60036
Viclat -0.18672 -0.18742 -0.13515 -0.27223 -0.66189  0.35546

Site scores (weighted averages of species scores)

        CAP1      PC1      PC2      PC3      PC4      PC5
X1  -1.19851 -0.13951  0.83107 -0.92749  0.30446  0.18771
X2  -0.80470 -0.33601  0.73210 -0.09861  0.08395  0.49669
…
X19 -0.06976 -0.23731 -1.47115  0.64386 -0.22138  0.47097
X20  1.36948  1.49536 -1.20732 -0.43657  0.05829 -0.27214

Site constraints (linear combinations of constraining variables)

        CAP1
X1  -0.54756
X2  -0.36059
…
X19 -0.30717
X20 -0.36059

Biplot scores for constraining variables

   CAP1
A1    1

Test for significance of all constrained eigenvalues
Pseudo-F:        2.26279
Significance:    0.006
Based on 1000 permutations under reduced model.
```

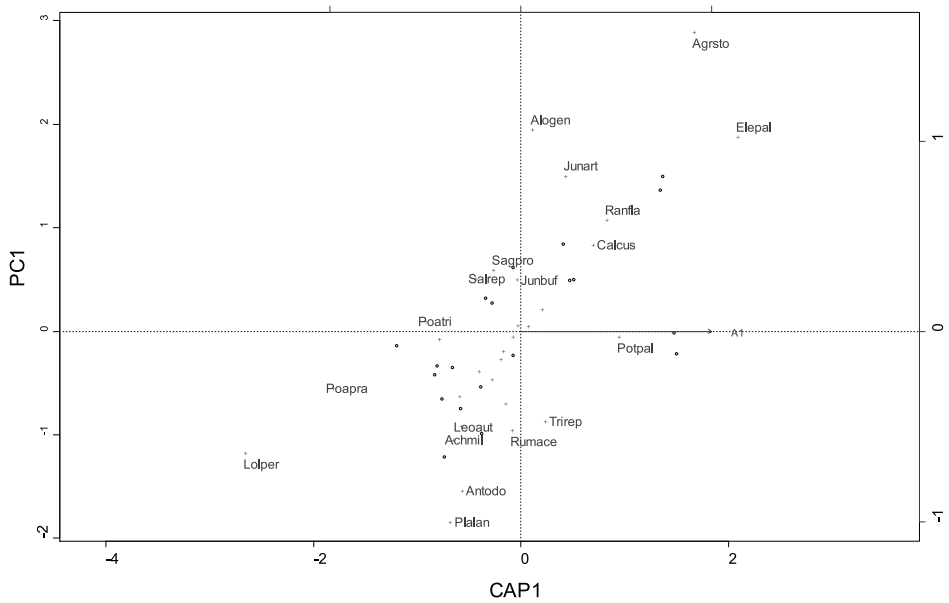


Figure 10.18 Db-RDA ordination graph for the first constrained and first residual axis for the dune meadow dataset using scaling method 1, the Bray-Curtis distance and the depth of the A1 horizon as a constraining variable. Sites are plotted as circles and species as crosses.

Choice of ordination method

As we saw in chapter 8, different ecological distance measures have different properties. The properties of the ordination method you use therefore depend on the properties of the distance measure on which it is based. The first rule should be that you use a distance measure that is a useful ecological distance measure. A good analysis practice is also to repeat the analysis with several good distance measures and investigate whether all these analyses lead to the same conclusion.

There are some methods to investigate how well the distances in the ordination graph represent the total distances as provided in a distance matrix. The first method calculates the percentage of variance that is displayed in the graph. The second method compares the ecological distance between sites with the distance between the positions of the sites in the ordination graph. The distances can be plotted against one another and so can you can check how good the correlation is. Figure 10.19 shows the correspondence between the Bray-Curtis distance on the first two axes of a PCoA (as shown in Figure 10.8), and the total Bray-Curtis distance among sites. You can see that the overall correlation is quite good, but that some sites are plotted relatively closer than other sites despite the same total distance.

A third method of investigating how well the distances in the ordination graph represent the total distance combines clustering results with ordination results. It is recommended to plot the results of single linkage (see Chapter 9) on top of an ordination graph. The results of single linkage produce a minimum spanning tree that shows how sites and clusters can be joined together by the smallest possible total length of segments among them. Figure 10.20 gives an example for the first two axes of PCoA. For example, sites X20 and X15 are plotted close together in the graph and were also joined by the minimum spanning tree. This is an indication that the graph provides a good representation of their ecological distance.

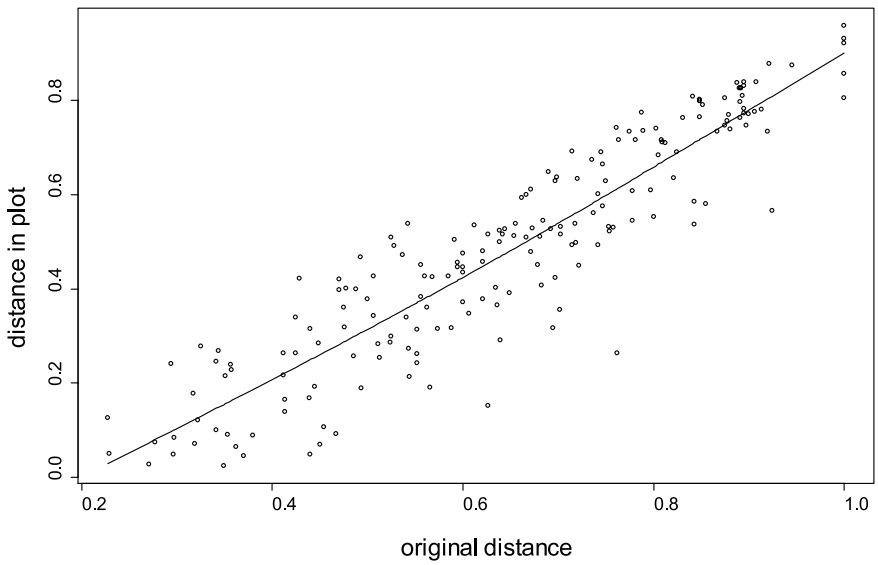


Figure 10.19 Relationship between distances between site positions in an ordination graph (Figure 10.8) and total Bray-Curtis distance between sites. The line shows the fit of a GAM (see chapters 6 and 7) between the original distances and the distances in the ordination graph.

If sites that had a small distance in the graph were joined through the minimum spanning tree at far distance in the graph, such as sites X1 and X2, this means that they are only joined much later in the process. The distance between X1 and X2 is thus not well presented in the graph. A better ordination graph will have a shorter length of the minimum spanning tree.

