Data Analysis in Vegetation Ecology

Otto Wildi

WILEY-BLACKWELL

Data Analysis in Vegetation Ecology

Otto Wildi

WSL Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, Switzerland

WILEY-BLACKWELL

A John Wiley & Sons, Ltd., Publication

Data Analysis
in Vegetation Ecology

Data Analysis in Vegetation Ecology

Otto Wildi

WSL Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, Switzerland

WILEY-BLACKWELL

A John Wiley & Sons, Ltd., Publication

This edition first published 2010

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Other Editorial Offices:
9600 Garsington Road, Oxford, OX4 2DQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloguing-in-Publication Data

Wildi, Otto.
Data analysis in vegetation ecology / Otto Wildi.
p. cm.
ISBN 978-0-470-66102-4 (pbk.)
1. Plant communities–Data processing. 2. Plant communities–Mathematical models. 3. Plant ecology–Data processing. 4. Plant ecology – Mathematical models. I. Title.
QK911.W523 2010
581.70285 – dc22
2010004441

ISBN: 9780470661017 (HB) 9780470661024 (PB)

A catalogue record for this book is available from the British Library.

Typeset in 10.5/13 Times by Laserwords Private Limited, Chennai, India
Printed in Great Britain by TJ International, Padstow, Cornwall.

First printing 2010

Contents

Plants are so unlike people that it's very difficult
for us to appreciate fully their complexity and sophistication.

Michael Pollan, *The Botany of Desire*

Preface

When starting to rearrange my lecture notes I had a ‘short introduction to multivariate vegetation analysis’ in mind. It ended up as a ‘not so short introduction’. The book now summarizes some of the well-known methods used in vegetation ecology. The matter presented is but a small selection of what is available to date. By focusing on methodological issues I try to explain what plant ecologists do, and why they measure and analyse data. Rather than just generating numbers and pretty graphs, the models and methods I discuss are a contribution to the understanding of the state and functioning of the ecosystems analysed. But because researchers are usually driven by their curiosity about the functioning of the systems I successively began to integrate examples encountered in my work. These now occupy a considerable portion of this book. I am convinced that the fascination of research lies in the perception of the real world and its amalgamation in the form of high-quality data with hidden content processed by a variety of methods reflecting our model view of the world. Neither my results nor my conclusions are final. Hoping that the reader will like some of my ideas and perspectives, I encourage them to use and to improve on them. There remains considerable scope for innovation.

The examples presented in this book all come from Central Europe. While this was not intended originally, I became convinced the topics they cover are of general relevance, as similar investigations exist almost everywhere in the world. An example is the pollen data set: pollen profiles offer the unique chance to study vegetation change over millennia. This is the time scale of processes such as climate change and the expansion of the human population. Another, much shorter time series than that of pollen data is found in permanent plot data originating from the Swiss National Park that I had the opportunity to look at. The unique feature of this is that it dates back to the year 1917, when Josias Braun-Banquet personally installed the first wooden poles, which are still in place. Records of the full set of species have been collected ever since in five-year steps. A totally different data set comes from the Swiss Forest Inventory, presented in the last chapter of this book. Whereas many vegetation surveys are merely preferential collections of plot data, this data set is an example of systematic sampling on a grid encompassing huge environmental gradients. It helps to assess which patterns really exist, and whether some of those described in papers or textbooks are real or merely reflect the imagination or preference of researchers scanning the landscape for nice locations. In this case the data set available for answering the question is still moderate in size, but handling of large data sets will eventually be needed in similar contexts. I used the Swiss wetland data set as an example for handling data of much larger size, in this case with n = 17608 relevés. Although this is outnumbered by others, it resides on a statistical sampling design.

Some basic knowledge of vegetation ecology might be needed to understand the examples presented in this book. Readers wishing to acquire this are advised to refer, for example, to the comprehensive volumes *Vegetation Ecology* by Eddy van der Maarel (2005) and *Aims and Methods of Vegetation Ecology* by Mueller-Dombois and Ellenberg (1974), presently available as a reprint. The structure of my book is influenced by Orlócis (1978) *Multivariate Analysis in Vegetation Research*, which I explored the first time when proofreading it in 1977. Various applications are found in the books of Gauch (1982), Pielou (1984) and Digby and Kempton (1987) and many multivariate methods used in vegetation ecology are introduced in Jongman *et al*. (1995). To study statistical methods used in this book in more detail, I strongly recommend the second edition of Numerical Ecology by Legendre and Legendre, probably the most comprehensive textbook existing today. Several books provide an introduction to the use of statistical packages, which are referred to in the appendix. For many reasons I decided to omit the software issue in the main text; upon the request of several reviewers I added a section to the appendix where I reveal how I calculated my examples and mention programs, program packages and databases.

I would like to express my thanks to all individuals that have contributed to the success of this book. First of all Rachel Wade from Wiley-Blackwell, who strongly supported the efforts to print the manuscript in time and organized all the technical work. I thank Tim West for careful copy-editing, and Robert Hambrook for managing the production process. My colleagues Anita C. Risch and Martin Schütz revised the entire text, providing corrections and suggestions. Meinrad Küchler helped in the computation of several examples. André F. Lotter provided the pollen data set. I cannot remember all the people who had an influence on the point of view presented here:

many ideas came from László Orlóci through our long lasting collaboration, others from Madhur Anand, Enrico Féoli, Valério de Patta Pillar, Janos Podani and Helene Wagner. I particularly thank my family for encouraging me to tackle this work and for their tolerance when I was working at night and on weekends to get it completed.

Otto Wildi
Birmensdorf, 1 December 2009

List of figures

List of tables

1 Introduction

This book is about understanding vegetation systems in a scientific context, one topic of vegetation ecology. It is written for researchers motivated by the curiosity and ambition to assess and understand vegetation dynamics. Vegetation, according to van der Maarel (2005) 'can be loosely defined as a system of largely spontaneously growing plants.' What humans grow in gardens and fields is hence excluded. The fascination of investigating vegetation resides in the mystery of what plants 'have in mind' when populating the world. The goal of all efforts in plant ecology, as in other fields of science, is to learn more about the rules governing the world. These rules are causing patterns, and the assessment of patterns is the recurrent theme of this book.

Unfortunately, our access to the *real world* is rather restricted and – as we know from experience – differs among individuals. To assure progress in research an image of the real world is needed: the *data world*. In this we get a description of the real world in the form of numbers. (An image can

Data Analysis in Vegetation Ecology Otto Wildi

be a spreadsheet filled with numbers, a digital photograph or a digital terrain model.) Upon analysis we then develop our *model world*, which represents our understanding of the real world. Typical elements are orders, patterns or processes governing systems. It is the aim of most analytical methods to identify patterns as elements of our model view.

Finding models reflecting the real world is a difficult task due to the complexity of systems. Complexity has its origin in a number of fairly well known phenomena, one being the scale effect. Any regularity in ecosystems will emerge at a specific spatial and temporal scale only: at short spacial distance competition and facilitation among plants can be detected (Connell & Slatyer 1977); these would remain undetected over a range of kilometres. In order to study the effect of global climate change (Orlóci 2001, Walther *et al.* 2002) the scale revealed by satellite photographs is probably more promising. Choosing the best scale for an investigation is a matter of decision, experience and often trial and error. For this a multi-step approach is needed, in which intermediate results are used to evaluate the next decision in the analysis. Poore (1955, 1962) called this *successive approximation* and Wildi & Orlóci (1991) *flexible analysis*. Hence, the variety and flexibility of methods is nothing but an answer to the complex nature of the systems. Once the proper scale is found there is still a need to consider an 'upper' and a 'lower' level of scale, because these usually also play a role. Parker & Pickett (1998) discuss this in the context of temporal scales and interpret the interaction as follows: 'The middle level represents the scale of investigation, and processes of slower rate act as the context and processes of faster rates reflect the mechanisms, initial conditions or variance.'

A second source of complexity is uncertainty in data measured. Data are restricted by trade-offs and practical limitations. A detailed vegetation survey is time-consuming, and while sampling, vegetation might already be changing (Wildi *et al.* 2004). Such data will therefore exhibit an undesired temporal trend. A specific bias causes variable selection. It is easier to measure components above ground than below ground (van der Maarel 2005, p. 6), a distinction vital in vegetation ecology. Once the measurements are complete they may reflect random fluctuation or chaotic behaviour (Kienast *et al.* 2007) while failing to capture deterministic components. It is a main objective in data analysis to distinguish random from deterministic components. Even if randomness is controlled there is *nonlinearity* in ecological relationships, a term used when linearity is no longer valid. This would not be a problem if we knew the kind of relationship that was hidden in the data (e.g. Gaussian, exponential, logarithmic, etc.), but finding a proper function is usually a challenging task.

Further, spatial and temporal interactions add to the complexity of vegetation systems. In space, the problem of order arises, as the order of objects depends on the direction considered. In most ecosystems, the environmental conditions, for example elevation or humidity, change across the area. Biological variables responding to this will also be altered and become *space-dependent* (Legendre & Legendre 1998). If there is no general dependency in space, a local phenomenon may exist: *spatial autocorrelation*. This means that sampling units in close neighbourhood are more similar than one could expect from ecological conditions. One cause for this comes from biological population processes: the chance that an individual of a population will occur in unfavourable conditions is increased if another member of the same population resides nearby. It will be shown later in this book how such a situation can be detected (Section 7.3.3). Similarly, correlation over time also occurs. In analogy to space, there is *temporal dependence* and *temporal autocorrelation*. This comes from the fact that many processes are temporally continuous. The systems will usually only change gradually, causing two subsequent states to be similar. Finally, time and space are not independent, but linked. Spatial patterns tend to change continuously over time. In terms of autocorrelation, spatial patterns observed within a short time period are expected to be similar. Similarly, a time series observed at one point in space will be similar to another series observed nearby.

In summary, all knowledge we generate by analysing the data world contributes to our model world. However, this is aimed at serving society. When translating this into practice we experience yet another world, the man-made *world of values*. This is people's perception and valuation of the world, which we know from experience is continuously changing. The results we derive in numerical analysis carry the potential to deliver input into value systems, but we should keep in mind what Diamond (1999) mentioned when talking about accepting innovations: 'Society accepts the solution if it is compatible with the society's values and other technologies.' Proving the existence of global warming, as an example, can be a matter of modelling. Convincing people of the practical relevance of the problem is a question of evaluation and communication, for which different skills may be required.