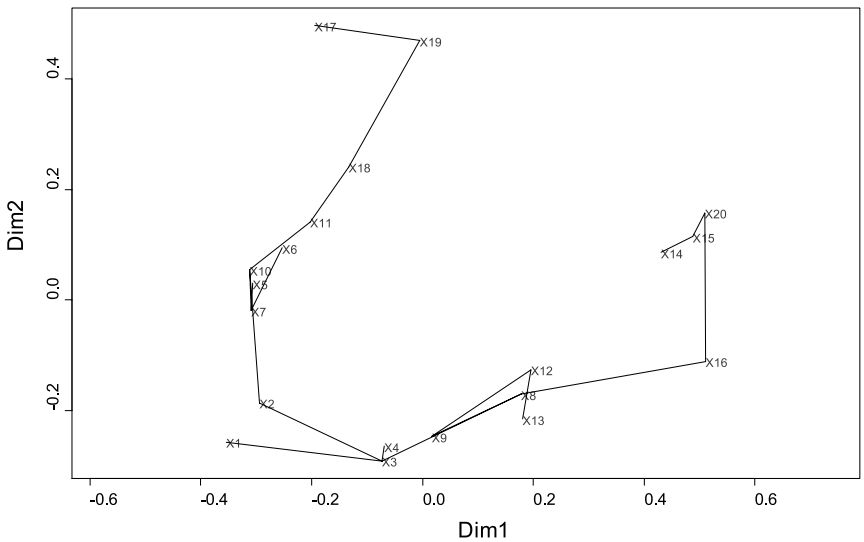


Figure 10.20 Plotting the cluster structure on top of an ordination graph (Figure 10.8) by a minimum spanning tree to investigate how well ecological distance is represented in the ordination graph.

Further interpretation of ordination graphs by indirect gradient analysis

As with constrained ordination, indirect gradient analysis methods also seek to understand the relationship between environmental variables of a site and their species composition. They are applied *after* an unconstrained ordination analysis. The key idea is to try to relate the pattern of sites in the ordination graph to environmental variables.

There are three methods of investigating quantitative environmental variables.

The first method calculates a fitted vector of an environmental variable with the ordination configuration. This vector shows the direction in the ordination graph where sites are expected with values that are higher than average for the environmental variable. This is a similar approach as the one shown earlier for calculating correlation scores for a species for a PCoA or NMS. The interpretation is also similar (see also Figure 10.4).

When you calculate the vector scores for the depth of the A1 horizon and the first two axes of a PCoA based on the Bray-Curtis distance for the dune meadow dataset, then you obtain the result shown below.

The scores for the head of the vector (listed in the result for Dim1 and Dim2) can be used to plot a vector for the environmental variable onto the ordination graph. Figure 10.21 shows how the results presented above can be presented graphically. You can see that we expect greater depth of the A1 horizon on sites X14, X15, X16 and X20.

The second method is to plot the values of the environmental variable as a bubble graph. This approach is more general than fitting a vector, as this method does not assume that the values will increase linearly on an axis. Figure 10.22 shows the bubble graph of the depth of the A1 horizon. Large bubbles indicate a larger value for the A1 horizon. The graph indicates that in general depth of the A1 horizon increases with Dim1, but there are exceptions. We can see large values for X14 and X16, but the value of X20 is not so large. There is no sign of depth of A1 changing with Dim2 except for X14 and X15. This example shows how the vector (Figure 10.21) picks out trends but does not show any detail of deviation from the trend.

The third method of investigating a quantitative environmental variable is to fit a surface that models how the environmental variable changes over the ordination graph. Rather than simply plotting the bubbles, we can try to describe the pattern in bubble size by a smooth surface, using a GAM as described in chapter 7.

When calculating the surface for the depth of the A1 horizon for the PCoA based on the Bray-Curtis distance for the dune meadow dataset, you obtain the result presented in Figure 10.23. This figure shows a similar picture as Figure 10.22, with increasing values from left to right in the ordination graph (the GAM approach can reveal more complex patterns, but in this case the algorithm fitted a linear trend). The lower value for site X20 is not reflected. If the residuals from the fitted surface were plotted (figure not included), then it would be clear that X20 is not represented well in the Figure 10.23. It is a good statistical practice to investigate residuals.

For categorical environmental variables, there are some other methods of indirect gradient analysis.

The first method is to use a different symbol for each category of the environmental variable. Figure 10.24 gives an example for the type of management for the PCoA ordination for the dune meadow dataset. We can see that all sites with nature management are plotted at the top