2
Patterns in vegetation ecology

2.1 Pattern recognition

Why search for patterns in vegetation ecology? Because the spatial and temporal distribution of species is non-random. The species are governed by rules causing detectable, regular patterns that can be described by mathematical functions, such as a straight line (e.g. a regression line), a hyperbola-shaped point cloud, or, in the case of a temporal pattern, an oscillation. But it might also be a complex shape that is familiar to us: Figure 2.1 shows the portrait of former US President Abraham Lincoln. Although drastically simplified, we immediately recognize his face. Typically, this picture contains more information than just the face: there is also the regular grid, best seen in the image on the right. This geometrically overlayed pattern tends to dominate our perception. The entire central image including the grid is

Data Analysis in Vegetation Ecology Otto Wildi

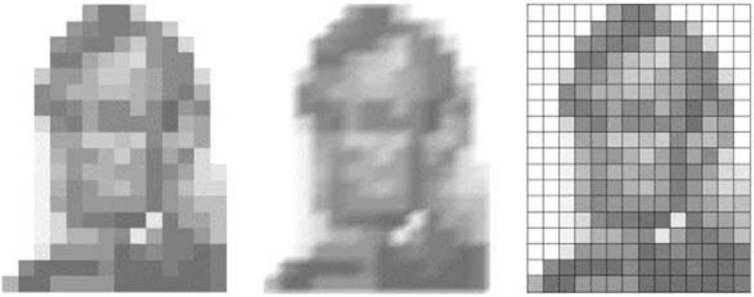


Figure 2.1 Portrait of Abraham Lincoln. Pixel image (left), blurred (centre), with superimposed raster (right).

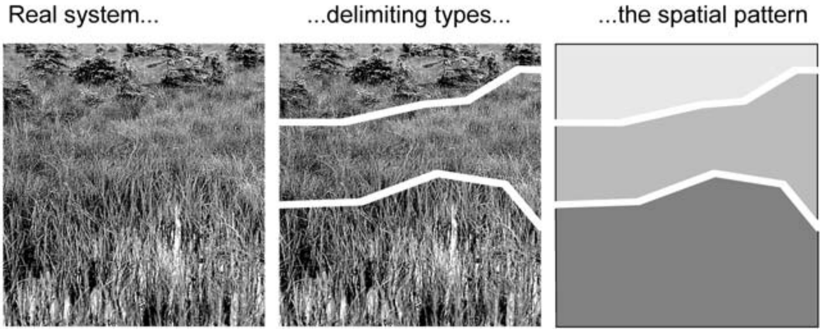


Figure 2.2 Vegetation mapping as a method for establishing a pattern (bog vegetation with a wetness gradient from the foreground to the background).

blurred, helping the human brain to recognize the face more easily. So patterns are frequently overlayed, and this also happens in ecosystems, where it is actually the rule. One of the aims of pattern recognition is in fact to separate superimposed patterns by partitioning the data in an appropriate way. A well-known application of pattern recognition is (vegetation) mapping. The usually inhomogeneous and complex vegetation cover of an area is reduced to a limited number of types. The picture in Figure 2.2 shows the centre of a peat bog in the Bavarian Pre-Alps. Three vegetation types of decreasing wetness are distinguished from the foreground to the background. Before drawing such a map the types have to be defined, a difficult task discussed in more detail in Chapter 6.

Patterns are often obscured not just by overlay, but by random variation (sometimes referred to as statistical noise) hiding the regularities. Methods are needed to divide the total variation into two components, one containing the regularity and one representing randomness.

One (statistical) property of any series of measurements is variance (s^2):

$$s^2 = \sum_{i=1}^{n} (x - \overline{x})^2$$

This is the sum of the squared deviation of all elements from the mean of vector $\vec{x}$. The mean can be interpreted as the deterministic component and the deviations as the random component of a measurement. Even in the simplest natural system the existence of a deterministic pattern and a random component can be expected. A typical example in vegetation ecology is the representation of a vegetation gradient as an ordination. A continuous change in underlying conditions, time or environmental factors leads to a nonlinear change in vegetation composition. When a vegetation gradient of this type is analysed, it will not manifest as a straight line but as a curve instead, also known as a horseshoe (see Section 5.5). What deviates from this can be considered statistical noise, but it can also come from yet another pattern. The issue is sketched in Figure 2.3 with data from a gradient in the Swiss National Park depicting the change from nutrient-rich pasture towards reforestation by *Pinus montana*. In this ordination the main pattern is the

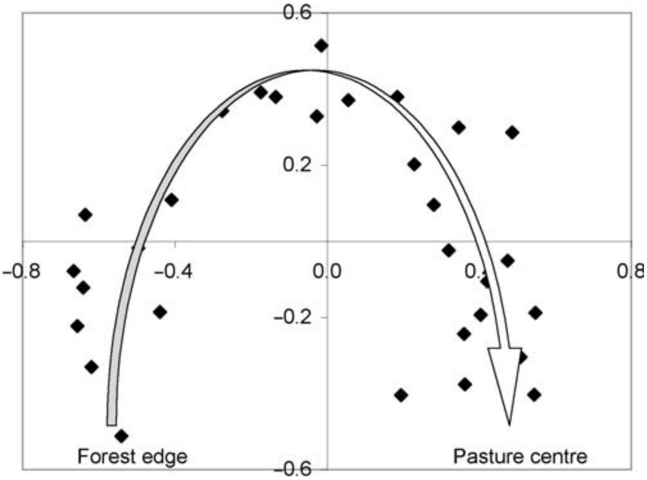


Figure 2.3 Ordination of a typical horseshoe-shaped vegetation gradient in the Swiss National Park. Relevés on the left-hand side are taken from the forest edge, those at the right-hand side from the centre of a pasture. If the arrow is assumed to represent the true trend then the distance of any one point from the arrow is caused by noise.

curved line and the random component comes from the deviations of the data points from this line. Alternatively, one may detect another pattern in the point cloud. As will be shown in Chapter 6, applying cluster analysis will result in determination of groups. This might be the preferred pattern for some practical applications like vegetation mapping.

I have shown so far that patterns refer to different kinds of regularities, some in space, some in time, others related to the similarity of objects, one-dimensional or multidimensional, deterministic or random. This book presents a strategy towards recognition of patterns. In Section 2.3 I refer to the sampling problem, a big issue as sampling yields the data and only these are subjected to analysis. Mathematical analysis starts with Chapter 3 on transformation, a step in any analysis that allows adjustment of the data to the objective of the investigation, while also partly overcoming restrictions imposed by the measurements. First, transformations address individual measurements (scalars), such as species cover, abundance or biomass, for which I frequently use the neutral term *species performance*. Second, vectors are subjected to transformation. A relevé vector includes all measurements belonging to this, including species performance scores and site factors. A species vector considers performance scores in all relevés where it occurs. In a synoptic table (Section 6.6) a relevé vector is a column and a species vector a row.

In Chapter 4 multivariate comparison is presented. Comparing two relevés, one has to include all species and all site factors. This can be done in many different ways. The same applies to the species vectors, depicting their occurrence across all the relevés, and the site vectors, doing the same. The resemblance pattern is then defined by comparing all pairs of species and relevé vectors. If the number of vectors involved is equal to n then the dimension of resemblance matrix including all pairwise similarities is $m = (n * (n - 1)/2)$. Because of the tremendous size of this matrix, further analysis is required.

Many of the subsequent analyses directly access similarity matrices, such as ordination (Chapter 5), showing similarity in reduced dimensional space, classification (Chapter 6), showing groups instead of single relevés, and ranking (e.g. Section 5.6), erasing relevés or species considered unimportant in the given context. These three approaches unveil patterns. Chapter 7 is devoted to the comparison of patterns, being biological, environmental, spatial or temporal. The analysis of temporal patterns is shown in Chapter 9 and is related to static (Chapter 8) and dynamic (Chapter 10) modelling, of which the very basic elements as well as examples are shown. Finally, two applications illustrate practical issues through specific data sets: the analysis of wetland vegetation in Switzerland in Chapter 11, as an example of handling large data sets, and the analysis of forest vegetation data in Chapter 12, focusing on the interpretation of ecological patterns.

2.2 Interpretation of patterns

Distinguishing pattern, process and mechanism (Anand 1997) is one way of proceeding towards interpretation of results. After identifying a pattern, one seeks a process that might have generated it. Identifying this process can be an easy task, as shown in Figure 2.4 (left). The opening in the forest was created on 26 December 1999, when the storm *Lothar* hit the Swiss Plateau. Figure 2.4 (right) depicts a different process: human impact, in this case hay production, prevents forest regrowth below the timber line. However, the case of the vegetation gradient in the Swiss National Park shown in Figure 2.3 is more complicated. At first glance one would expect a strong nutrient gradient to which vegetation has responded. But long-term investigations have shown that it is actually the outcome of species movements in the direction from the forest edge towards the centre of an ancient pasture (Wildi & Schütz 2000) (see Section 9.4.2 for further explanations). This illustrates why it is sometimes difficult to distinguish between spatial and temporal processes.

Behind processes there are often mechanisms – that is, natural laws – acting as drivers. One such law is gravity, which lets an apple fall from a tree. Dynamic wind forces have caused the trees to break in the

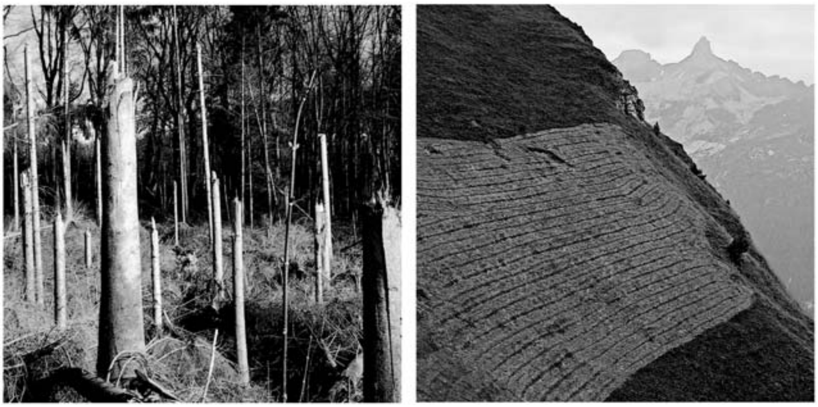


Figure 2.4 Left: a natural event – forest gap caused by storm Lothar, 26 Dec. 1999. Right: man-made – a meadow just below the timber line.

opening shown in Figure 2.4 (left). Why did the trees break, instead of being uprooted as usual? Why has the area of damage an almost circular shape, while the neighbouring trees were not damaged? A nonlinear physical process – the turbulent flow of air – seems to be the force that caused the pattern. This illustrates that sometimes a physical process must be understood in order to interpret the outcome. A mechanism can also be biological: in the case of the pastures of the Swiss National Park, one cause is probably the browsing behaviour of animals. Before 1914 the pasture was grazed by cattle, which preferred the centre of the forest clearing. After 1940 red deer were invading the park and we know form investigations that they browse the pasture more evenly. In this case the behaviour of the animals is a mechanism governing the process of vegetation change.

Space and time almost always interact, resulting in space–time patterns. The roles space and time play can differ considerably, as shown in the two examples below. The first is presented in Figure 2.5, where net primary production was measured at three different time intervals in 2001 by the US MODIS sensor. The pictures illustrate the seasonal changes leading to complex and fast shifting spatial patterns of primary production all across Europe. Shifting spatial patterns occur everywhere and are not only caused by seasonality, but by weather fluctuations in general.

In the second example a persisting spatial pattern reveals a process dating far back. In the year 2001, Mátyás and Sperisen published a map of Switzerland showing the distribution of oak trees. Based on chloroplast DNA they distinguished seven haplotypes, among which two dominated: no. 1 (light) and no. 7 (dark) in Figure 2.6. Historic studies suggested that this was not

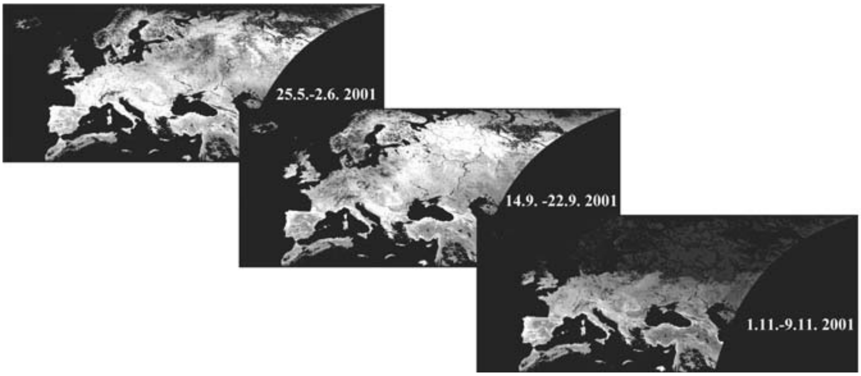


Figure 2.5 Primary production of the vegetation of Europe measured by the MODIS sensor at three time intervals in 2001. Light areas have high weekly primary production. http://modis.gsfc.nasa.gov/

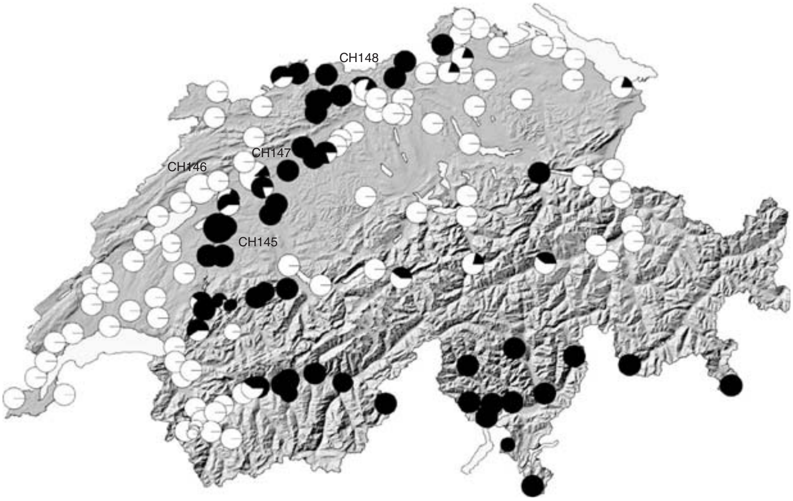


Figure 2.6 Distribution pattern of oak haplotypes in Switzerland according to Mátyás & Sperisen (2001). This reveals the post-glacial invasion route.

the result of forest management, but the effect of re-colonization of Central Europe by oak *(Quercus sp.)* some 8000 years ago. The known retreat areas for oak during glaciation were Spain, southern Italy, the Balkan peninsula and probably Greece. Surprisingly, the genetic pattern found concerns three species simultaneously: *Quercus robur, Q. petraea* and *Q. pubescens*. In other words, all retreats hosted more than one of today's oak species.

Haplotype no. 7 (dark circles) arrived from the Balkan peninsula, invading southern Switzerland first, then crossing the western Alps and further progressing north towards France and Germany. The remaining haplotypes (white circles) originate from southern France (left-hand side in Figure 2.6). Genetic patterns of this kind recently helped reveal the spread of many species, including the modern *Homo sapiens sapiens*.

2.3 Sampling for pattern recognition

2.3.1 Getting a sample

The aim of data sampling is to generate a numerical description of the real system we wish to analyse. That is what a 'good' sampling design does. A 'bad' design includes the risk of generating a pattern which is absent in the real world. Generating a sampling design means that the sampling elements have to be chosen, which is explained below. In this section, sampling is not presented in detail. The elements are introduced because these determine the organization of the data sets. It must be noted that the selection and definition of these elements is a central issue in any investigation as it will determine the contents and relevance of the results. There exist few guidelines to help find a good sampling design and much is left to the intuition of the researcher.

The terminology used throughout this text is shown in Table 2.1 and applied to an object in Figure 2.7. Many of the terms are prone to confusion,

Table 2.1 Terms used in sampling design (International Statistical Institute 2009) **E**nglish, **F**rench, **G**erman, **S**panish.

E / F / G / S	Meaning	Example
Population (universe) Population (population) Population (Grundgesamtheit) Población (universo)	All measurable items	All plants in an investigation area
Sample Échantillon Stichprobe Muestra	All measurements taken within the investigation area	A vegetation table
Sampling unit Unité d'échantillon Stichprobeneinheit Unidad de muestra	One element of a sample	A relevé
Attribute Attribut Merkmal, Attribut Atributo	Descriptors of the sampling units	Plant species, site factors
Sampling plan Plan d'échantillonage Stichprobenplan Planeo de la muestra	Location of units, size and shape	Sampling grid
Stratum Strate, couche Schicht, Stratum Estrato	Subset of the sample	Relevés between 600 m and 800 m a.s.l.

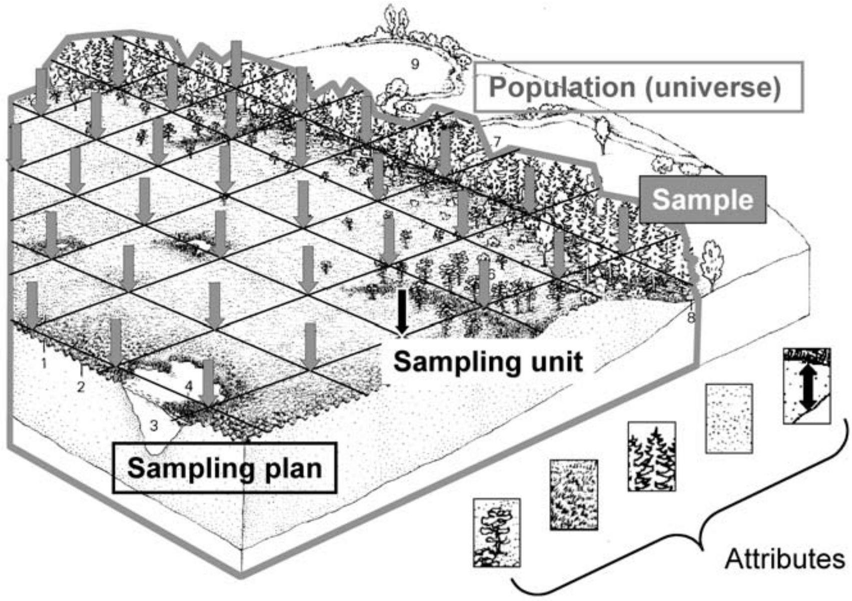


Figure 2.7 The elements of sampling design. In this example, a systematic sampling design is used to assess the state of a peat bog.

for example a sample in some textbooks is the same as a sampling unit in others. The translations given in Table 2.1 intend to foster communication in some languages (International Statistical Institute 2009). The first step in sampling is the delineation of the *population*, which is the object to be investigated; that is, the full investigation area (not to be confounded with the population in the biological sense). The results will be valid for this population in terms of time, space and content. In theory, all items belonging to the population could be measured, such as the diameter and height of all trees in a forest, for example. In practice, however, the costs of such a strategy (termed *full enumeration*) would be excessive and much of the energy and money would be wasted. Instead, a subset of all measurable items is taken: the *sample*. In the terminology used here, the sample is the full set of measurements taken from the population. It provides an estimate of real values of parameters of interest. The sample consists of *sampling units*. In vegetation science, a sampling unit is often a plot of pre-defined size and shape (Kent & Coker 1992), as indicated in Figure 2.7. Each sampling unit is characterized by *attributes*, such as percentage vegetation cover. One can measure just one attribute per sampling unit. In practice, the number of attributes is often rather high. This is the case when relevés are taken where all the species occurring in the plot are recorded.

There are many more decisions needed to accomplish a full sampling design, one of them being the sampling plan. Plots can be arranged systematically, as seen in Figure 2.7; for other applications, a random arrangement is the best option; while in more complex situations, a *stratification* of the entire surface is suggested. When stratifying its surface, the investigation area is divided into subspaces, the *strata*, which are formed based on available information on the investigation area, such as thematic maps. To increase the efficiency of sampling, different sampling plans can be applied to the individual strata. If small strata are more densely sampled than large strata, the sampling intensity eventually becomes equal for all strata. Not mentioned in Table 2.1 are plot size, plot shape and the time of sampling.

2.3.2 Organizing the data

At first glance, organizing the data appears to be a technical matter only: the sample is usually presented in a rectangular matrix, where the columns are reserved for the sampling units and the rows are the attributes (or vice versa). However, in natural space–time systems, the variables can be grouped by type. For this, the concept of space is used. A data table of the kind presented here forms the *data space*. As will be shown later (Chapter 4), there are other, even more abstract spaces such as the resemblance space.

At this point, some subtypes of data space are considered:

The biological space. This consists of the attribute vectors describing the biotic part of the system, such as plant species, plant cover, animal species, population sizes, life forms, etc. In many models of data analysis these function as dependent variables.

The environmental space. The attributes involved measure the environmental conditions, such as climate, nutrients, the substrate or disturbances such as fire or land-use. They are often considered explanatory or independent variables.

The physical space in two or three dimensions. In the sample space, each sampling unit is described by its x-, y- and z-coordinate. By assigning this, the sampling plan also becomes part of the sample. Specific methods exist for the analysis of spatial effects.

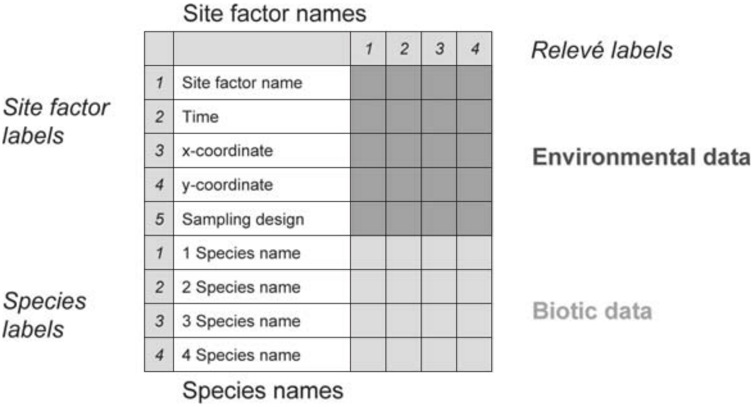


Figure 2.8 Organization scheme of sample data. The environmental data include spatial and temporal attributes as well.