```
      Dim1    Dim2     r2 Pr(>r)
A1 0.98806 0.15404 0.3845 < 0.01 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
P values based on 100 permutations.
```

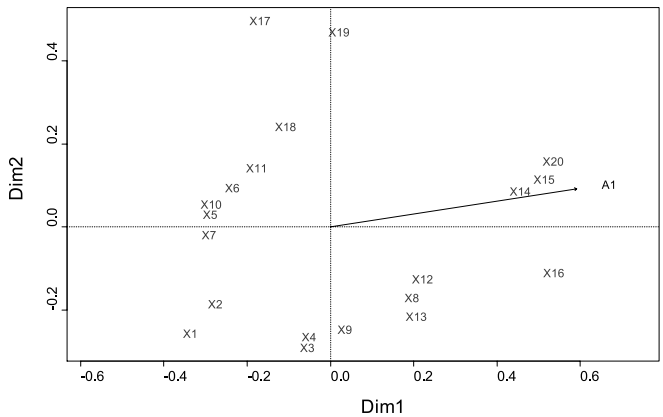


Figure 10.21 Plotting a vector for a quantitative environmental variable onto an ordination graph (Figure 10.8).

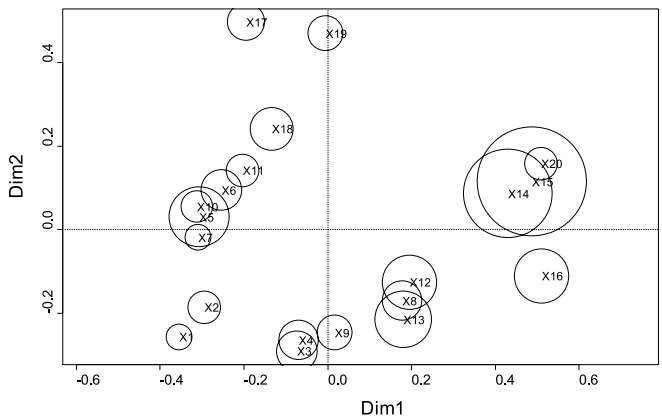


Figure 10.22 Plotting a bubble graph for a quantitative environmental variable onto an ordination graph (Figure 10.8).

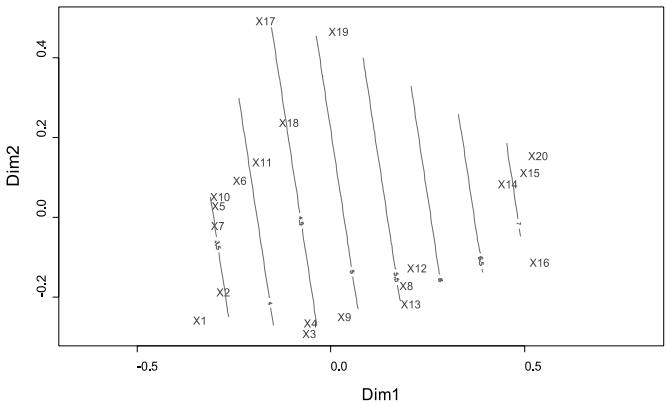


Figure 10.23 Plotting a contour for a quantitative environmental variable onto an ordination graph (Figure 10.8).

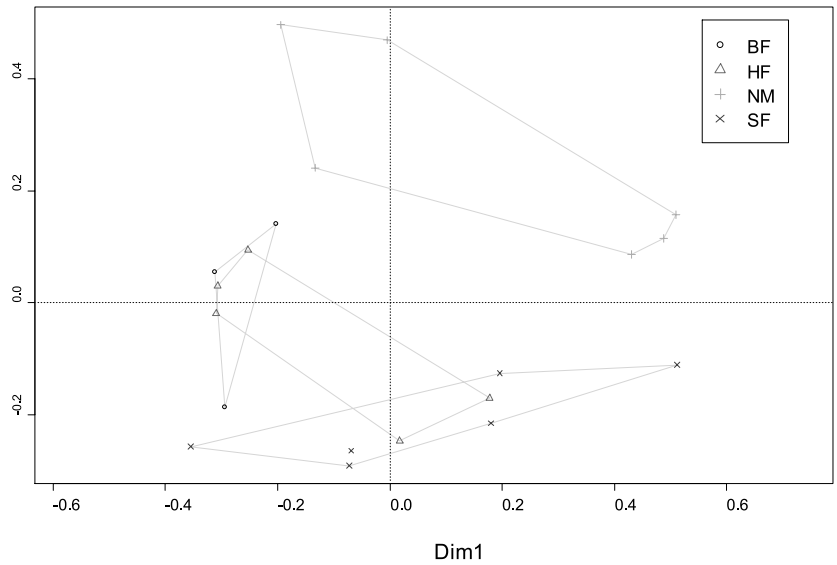


Figure 10.24 Plotting different symbols for different categories of a categorical environmental variable onto an ordination graph (Figure 10.8). A convex hull encloses all sites of the same category.

part of the graph. We can also see that some sites with hobby farming are more similar in species composition to sites with biological farming, whereas other sites with hobby farming are more similar in species composition to sites with standard farming.

The second method for categorical environmental variables is to calculate the average plotting position for each category (the centroids). When you calculate the average positions for management for the dune meadow dataset and the PCoA graph, then you obtain the following result:

```
Centroids:
       Dim1    Dim2
PBF -0.2694  0.0032
PHF -0.1350 -0.0623
PNM  0.1822  0.2609
PSF  0.0650 -0.2106

Goodness of fit:
      r2 Pr(>r)
P 0.4482   0.01 **
```

These centroid positions can be plotted onto the ordination graph (Figure 10.25). By connecting each site with the centroid of the same category (a spiderplot), you can investigate whether some sites are outliers. We can observe again that sites with nature management have a different species composition, and that sites with hobby farming are either more similar to sites with standard farming or sites with biological farming. Since the convex hulls (Figure 10.24) are more sensitive to outliers, you need to be careful in making conclusions that species composition is similar when convex hulls overlap. When convex hulls do not overlap, this provides evidence that species composition is dissimilar.

A third method is to calculate confidence ellipses that predict where sites of a certain category will occur. This method estimates a confidence interval for sites of each category, using the positions of the sites on the horizontal and vertical axes as input variables. This approach is thus more sophisticated than the previous methods for categorical variables.

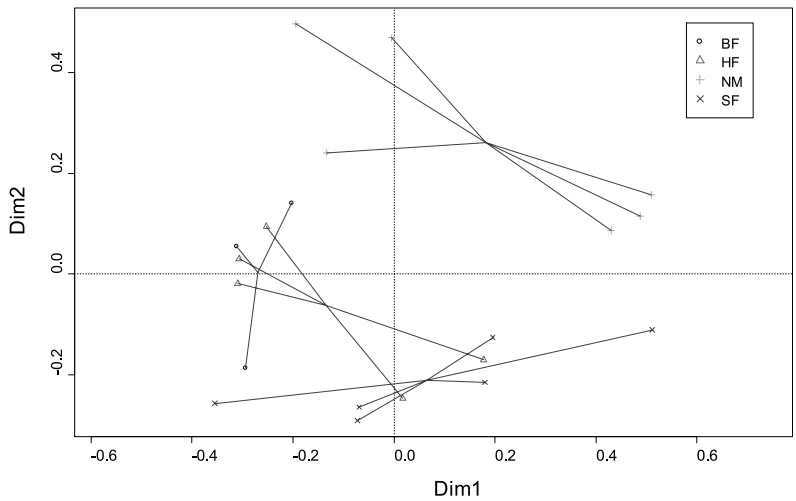


Figure 10.25 Connecting sites to the centroid of each category onto an ordination graph (Figure 10.8).

Figure 10.26 gives an example for the type of management for the PCoA ordination for the dune meadow dataset. The ellipses indicate where 95% of sites of the same category are expected to occur. Different symbols were also used for the different categories. You can see again that sites with nature management occur at the top of the graph, and that there is an overlap for sites with hobby farming and the other two categories of standard farming and biological farming.

Further interpretation of ordination graphs for individual species

By analogy with indirect gradient analysis methods, patterns of some individual species can be analysed *after* an ordination analysis. You can use the ordination method to check whether sites are different in species composition, and then check for the species that contribute most to the differences.

Figure 10.27 shows the results of the CAP

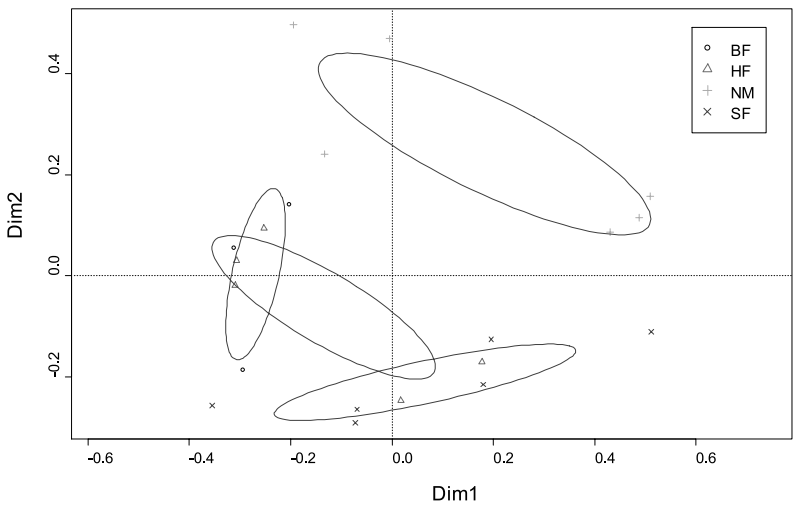


Figure 10.26 Drawing confidence ellipses for each category onto an ordination graph (Figure 10.8).