Time space. This has just one dimension, the time axis. As in physical space, there are special methods to analyse time series data.

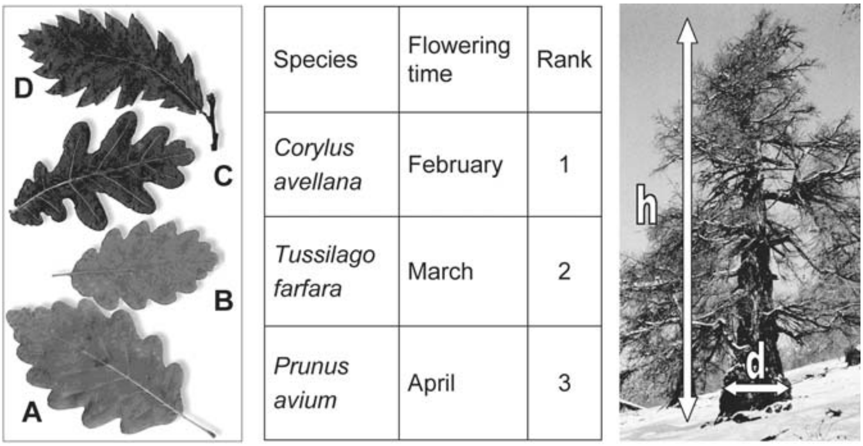
In traditional phytosociology (Braun-Blanquet 1964, Dengler *et al*. 2008) there is a convention to put ecological, spatial and temporal attributes on top of data tables. The biological ones are then added in the form of species lists. An example is shown in Figure 2.8. For improved presentation and interpretation of results, the sampling units and all the attributes are numbered and, if available, identified by names. This organization allows easy access to the data by most software packages.

3 Transformation

3.1 Data types

As mentioned in Chapter 1, the aim of measurement is to generate a numerical description of the real world. This sounds like a merely technical issue; on closer inspection, however, data often mirror the tool that has been used for the measurement. We measure what we can measure and we omit what we cannot. Sometimes we also have a choice in the method we use to obtain some particular information, as for example in measuring the colour of light. We can either use a scale with discrete states (red, blue, yellow, etc.) or measure the wavelength of electromagnetic radiation. In the first case the measurement addresses a type of colour, in the second we get a number, representing a totally different data type. We need to distinguish different data types as their numerical analyses require different treatments. In some cases the transformation of one type into another may be necessary

Data Analysis in Vegetation Ecology Otto Wildi



Species	Flowering time	Rank
Corylus avellana	February	1
Tussilago farfara	March	2
Prunus avium	April	3

Figure 3.1 An example of three data types. Left: nominal data (leaves of (a) *Quercus petraea*, (b) *Q. pubescens*, (c) *Q. robur* and (d) *Q. cerris*, photograph WSL, Genetic ecology). Centre: rank data (flowering order of three plant species). Right: metric data (height and stem diameter of a tree).

(e.g. in Table 3.3). Some textbooks distinguish between quite a few; however, a very simple classification would be the one in Figure 3.1:

Nominal data are recorded according to a list of possible states. Four leaf types are distinguished in Figure 3.1 and labelled by letters. These are (a) *Quercus petraea*, (b) *Q. pubescens*, (c) *Q. robur* and (d) *Q. cerris*. Data of this type are restricted in the application of mathematical operations. Leaf types are either the same or different, thus the operations to be applied are $=$ and $\neq$.

Ordinal data are measurements on a rank scale. The three plant species noted in the centre of Figure 3.1 flower at different times of the year. In a warm winter, flowering of *Corylus avellana* may start in December of the preceding year. However, if cold weather conditions prevail the first flowers may show in late February. Yet the order remains always the same: *Corylus* will flower before *Tussilago* and *Prunus* will be the latest. Hence, there is a natural order irrespective of weather conditions. The operations applicable to nominal data also apply for ordinal data. In addition, calculating a difference in ranks makes sense. A large difference in ranks usually means lower similarity of the two elements.

Metric data are measurements of distance, volume, weight, force and so on. The example in Figure 3.1 shows the height and diameter of a tree. In metric data all arithmetic operations make sense, including the ones allowed for the previously mentioned types. For example, the height and stem diameter of a tree allows calculation of the approximate volume of the trunk.

One simple rule for the transformation of data types concerns the direction in which this is done. It is easy to transform from metric to ordinal and further to nominal (with loss of information, however), but the opposite direction requires additional assumptions about the meaning of the measurements. This is a common practice when analysing plant cover-abundance data, as will be shown in Section 3.4. The transformations presented in the following two sections apply to ordinal and metric data only. In classical statistics (Sampford 1962) there are formal rules that have to be applied when using transformation, such as correcting for non-normal distributions of the data. In fact, transformation generally is used to adapt data to statistical models. Yet I present a slightly different view here: attributes are measured at a specific scale (given by the measuring device used). This scale does not necessarily serve the objective of the investigation. Often, the *perspective* has to be adjusted: one metre when seen from two metres away may appear large, but when seen from one kilometre's distance will hardly be visible. Hence, when talking about transformation, we will have to keep the purpose of our measurement in mind.

3.2 Scalar transformation and the species enigma

When transformations are applied to individual measurements, I call them scalar transformations. Scalar transformation means that the scale used for measuring is adjusted according to our intention. Such transformations are widespread in environmental science. Often a relationship between two variables only emerges after proper transformation. Figure 3.2 illustrates this in a biological example. It is generally assumed that the survival of plant and animal populations depends on appropriate environmental conditions. When the conditions are favourable, populations may grow. Under less favourable conditions, they are likely to remain small. A small population may, for example, consist of five individuals. But 'large' is not, say, 20, but 100 or even more. When correlating population size with an environmental

Biological population			
Population size n	5	25	100
nominal	small	medium	large
rank	1	2	3
n' = $n^{0.25}$	1.49	2.23	3.16

Figure 3.2 Scalar transformation of population size to optimize for correlation with environmental factors.

variable, for example temperature, a transformed number of individuals may be a better measure of population size. When taking $n' = n^{0.25}$ for example, we adopt a more qualitative view of the size: 5 will become 1.49 (small), 20 will be 2.23 (average) and 100 is 3.16 (large). Correlating these values with temperature could easily yield a good linear relationship.

Another way of reasoning is that scalar transformation changes the perspective of objects: in many ways they appear smaller when seen from a distance, as illustrated in Figure 3.3: the trees are just a series of points in two-dimensional space, connected by a line. On the left, the coordinates are untransformed and all trees have the same height. In the middle and on the right the coordinates have been transformed and this obviously affects the perspective by reducing the importance of high values compared to low values.

Transformation may sometimes contribute to the solution of problems inherent in ecosystems, such as poor correlation of species occurrence under

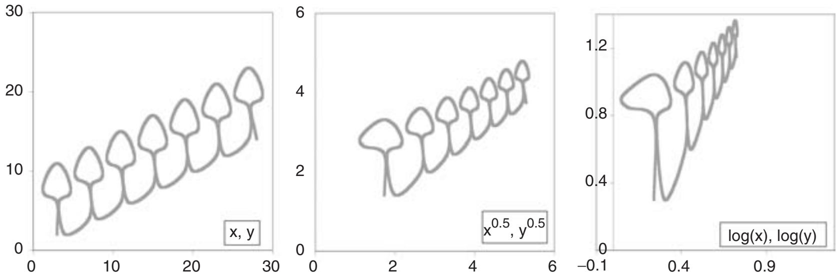


Figure 3.3 Scalar transformation of the coordinates of a graph. These transformations affect the perspective adopted during the course of the analysis.

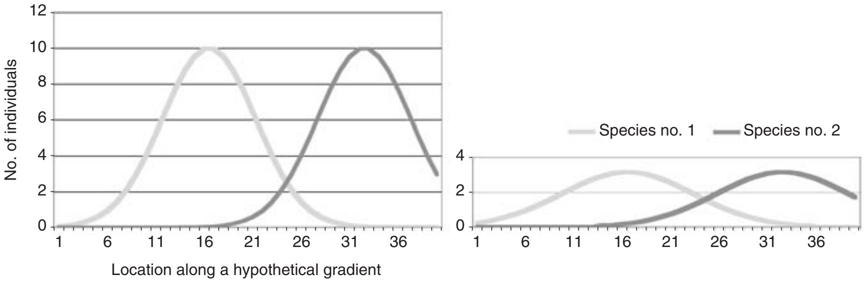


Figure 3.4 Overlap of two species with Gaussian response along a hypothetical gradient. Left graph: species scores on a 0—10 performance scale. Right graph: the same scores, but square-root transformed.

similar site conditions (Chapter 1). Despite the hope of many practitioners that species will form groups, thereby enabling the identification of vegetation types, reality differs. When inspecting synoptic tables (Section 6.6) many species overlap nicely, but they hardly ever cover the same niche. Even worse, apparently species tend to avoid common distribution (Clarke 1993). As claimed by Gleason (1926, 1939) in his 'individualistic concept of the plant association', species behave like loners. And in fact if the formation of an ecological niche is the result of Darwinian struggle for life then species are prone to ecological differentiation. I attempted to sketch a typical case of two overlapping species in Figure 3.4. The response of both species to the hypothetical gradient is Gaussian (Section 8.2.2). Despite the shifted optima there is a small area of overlap. On the right, the same situation is shown, but this time with performance scores square-root transformed to let the high scores shrink. Transformation in this case affects the *relative* overlap as this is now larger than in the left graph. In practice this may be most welcome as co-occurrence measures of species are often unpleasantly low. As will be shown later (Section 7.2.3), transformations towards presence-absence are frequently a good choice when revealing ecological patterns.

3.3 Vector transformation

As shown in Section 2.3.2 data are traditionally organized in two-dimensional data matrices. The column vectors are the sampling units and the row vectors are the attributes. Transformation of vectors therefore concerns rows, columns or both simultaneously. The aim in either case resides in obtaining similar properties of vectors. When sampling unit transformation, it is frequently the intention to achieve equal weight of all

Table 3.1 Effects of different vector transformations on the properties of data.

Term	Formula	Explanation
Centring	$x_i' = x_i - \overline{x}$	Adjusts mean to zero
Normalizing	$x_i' = \frac{x_i}{\sqrt{\sum_{x^2}}}$	Adjusts vector length to 1.0
Standardizing	$x_i' = \frac{x_i - \overline{x}}{\sqrt{\frac{1}{n}\sum_{(x_i - \overline{x})^2}}}$	Adjusts mean to zero and variance to 1.0
Range adjustment	$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$	This is a fuzzy transformation (range 0.0–1.0)

Table 3.2 Numerical example of vector transformation (two vectors).

	x_1	x_2	x_3	x_4	x_5	$\sum$	$\overline{x}$	$\sum_{x^2}$	$\sqrt{\sum_{x^2}}$	S_x
Raw	2.00	0.00	5.00	4.00	6.00	17.00	3.40	81.00	9.00	2.15
	0.00	0.00	1.00	2.00	2.00	5.00	1.00	9.00	3.00	0.89
Centred	−1.40	−3.40	1.60	0.60	2.60	0.00	0.00	23.20	4.82	2.15
	−1.00	−1.00	0.00	1.00	1.00	0.00	0.00	4.00	2.00	0.89
Normalized	0.22	0.00	0.56	0.44	0.67	1.89	0.38	1.00	1.00	0.24
	0.00	0.00	0.33	0.67	0.67	1.67	0.33	1.00	1.00	0.30
Standardized	–0.65	−1.58	0.74	0.28	1.21	0.00	0.00	5.00	2.24	1.00
	−1.12	−1.12	0.00	1.12	1.12	0.00	0.00	5.00	2.24	1.00
Fuzzyfied	0.33	0.00	0.83	0.67	1.00	2.83	0.57	2.25	1.50	0.36
	0.00	0.00	0.50	1.00	1.00	2.50	0.50	2.25	1.50	0.45

samples. Attribute transformation results in obtaining the same potential weight in describing the sampling units. Some of the most frequently applied vector transformations are shown in Table 3.1, with a numerical example given in Table 3.2.

A first step, rarely used alone, is *centring*. The mean of the vector is deduced from each element. As a result, the new mean and the new sum both become zero. The sum of squares also changes, without becoming zero. The variance, however, remains unchanged.

Normalizing is a different method of transformation. Each element of the vector is divided by its (Euclidean) length. The vector sum, the vector mean change and the vector length are now 1.0. As shown in Table 3.2, the vectors become more similar in many ways while the variances still differ.

A most rigorous transformation is *standardizing*. This is a combination of centring and normalizing. As a result, the vector mean is zero and the standard deviation (and the variance) becomes 1.0. The length of the vector is equal to the square root of the number of elements. Standardization is used to compare different scaled measurements, such as temperature and the height of trees, for example. However, standardization has a downside: if the information is hidden in the variance then it will be lost.

Fuzzyfying is a simple transformation (Boyce & Ellison 2001). The elements are adjusted to range from zero (lowest score) to 1.0 (highest score). It should be used only if you intend to adopt this view of the data. Aberrant values can set the boundaries in an undesirable way, deteriorating the observations completely. Fuzzy transformation is not an alternative to normalizing or standardizing, but rather is applied in combination with these.

3.4 Example: Transformation of plant cover data

In phytosociology, Braun-Blanquet (1932) established a scale for measuring the quantity of plant species – that is, species performance – in vegetation relevés. He released his first comprehensive book on that topic in 1928 (English version in 1932). From the point of view of modern data analysis this scale (the so-called cover-abundance scale) is a mixture of form and content. At lower species densities, it expresses the abundance of individuals. At high densities, it directly translates to plant cover percentage. As shown in Table 3.3, it starts with a nominal notation in the form of the symbol 'empty' (in Table 3.3 a minus sign), followed by '+'. Then it continues with a rank scale from 1 to 5. In the past hundred years, huge data sets have been collected all over the globe using this scale (Dengler *et al*. 2008). Handling such data is therefore an issue in data analysis. Table 3.3 demonstrates how it could be done based on an idea published by Maarel (1979).

In the first step the code is transformed into a proper rank scale with a range from 0 to 6 (column three in Table 3.3). The ranks are then treated as if they were metric. The justification for this is shown in the right-hand columns, where the rank scale is further transformed according to:

$$x' = x^y \tag{3.1}$$

Table 3.3 Transformation of cover-abundance values in phytosociology.

Code	Cover %	$x^{1.0}$ (rank)	$x^{0.1}$	$x^{0.25}$	$x^{2.5}$ (cover)
−	0	0	0	0	0
+	<1	1	1	1	1
1	5	2	1.07	1.19	5.65
2	17.5	3	1.12	1.31	15.58
3	37.5	4	1.15	1.41	32.00
4	62.5	5	1.17	1.50	55.90
5	87.5	6	1.19	1.57	88.18

where x' is the transformed score. When $y < 1$ the data approach a binary state $\{0, 1\}$. Near $y = 2.5$ it can be seen that this approximates the initial cover percentages. By choosing the appropriate value for y the scope of the analysis can hence be altered to emphasize either the qualitative or the quantitative aspect. For many applications, choosing $y = 0.25$ turns out to be a good compromise as this expresses the qualitative view while considering the quantitative sufficiently as well (see Section 7.2.3).

4 Multivariate comparison

4.1 Resemblance in multivariate space

When talking about resemblance we address two types of measurement: *similarity*, where high values signify a high proportion of common features, and *distance*, where high values signify dissimilarity. As long as sampling units are described by one species or one site factor only, comparison is straightforward and the operational rules discussed in Section 3.1 on data types are valid. When more attributes exist, the technique is no longer trivial and several questions need to be answered in advance of data analysis:

- Are the attributes of the same type or is treatment necessary?
- Do the attributes have the same weight or is transformation necessary?

Data Analysis in Vegetation Ecology Otto Wildi

- Are the attributes measured on the same scale or have the scales to be adjusted by transformation?

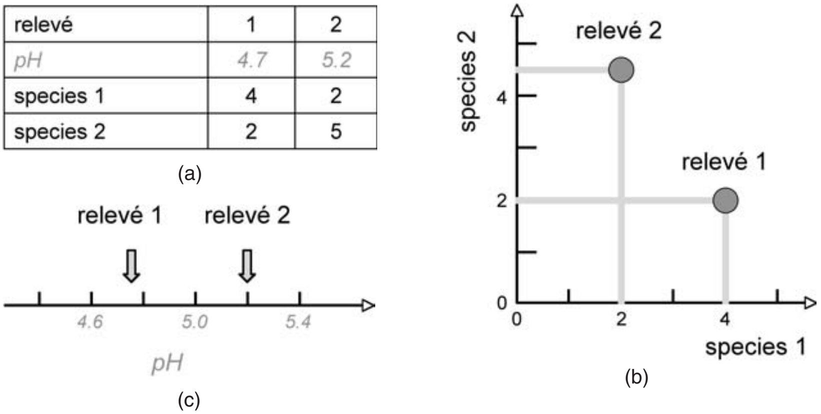
- Are some of the attributes correlated and therefore partly carrying the same information?

Due to the multivariate nature of data, several attributes or sampling units have to be taken into account simultaneously. A first and really illustrative approach to resemblance is the geometric, where attributes function as axes in a scatter diagram. Attribute scores are therefore the coordinates of the sampling units, which are points located in space (Figure 4.1).

A second way of measuring resemblance is a statistical one, specifically suited for species lists where presence and absence are major issues. The joint occurrences can be counted and statistical measures will help to decide whether the frequency assessed is higher or lower than expected compared to a random situation.

Probably the most common technique is the use of product moments, among which the better known are correlation and covariance. If much of the variance of two sampling units is shared then covariance is high and they are considered similar.

Of course, there are more approaches to the comparison of sampling units or species, such as measures relying on information theory (Rényi 1961, Orlóci 1978). I will not discuss these in the following sections.



relevé	1	2
pH	*4.7*	*5.2*
species 1	4	2
species 2	2	5

(a)

Figure 4.1 Presentation of data in the Euclidean space. The data are shown in (a). In (b), the biological attributes are used to represent the relevés in two-dimensional (biological) space. (c) shows the one-dimensional environmental space.

4.2 Geometric approach

Multivariate similarity can easily be related to geometry, because geometry considers dimensionality. Geometrical space may be one-dimensional (a straight line), two-dimensional (a surface) or three-dimensional (a volume). The dimensions can also be extended to any number, say four or a hundred.

In practice, there are at least two constraints. First, it is assumed that the dimensions (i.e. the axes) are based on the same scale. Second, the weight of the axes is the same. The latter is *only* the case if the attributes (and hence the axes) are uncorrelated. If the attributes are perfectly correlated then they carry identical information. The use of the multivariate Euclidean space is only justified if the attributes are equally scaled, as is the case when using the Braun-Blanquet code, for example Table 3.3. The principle is shown in Figure 4.1, where the data are given in (a). It is assumed that the pH values are part of the environmental space, whereas the species scores form the biological space. In (b) the two-dimensional biological space is shown, where the relevés are points in the scatter diagram. Whenever comprehensive species lists are used, the biological space is extremely high-dimensional, with each species forming its own dimension. In (c), however, it can be seen that a space may also be one-dimensional only. The relevés are still points, but on a one-dimensional vector, in this case pH.

Resemblance of any two sampling units in Euclidean space is most easily measured as a distance. If the distance is short then any two relevés are similar. If the distance is long, the relevés involved diverge in many possible ways. There are different methods of calculating distance, as shown in Figure 4.2. A straightforward measure is *Euclidean distance*. The Euclidean distance between relevé 1 and relevé 2 is calculated by:

$$De_{1,2} = \sqrt{\sum_{j=1}^{p} (x1_j - x2_j)^2} \tag{4.1}$$

In the left-hand side of Figure 4.2, this is the direct distance between the corresponding data points. Equation (4.1) is written for p species and is therefore valid for any number of dimensions. The lower bound of $De_{1,2}$ is zero for identical relevés; the upper bound has no limit. When the number of dimensions (species) increases, the Euclidean distance tends to become larger.

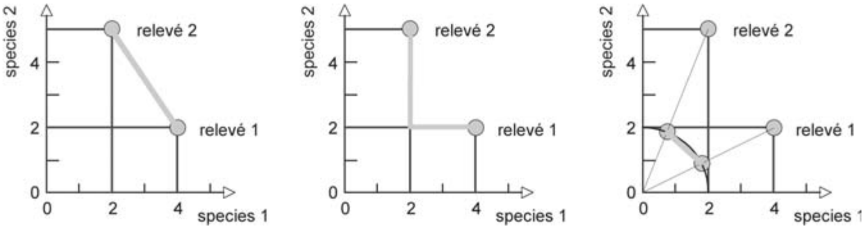


Figure 4.2 Three ways of measuring distance. Left: Euclidean distance. Centre: Manhattan distance. Right: Chord distance.

A second possible measure is *Manhattan distance*, which is the sum of the differences of the scores calculated on all axes, that is:

$$Dm_{1,2} = \sum_{j=1}^{p} |x1_j - x2_j| \tag{4.2}$$

The Manhattan distance (Equation 4.2) has similar properties to the Euclidean. As shown in Figure 4.2, centre, the Manhattan distance is generally somewhat longer than the Euclidean distance.