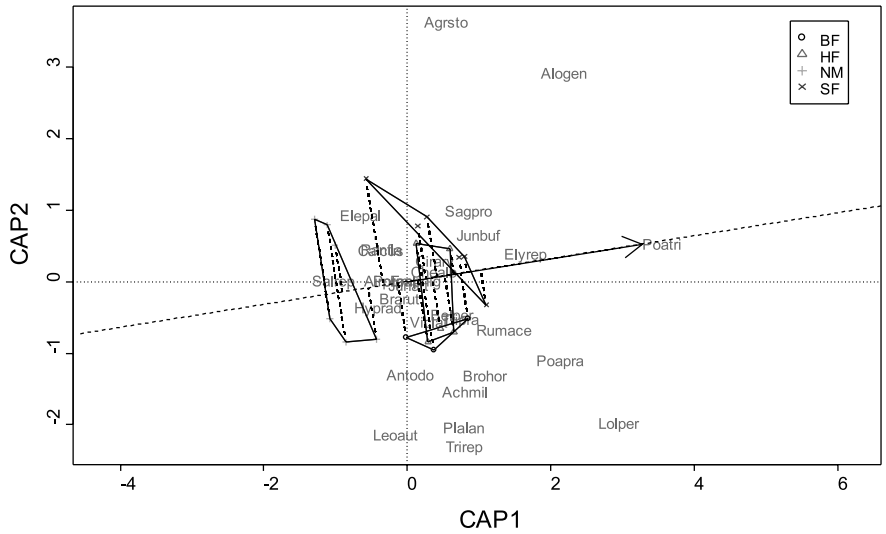


Figure 10.27 Investigating for important species that contribute to the differences in species composition. The first axes of a CAP analysis of the dune meadow dataset based on the Bray-Curtis distance and with type of management as explanatory factor.

analysis for the dune meadow dataset (based on the Bray-Curtis distance) for differences in species composition related to differences in management. The results of this analysis were provided above (including Figure 10.17). Added to the earlier results are the hulls for the different management categories, symbols for the different categories and the interpretation for species *Poa trivialis*. Since this species has a long species vector, we expect that it contributes to differences in species composition between types of management. We can also expect that abundances for this species will be lower for nature management, since the projected scores for this species are lower for this type of management.

The formal way of testing for the differences for *Poa trivialis* for the different types of management is a regression analysis, as seen in chapter 7. A GLM regression with log link and quasipoisson variance functions gives the result shown on the next page.

We can see from the regression coefficients that fewer individuals of *Poa trivialis* are predicted for nature management (checking the dune meadow datasets reveals that the species does not occur on the six quadrats with nature management). The ANOVA provides evidence for differences among categories for management. The standard errors and significance levels for the regression coefficients are large, however. What is happening? The reason is that the sample size is quite small in comparison to the four categories of management to investigate differences for individual species. In this case, we can observe actual differences in abundance, but we can not confirm that these differences were not observed by chance. This means that we could demonstrate that the sites are different in composition, but could not check for an individual species. This result is not entirely surprising, since the occurrences of species are correlated with each other. As we investigate several species at the same time in an ordination analysis, the investigation becomes more powerful.