In some cases, methods differ by the intrinsic transformation applied. *Chord distance* is an example. It is identical to the Euclidean distance, but after normalizing the vectors. Combining these two operations yields the corresponding formula:

$$Dc_{1,2} = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} x1_j x2_j}{\sqrt{\sum_{j=1}^{p} x1_j^2 \sum_{j=1}^{p} x2_j^2}}\right)} \tag{4.3}$$

Chord distance has a lower bound of zero (for identical relevés or species vectors). Unlike the previous measures, there is now a maximum value of square root of two; that is, 1.414213. This is the case when relevés have no species in common. It is difficult to decide whether the normalizing involved is ideal for applications: when transformation is really needed, many researchers prefer standardization (adjusting vector length and variance) to normalization (adjusting vector length only). This idea will be discussed in the context of the product moment measures (Section 4.4).

4.3 Contingency testing

Contingency testing is a statistical approach, focusing on the joint occurrence of objects. In the case of relevés, these are common species. If there are many, one assumes that the relevés are similar. From a statistical point of view the question arises whether the number of common species is above, equal to or below expectation. Hence, we will have to deal with the meaning of 'expectation'.

The standard setup for this type of measurement is the contingency table, as shown in Table 4.1. This explains how relevés are compared. For each species, common occurrence is counted in cell a. When a species occurs in one relevé only, it is counted in either cell b or cell c. If a species occurs in neither relevé 1 nor 2, it contributes to cell d.

The row and column sums yield useful numbers as well. The sum in row $a + b$ is the total number of species found in relevé 2; the sum in column $a + c$ for relevé 1. The sum in row $c + d$ is the number of species that do not occur in relevé 1 and the sum of column $b + d$ the number that do not occur in relevé 2. The grand total Σ is the total number of species considered for calculations, including those occurring in neither relevé 1 nor 2.

Using such counts from contingency tables, an almost unlimited number of coefficients can be calculated. Many of these are listed in Legendre & Legendre (1998), pp. 275–276. They differ in their properties and some are related to other types of resemblance measure. Four of them are shown in Table 4.2.

The *Jaccard coefficient* S_J is the oldest, published in 1901. It counts the number of common species and the total number of species present in either of the two relevés. The range is from zero (no species in common) to one (all species in common). When 50% of the species are common, $S_J = 0.50$.

Table 4.1 Notations in contingency tables. *a*, *b*, *c* and *d* are frequency counts.

relevé 1

relevé 2	+	−	
+	a	b	$a + b$
−	c	d	$c + d$
	$a + c$	$b + d$	Σ

Table 4.2 Resemblance measures using the notations in Table 4.1.

Name	Formula	Distance measure	property
Jaccard	$S_J = \frac{a}{a+b+c}$	$D_J = 1 - S_J$	metric
Soerensen	$S_S = \frac{2a}{2a+b+c}$	$D_S = 1 - S_S$	semimetric
Simple maching	$S_{SM} = \frac{a+d}{a+b+c+d}$	$D_{SM} = 1 - S_{SM}$	metric
Chi squared	$\chi^2 = \left(\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}\right)^2$	$D_{\chi^2} = 1 - \chi^2$	metric

The second is the *Soerensen coefficient* S_S. This differs from the Jaccard coefficient in that common species have double weight. The range is also zero to one, but when 50% of the species are in common, $S_S = 0.667$. The derived distance measure (the complement) is called semimetric, because it may happen that the distance configuration of three or more relevés cannot be presented in Euclidean space (i.e. the triangular unequality is violated), limiting its application in some methods.

In the *Simple maching coefficient* S_{SM}, frequency d is used as well. When analysing a sample, such as a synoptic table (Section 6.6), the total number of species considered remains the same for all pairs of relevés. However, when using different lists of species, S_{SM} differs for the same pair of relevés.

The fourth coefficient is the *Chi squared* (χ^2), as known from statistics. This is the sum of squared differences from the expected frequencies when independence is assumed. The probability distribution of the χ^2 can be found in most statistical textbooks. This allows it to be used for significance tests – as long as data are based on statistical sampling. When analysing vegetation data the χ^2 is rarely used in the statistical sense, but rather as yet another similarity measure with a lower bound of zero and no finite upper bound.

4.4 Product moments

Product moments are a flexible group of measures. They express the degree to which vectors point in the same direction. This conforms with the basic concept of variance (the variance within one vector) and covariance (the variance shared by two vectors). Four related measures that differ in their implicit transformation only are listed in Table 4.3.

Table 4.3 Product moments. Types differ in the mode of implicit data transformation.

Name	Formula	Transformation
Scalar product	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = X_{hj}$
Centred scalar product	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = X_{hj} - \overline{X}_h$
Covariance	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = (X_{hj} - \overline{X}_h)/\sqrt{n-1}$
Correlation	$S_{jk} = \sum_{h=1}^{p} A_{hj} A_{hk}$	$A_{hj} = \frac{(X_{hj} - \overline{X}_h)}{\left(\sum_{e=1}^{n} X_{he} - \overline{X}_h\right)^{1/2}}$

The *scalar product* is the vector product with no further transformation involved. If all scores are positive it ranges from zero to infinity. The more attributes involved, the larger the scalar product.

The *centred scalar product* involves centring of the observational vectors; thus the mean of any vector will be zero. On average, half of the coefficients will be negative with no upper and lower bound.

Covariance does the same thing as the centred scalar product, but in addition it offers a correction for the number of elements, n. It is used in analysis of variance. Note that $n - 1$ corrects for the underestimation of variance in small samples n.

The *product moment correlation coefficient* (termed *correlation* in Table 4.3) standardizes the observational vectors implicitly. Their mean is zero and the standard deviation is equal to one. This has the practical advantage that there are fixed upper and lower bounds: $-1 \leq r \leq +1$. This is shown in Figure 4.3 in the form of a geometrical interpretation. When two vectors show the same trend but in the opposite direction, correlation approaches $cos\alpha \approx -1$. When they are independent, it is around zero. When they point in the same direction it approaches $cos\alpha \approx +1$.

Many standard statistical packages use the correlation coefficient as a default measure for the majority of methods. Thus, measurements taken at different scales become comparable, and the variance is adjusted. If this is not desirable because one expects important information from variance differences then another option should be considered. An example where this frequently is suggested is the comparison of species-rich relevés versus species-poor relevés.

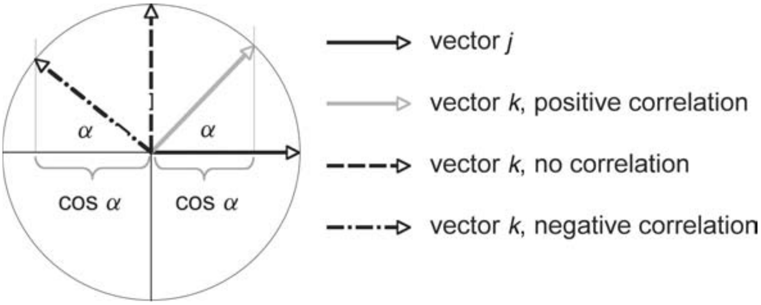

α
α
cos α
cos α
vector *j*
vector *k*, positive correlation
vector *k*, no correlation
vector *k*, negative correlation

Figure 4.3 The correlation of vector *j* with vector *k*. The correlation coefficient is the cosine of the angle α between any two observational vectors *j* and *k*.

4.5 The resemblance matrix

Whereas pairwise comparison of observational vectors like relevés or species is useful for many purposes, assessing the pattern of an entire sample involves the computation of a resemblance matrix. This is done by comparing all possible pairs of sampling units, resulting in an $n * n$ matrix of resemblance coefficients. Such a matrix (Figure 4.4) is generally symmetric and only the lower-left triangle (or the upper-right) has to be considered. Depending on the resemblance measure used, the diagonal elements, the self-similarity of the sampling units, may be of interest or not. When using Euclidean distance, for example, they are all zero; when using the correlation coefficient they all equal 1.0. When using covariance, however, they carry the variances of the sampling units and these usually vary.

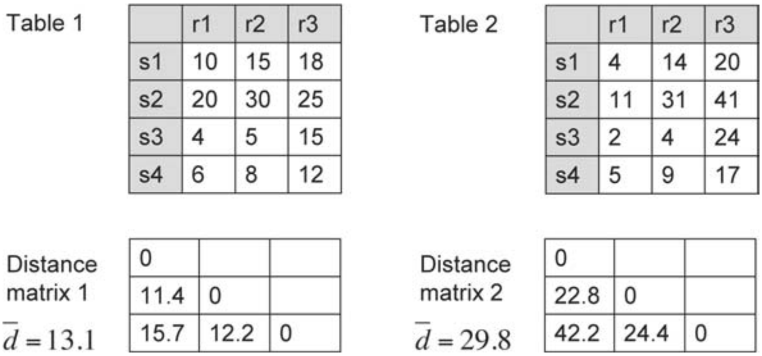


Table 1

	r1	r2	r3
s1	10	15	18
s2	20	30	25
s3	4	5	15
s4	6	8	12

Table 2

	r1	r2	r3
s1	4	14	20
s2	11	31	41
s3	2	4	24
s4	5	9	17

Distance matrix 1

$\bar{d} = 13.1$

0		
11.4	0	
15.7	12.2	0

Distance matrix 2

$\bar{d} = 29.8$

0		
22.8	0	
42.2	24.4	0

Figure 4.4 The average distance of a distance matrix is a perfect measure for homogeneity of a sample. Left: high homogeneity. Right: low homogeneity.

Resemblance matrices may become very large. When computing the triangular matrix only, without the diagonals, the number of elements is $(n * (n - 1))/2$. This is far too great for immediate interpretation. The matrix therefore has to be processed further with the aim of pattern recognition, by component analysis (Chapter 5), cluster analysis (Chapter 6) or ranking (Section 5.6), for example.

A simple and yet most useful application is shown in Figure 4.4, lower part. The aim is to determine the *homogeneity* of a sample. From the data matrices in the upper row, the distance matrices are calculated and the mean Euclidean distance is computed. This is as a measure of distance or dissimilarity of the total set of relevés. Table 1, with a relatively low average distance, is hence more homogeneous than Table 2.

4.6 Assessing the quality of classifications

Under specific circumstances a resemblance matrix can be used to evaluate group patterns, as shown in Figure 4.5. This is a graphical representation of the similarities within and between 71 forest vegetation types in Switzerland, distinguished by Ellenberg & Klötzli (1972). The underlying data have been reconstructed from the original notes of the authors and the relevés found in the literature (Keller *et al.* 1998). From these 2533 relevés we know the corresponding classification used for definition of the forest types. The coefficients in Figure 4.5 are not just pairwise similarities, but average similarities between all relevés of the 71 groups involved. The diagonal elements are the average similarities within the groups and thus a measure of homogeneity, as explained in Figure 4.4.

Let us first look at some findings concerning the *diagonal* elements. There are examples of vegetation types exhibiting high internal homogeneity: the average similarity of relevés is high and therefore the symbol is large. Typical examples are forest types 49 (Equiseto-Abietetum), 56 (Sphagno-Piceetum typicum) and 70 (Rhododendro ferruginei-Pinetum montanae). The opposite is true for forest types 11 (Aro-Fagetum), 44 (Carici elongatae-Alnetum Glutinosae) and 64 (Cytiso-Pinetum silvestris). When inspecting all diagonal elements it becomes clear that the internal homogeneity of the different vegetation types varies amazingly: large symbols, indicating homogeneous groups, alternate with small symbols, indicating heterogeneity. In practice this means that there are types that are easy to recognize in the field (homogeneous ones) and others that are difficult to recognize (heterogeneous ones).

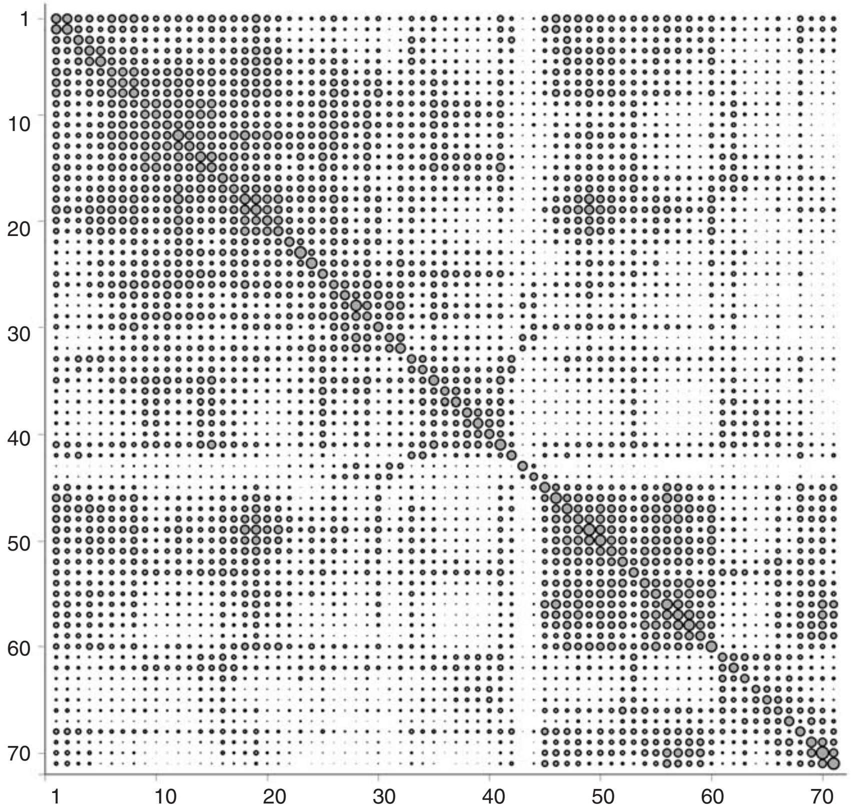


Figure 4.5 Similarities within and between the forest types of Switzerland according to Ellenberg & Klötzli (1972) based on the revision of Keller *et al.* (1998).

The *off-diagonal* elements show which of the vegetation types are difficult to distinguish from others (large symbols) and which are easily differentiated (small symbols). Forest types 1–21 form a block with large off-diagonal symbols. These are beech (*Fagus sylvatica*) forests. Differences in species composition between these types are minor and careful inspection of the species lists is required for proper identification. A similar example is seen in spruce and fir forests (*Picea abies* and *Abies alba*), forest types 45–60. Interestingly, there are also certain forest types which bridge the two blocks when taking species composition into account (types 19 and 49). As can be seen from this example, a similarity matrix presented graphically is an excellent tool for predicting problems in practical applications of classifications such as vegetation mapping. A real-world example is shown in Section 11.5, where the quality of a phytosociological classification system is evaluated.

5
Ordination

5.1 Why ordination?

Ordination is a graphical representation of the similarity of sampling units and/or attributes in resemblance space. An example of an ordination in two-dimensional space has been shown in Figure 4.1 (b), where the axes represent two plant species and the data points relevés. This graph displays the similarity of the two relevés involved, a rather trivial case as it presents the full configuration given in the raw numbers without improving insight into the system. Hence, ordination is a tool for analysing and visualizing complex data sets including a high number of sampling units with many attributes involved.

Recognizing patterns in large multivariate data sets inevitably means operating in a resemblance space of high dimension. When considering four species, for example, the configuration of resemblance is three-dimensional.

Data Analysis in Vegetation Ecology Otto Wildi

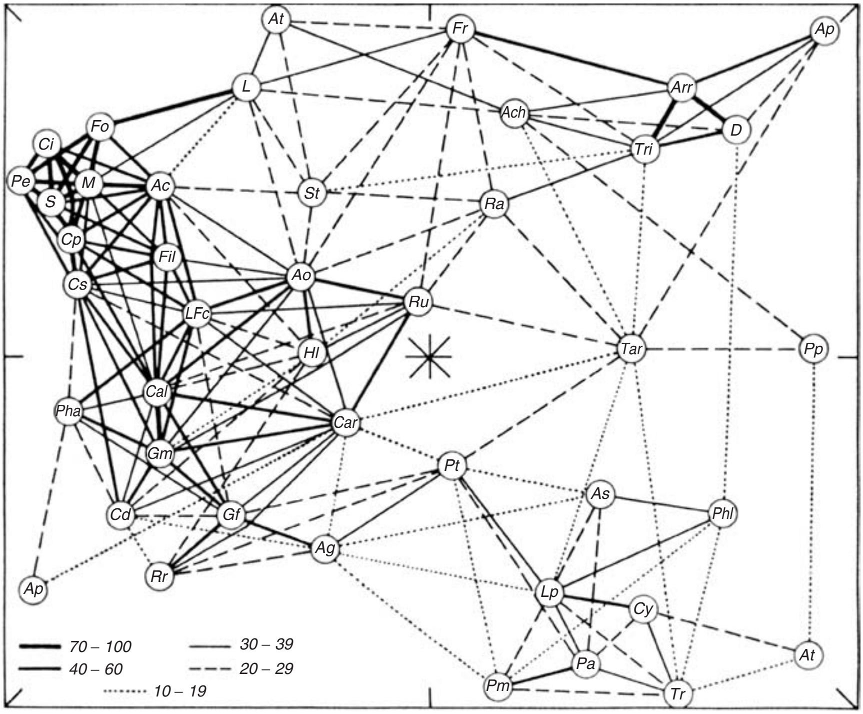
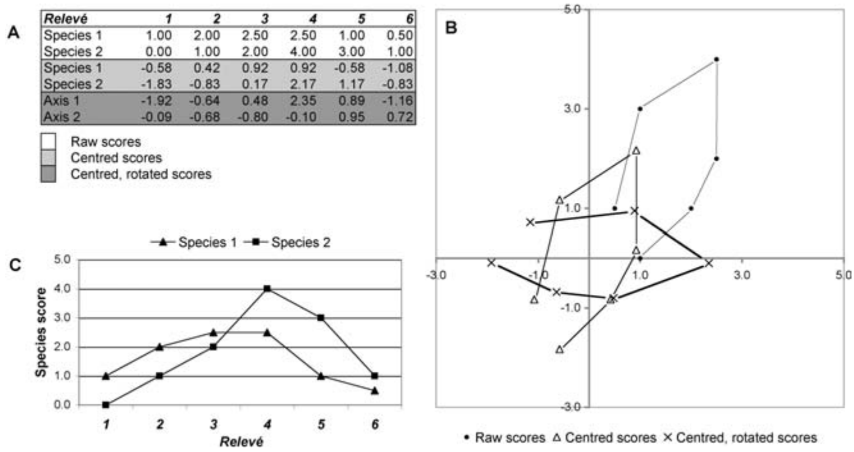


Figure 5.1 Three-dimensional representation of similarity relationships (Mueller-Dombois & Ellenberg 1974).

While four dimensions are already difficult to display graphically, in vegetation ecology large data sets often include hundreds of species, requiring hundreds of dimensions to be analysed. The aim of ordination is to reduce this number, to derive a graph that can be plotted or inspected dynamically as a three- or four-dimensional rotating point cloud. This is why ordination has always played a key role in vegetation ecology, and Figure 5.1 is a historical example of this effort, taken from Mueller-Dombois & Ellenberg (1974), illustrating the effort to reveal all the important relationships found in a multi-dimensional configuration. Although the methods have since evolved, the mode of interpretation has remained unchanged.

Most methods for gaining the desired insight into the similarity patterns of multivariate data sets roughly proceed through the steps illustrated in Figure 5.2:

- Centre the attributes in a data matrix to shift the origin of the new coordinate system into the centre of the point cloud.



Relevé	1	2	3	4	5	6
Species 1	1.00	2.00	2.50	2.50	1.00	0.50
Species 2	0.00	1.00	2.00	4.00	3.00	1.00
Species 1	-0.58	0.42	0.92	0.92	-0.58	-1.08
Species 2	-1.83	-0.83	0.17	2.17	1.17	-0.83
Axis 1	-1.92	-0.64	0.48	2.35	0.89	-1.16
Axis 2	-0.09	-0.68	-0.80	-0.10	0.95	0.72

Figure 5.2 Main functions of PCA. (a) The data table (artificial data) with original scores (species 1 and 2, white background), centred species vectors (species 1 and 2, light grey background) and centred as well as rotated scores (axes 1 and 2, dark grey). (b) Representation of the point configuration in x- and y-space. (c) Species scores as a function of relevé order (response functions).

- Rotate the point cloud such that the maximum possible variance is found along the first axis.
- Continue rotating the point cloud, while keeping the first axis fixed, and maximize the remaining variance on the second axis.
- Continue this process until all the axes are processed.
- Represent the result graphically by omitting higher dimensions.

In this procedure, conforming to principal component analysis (PCA), the point pattern hidden in raw data is maintained. As will be shown later, there are methods that are changing the pattern. Hence, choosing the proper method and understanding what it does to data is crucial, and offers flexibility in defining the goal of the analysis.

5.2 Principal component analysis (PCA)

Principal component analysis is a basic procedure that operates as described in Section 5.1. First of all, it strictly relies on *linear correlation* of attributes.