```
glm(formula = Poatri ~ Management, family = quasipoisson(link = log),
    data = dune.env, na.action = na.exclude)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.708013  -0.331340  -0.000091   0.157273   1.776335

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.2993     0.2908   4.468 0.000388 ***
Management[T.HF]    0.2693     0.3512   0.767 0.454260
Management[T.NM]  -20.6019  3711.5165  -0.006 0.995640
Management[T.SF]    0.2412     0.3432   0.703 0.492330

(Dispersion parameter for quasipoisson family taken to be 0.9301272)

    Null deviance: 63.412  on 19  degrees of freedom
Residual deviance: 17.842  on 16  degrees of freedom


Analysis of Deviance Table

           Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                         19     63.412
Management  3   45.570        16     17.842 16.331 3.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

References

Anderson MJ and Willis TJ. 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* 84, 511–525.

Borcard D, Legendre P and Drapeau P 1992. Partialling out the spatial component of ecological variation. *Ecology* 73: 1045-1055.

Gotelli NJ and Ellison AM. 2004. *A primer of ecological statistics.* Sunderland: Sinauer Associates.

Jongman RH, ter Braak CJF and Van Tongeren OFR. 1995. *Data analysis in community and landscape ecology.* Cambridge: Cambridge University Press.

Kent M and Coker P. 1992. *Vegetation description and analysis: a practical approach.* London: Belhaven Press.

Legendre P and Anderson MJ. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* 69: 1-24.

Legendre P and Gallagher ED. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271-280.

Legendre P and Legendre L. 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV. (recommended as first priority for reading)

Makarenkov V and Legendre P. 2002. Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. *Ecology* 83: 1156-1161.

McGarigal K, Cushman S and Stafford S. 2000. *Multivariate statistics for wildlife and ecology research.* New York: Springer-Verlag.

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press.

Shaw PJA. 2003. *Multivariate statistics for the environmental sciences.* London: Hodder Arnold.

ter Braak CJF. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67: 1167-1179.

Doing the analyses with the menu options of Biodiversity.R

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 meadow dataset

Calculating a principal component analysis (PCA):

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: PCA (or PCA (prcomp))

→scaling: 1

→Plot method: ordiplot

→Plot method: text sites

→Plot method: text species

→Plot method: equilibrium circle

Calculating a PCA on a transformed matrix:

Biodiversity > Community matrix > Transform community matrix…

→Method: Hellinger

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: PCA

Conducting a principal coordinates analysis (PCoA)

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: PCoA (or PCoA (Caillez))

→Distance: bray

Calculating a non-metric multidimensional scaling (NMS)

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: NMS (or NMS (standard))

→Distance: bray

→NMS axes: 2

→NMS permutations: 100

Calculating a correspondence analysis (CA or WA)

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: CA

→scaling: 1

Calculating a redundancy analysis (RDA):

Biodiversity > Analysis of ecological distance > Constrained ordination…

→Ordination method: RDA

→scaling: 1

→permutations: 100

→Explanatory: Management

Calculating a canonical correspondence analysis (CCA)

Biodiversity > Analysis of ecological distance > Constrained ordination…

→Ordination method: CCA

→scaling: 2

→permutations: 100

→Explanatory: Management

Calculating distance-based redundancy analysis (db-RDA)

Biodiversity > Analysis of ecological distance > Constrained ordination…

→Ordination method: capscale

→distance: bray

→permutations: 100

→Explanatory: Management

Calculating the correlation between distance in an ordination graph and total distance

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: PCoA

→Distance: bray

→Plot method: ordiplot

→Plot method: distance displayed

Plotting clustering results onto an ordination graph:

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: PCoA

→Distance: bray

→Plot method: ordiplot

→Plot method: ordicluster

Plotting quantitative environmental variables onto an ordination graph:

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: PCoA

→Distance: bray

→Plot variable: A1

→Plot method: ordiplot

→Plot method: vectorfit

→Plot method: ordibubble

→Plot method: ordisurf

Plotting categorical environmental variables onto an ordination graph:

Biodiversity > Analysis of ecological distance > Unconstrained ordination…

→Ordination method: PCoA

→Distance: bray

→Plot variable: Management

→Plot method: ordiplot

→Plot method: factorfit

→Plot method: ordihull

→Plot method: ordispider

→Plot method: ordiellipse

→Plot method: ordisymbol

Doing the analyses with the command options of Biodiversity.R

Calculating a principal component analysis (PCA)

```
Ordination.model1 <- rda(dune)
summary(Ordination.model1, scaling=1)
plot1 <- ordiplot(Ordination.model1, scaling=1, type="text")
plot2 <- ordiplot(Ordination.model1, scaling=2, type="text")
```

Calculating the variance of each species of the species matrix

```
inertcomp(Ordination.model1, display='species',
  statistic='explained', proportional=F)
```

Calculating the proportion of variance explained for an ordination graph

```
goodness(Ordination.model1, display='sites', choices=c(1:2),
  statistic='explained')
```

Adding a vector and perpendicular lines for a particular species to an ordination plot

```
ordivector(plot1,"Agrsto",lty=2)
```

Calculating correlations among vectors:

```
cor.test(dune[,"Alogen"],dune[,"Agrsto"])
```

Calculating the number of ecologically meaningful principal components:

```
PCAsignificance(Ordination.model1,axes=30)
```

Drawing an equilibrium circle

```
ordiequilibriumcircle(Ordination.model1,plot1)
```

Calculating a PCA on a transformed matrix

```
Community.1 <- disttransform(dune, method='Hellinger')
Ordination.model2 <- rda(Community.1)
summary(Ordination.model2, scaling=1)
plot3 <- ordiplot(Ordination.model2, scaling=1, type="text")
```

Calculating a principal coordinates analysis (PCoA)

```
distmatrix <- vegdist(dune,method='bray')
Ordination.model3 <- cmdscale(distmatrix, k=nrow(dune)-1,
  eig=T, add=F)
Ordination.model3 <- add.spec.scores( Ordination.model3, dune,
  method='pcoa.scores', Rscale=T, scaling=1, multi=1)
plot4 <- ordiplot(Ordination.model3, type='text')
```

Calculating a non-metric multidimensional scaling (NMS)

```
distmatrix <- vegdist(dune, method='bray')
initNMS <- NMSrandom(distmatrix, perm=100, k=2)
Ordination.model4 <- postMDS(initNMS, distmatrix)
Ordination.model4 <- add.spec.scores( Ordination.model4, dune,
  method='wa.scores')
Ordination.model4
plot5 <- ordiplot(Ordination.model4)
```

Calculating a correspondence analysis (CA or WA)

```
Ordination.model5 <- cca(dune)
summary(Ordination.model5, scaling=1)
plot6 <- ordiplot(Ordination.model5, type='text', scaling=1)
```

Calculating a redundancy analysis (RDA)

```
Ordination.model6 <- rda(dune ~ Management, dune.env)
summary(Ordination.model6, scaling=1)
permutest.cca(Ordination.model6, permutations=1000)
plot7 <- ordiplot(Ordination.model1, type='text', scaling=1)
```

Calculating a canonical correspondence analysis (CCA)

```
Ordination.model7 <- cca(dune ~ Management, dune.env)
summary(Ordination.model7, scaling=2)
permutest.cca(Ordination.model7, permutations=1000)
plot8 <- ordiplot(Ordination.model7, type='text', scaling=1)
```

Calculating distance-based redundancy analysis (db-RDA)

```
Ordination.model8 <- capscale(dune ~ Management, dune.env)
summary(Ordination.model8, scaling=1)
permutest.cca(Ordination.model8, permutations=1000)
plot9 <- ordiplot(Ordination.model8, type='text', scaling=1)
```

Calculating the correlation between distance in an ordination graph and total distance

```
distdisplayed(dune, plot4, distx='bray')
```

Plotting clustering results onto an ordination graph

```
distmatrix <- vegdist(dune, method='bray')
cluster <- hclust(distmatrix, method='single')
ordicluster(plot4, cluster)
```

Plotting quantitative environmental variables onto an ordination graph

```
fitted <- envfit(plot4, A1, permutations=100)
plot(fitted)
fitted <- vectorfit(plot4, A1, permutations=100)
ordibubble(plot4, A1)
ordisurf(plot4, A1)
```

Plotting categorical environmental variables onto an ordination graph

```
ordisymbol(plot4, dune.env, 'Management', legend=T)
fitted <- envfit(plot4, Management, permutations=100)
plot(fitted)
fitted <- factorfit(plot4, Management)
ordihull(plot4, Management)
ordispider(plot4, Management)
ordiellipse(plot4, Management)
```

Effective data analysis requires familiarity with basic concepts and an ability to use a set of standard tools, as well as creativity and imagination. *Tree diversity analysis* provides a solid practical foundation for training in statistical methods for ecological and biodiversity studies.

This manual arose from training researchers to analyse tree diversity data collected on African farms, yet the statistical methods can be used for a wider range of organisms, for different hierarchical levels of biodiversity and for a variety of environments — making it an invaluable tool for scientists and students alike.

Focusing on the analysis of species survey data, *Tree diversity analysis* provides a comprehensive review of the methods that are most often used in recent diversity and community ecology literature including:

- Species accumulation curves for site-based and individual-based species accumulation, including a new technique for exact calculation of site-based species accumulation.
- Description of appropriate methods for investigating differences in diversity and evenness such as Rényi diversity profiles, including methods of rarefaction to the same sample size for different subsets of the data.
- Modern regression methods of generalized linear models and generalized additive models that are often appropriate for investigating patterns of species occurrence and species counts.
- Methods of ordination for investigating community structure and the influence of environmental characteristics, including recent methods such as distance-based redundancy analysis and constrained analysis of principal coordinates.

The manual also introduces a powerful new software programme, *Biodiversity.R*, that is capable of performing all the statistical analyses described in the book. The software is built using the free R language and environment for statistical computing, and several of its libraries such as the vegan community ecology package and the R-commander graphical user interface. The software is provided on CD.

Roeland Kindt is an ecologist with interests in community ecology, conservation, ecoagriculture and training. He is currently leading a World Agroforestry Centre project on enhanced utilization of tree diversity at the landscape level.

Richard Coe is a biometrician and has more than 25 years experience in providing research support to scientists in the fields of ecology, agroforestry and natural resource management. He is leading the Research Support Unit of the World Agroforestry Centre.

VVOB

DFID Department for International Development

USAID

Ministerie van Buitenlandse Zaken

World Agroforestry Centre, United Nations Avenue, Gigiri, PO Box 30677- 00100 Nairobi, Kenya
Tel: +254 20 7224000 • Fax: +254 20 7224001 or via USA: Tel: +1 650 833 6645 • Fax: +1 650 833 6646
Email: icraf@cgiar.org • www.worldagroforestry.org

Barcode
ISBN92 9059 179 X