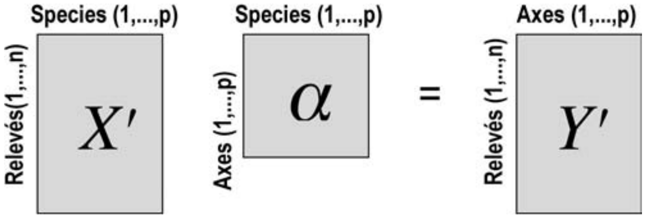


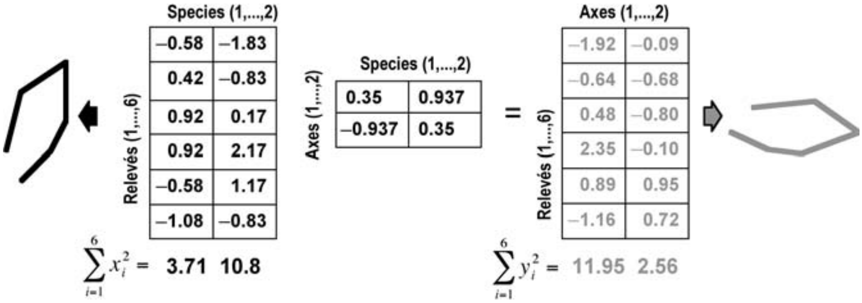
Figure 5.3 Projecting data into ordination space in PCA. The scalar product of the species by relevé matrix X' and the species by axes matrix α (the Eigenvectors) yields the axes by relevé matrix Y' (the ordination scores).

It operates in the orthogonal Euclidean space and searches for useful projections of point clouds. Because it is based on the concept of variance partitioning and the variance is maximized along the axes, the result it finds is reproducible – even when using different computers and computer programs (e.g. independent of any initial order of data). Orthogonality (i.e. absence of correlation) also means that the variance carried by the axes is *additive*. Whatever projection of a point is chosen, the variance explained by the graph is equal to the sum of the explanatory power of the axes involved.

Whereas centring the data is a trivial task, finding the best method of rotation is more demanding. As seen in Figure 4.1, metric data can be used directly as coordinates, where data points are sampling units and species are axes. In mathematical terms, generating a new projection is just a transformation of a coordinate system and is achieved by multiplying two matrices. In the case of PCA, this is shown in Figure 5.3, where matrix X contains the original data and X' the centred. This is multiplied by a new square matrix α, according to:

$$X'^{n*p}\alpha^{p*p} = Y'^{n*p} \tag{5.1}$$

Matrix α holds the Eigenvectors. It is a squared matrix with the number of *species* by the number of *axes* as dimensions. X' and α have one dimension in common, the species. The new matrix Y' still has the relevés as rows, but the attributes are now the new axes. The matrix of Eigenvectors α is obtained from the original data by Eigenanalysis (Batschelet 1975), yielding the desired properties of the final result – orthogonality of axes – and maximizing variance on first axes. Eigenanalysis is performed on the variance or correlation matrix of the species, and as a result there are as many Eigenvalues as there are species (although some may be zero).



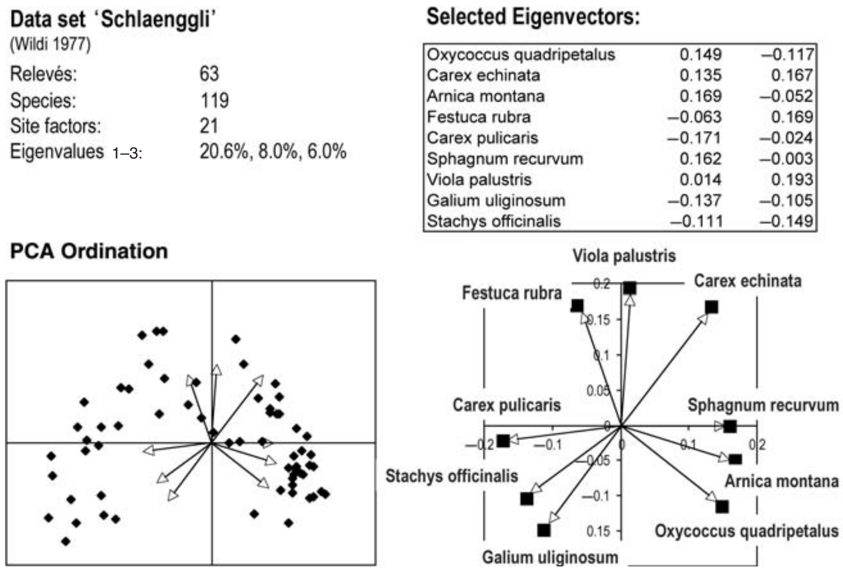
$$\begin{array}{c|cc} & \multicolumn{2}{c}{\text{Species }(1,\ldots,2)} \\ \hline \text{Relevés }(1,\ldots,6) & -0.58 & -1.83 \\ & 0.42 & -0.83 \\ & 0.92 & 0.17 \\ & 0.92 & 2.17 \\ & -0.58 & 1.17 \\ & -1.08 & -0.83 \\ \hline \sum_{i=1}^{6} x_i^2 = & 3.71 & 10.8 \end{array} \quad \begin{array}{c|cc} & \multicolumn{2}{c}{\text{Species }(1,\ldots,2)} \\ \hline \text{Axes }(1,\ldots,2) & 0.35 & 0.937 \\ & -0.937 & 0.35 \end{array} \quad = \quad \begin{array}{c|cc} & \multicolumn{2}{c}{\text{Axes }(1,\ldots,2)} \\ \hline \text{Relevés }(1,\ldots,6) & -1.92 & -0.09 \\ & -0.64 & -0.68 \\ & 0.48 & -0.80 \\ & 2.35 & -0.10 \\ & 0.89 & 0.95 \\ & -1.16 & 0.72 \\ \hline \sum_{i=1}^{6} y_i^2 = & 11.95 & 2.56 \end{array}$$

Figure 5.4 Numerical example of PCA. The centred data matrix (left) is multiplied by the matrix of Eigenvectors (centre) to yield the ordination coordinates (right). The variances of the ordination axes are the corresponding Eigenvalues.

The numerical example illustrated in Figure 5.4 is carried over from Figure 5.2. The elements of the Eigenvector matrix are correlation coefficients (by definition) between the original attributes (the species) and the new ordination axes. Element 0.35 signifies that the first species has a positive correlation of $r = 0.35$ with the first ordination axis. The correlation with the second axis is $r = -0.937$. The second species correlates with the first axis by $r = 0.937$ and with the second axis by $r = 0.35$. Hence, the Eigenvectors are a useful tool for interpreting the final ordination.

In Figure 5.4 the variances of the attributes are also shown. In the original data, species 2 has the highest variance with 10.8. According to the definitions in PCA, the highest variance in the ordination is attributed to the first axis. This is 11.95 and is the first *Eigenvalue* in the Eigenanalysis. Because variance on any axis is a linear combination of the variance of many species, it generally exceeds the variance of any individual species. However, the total variance remains unchanged as the point pattern as a whole is not affected by PCA, and only its projection is adjusted.

The result of PCA deserves careful interpretation, as illustrated in Figure 5.5, a data set consisting of 63 sampling units (relevés) and 119 attributes (species) (‘Schlaenggli’, see Appendix B). The environmental factors are not analysed in this example (but will be in later sections). The absolute magnitudes of the Eigenvalues depend on the size of the sample and are therefore not useful in the interpretation. The relative proportions are most crucial, as they inform us about the explanatory power of axes. Here, the x-axis explains 20.6% of the variance and the y-axis 8.0%. The ordination shown in Figure 5.5 uses the scores of the first two axes of PCA as coordinates, hence explaining 28.6% of the total variance, as the



Data set 'Schlaenggli'
(Wildi 1977)

Relevés:	63
Species:	119
Site factors:	21
Eigenvalues 1–3:	20.6%, 8.0%, 6.0%

Selected Eigenvectors:

Oxycoccus quadripetalus	0.149	−0.117
Carex echinata	0.135	0.167
Arnica montana	0.169	−0.052
Festuca rubra	−0.063	0.169
Carex pulicaris	−0.171	−0.024
Sphagnum recurvum	0.162	−0.003
Viola palustris	0.014	0.193
Galium uliginosum	−0.137	−0.105
Stachys officinalis	−0.111	−0.149

Figure 5.5 Main results of a PCA using real data. The Eigenvectors are used to help in the interpretation of the ordination by pointing in the direction of the centres of species occurrences.

variances are strictly additive. A three-dimensional plot would explain another 6% – a total of 34.6%.

Is 28.6% explained variance good or poor for a two-dimensional ordination? Peres-Neto *et al.* (2005) have written a review on papers discussing the issue of 'nontrivial axes'. The authors suggest a randomization test to identify the number of relevant axes. It may be safe, however, to screen for patterns beyond this number. The proportion of explained variance depends on the type of data analysed and of course on the number of axes considered for viewing. The total dimensionality of the data set is 63 (although there are 119 species involved, 63 data points can be presented in a maximum of 63 dimension without loss of information. A detailed inspection of the Eigenvalues would show that all beyond 63 are zero!). For this size of sample, experience suggests that 28.6% usually reveals the dominating pattern, which in this case is a classical horseshoe, indicating that a (nonlinear) gradient exists. However, it is good practice to inspect the third and the fourth dimension as well. From many more examples it can be infered that data sets of several hundreds of relevés usually result in a first Eigenvalue explaining around 10% of the total variance or even less (see Chapter 11 for examples). As a rule the explanatory power of the first axis will decrease as the sample size is increasing.

The interpretation of the point cloud is simple as it displays the similarity space, with the only complication arising from the high dimensionality. If any two data points are in close neighbourhood then they are similar. However, they may still be distant in the third dimension, which is not visible. Even in simple cases it is suggested that a computer program which displays three-dimensional point clouds be used, either as a stereogram or as a spinning graph.

For proper interpretation of PCA results the Eigenvectors (also known as component coefficients) have to be considered as well (Figure 5.5). As explained above, they are the correlations of the species with the ordination axes. Due to the very high dimensionality of the resemblance space, most correlations are rather low, with none even reaching $r = 0.2$. Geometrically, correlation coefficients are cosines of vectors (see Figure 4.3). Therefore, they can be used for drawing species vectors (Figure 5.5, lower-right graph). Their scaling differs entirely from the ordination diagram, but the graph of vectors can be enlarged or reduced to fit into ordination by superimposing the origins of the two diagrams (lower-left graph). The species arrows now point in the directions of their centres of occurrence. As done here, selecting just a few species for display will avoid a proliferation of information in the graph.

The detailed interpretation of Figure 5.5 proceeds as follows: the relevés (i.e data points) in the lower-right quadrant of the ordination are characterized by high values of *Oxycoccus quadripetalus*. In the relevés in the lower-left quadrant *Stachys officinalis* and *Galium uliginosum* occur frequently. The horseshoe-shaped point cloud reveals a gradient from the lower left (with high soil pH values, not shown here) towards the lower right (with low pH values). Assessing the relationship to pH, however, requires other methods such as constrained ordination (Section 7.5).

5.3 Principal coordinates analysis (PCOA)

This method is similar to PCA, but since it accepts almost any kind of similarity or distance measure, it is of broad practical use. The method, first published by Gower (1966), is not only known as *principal coordinates analysis* but also as *principal axis analysis* and *metric multidimensional scaling*. In PCOA, when the relevés are ordinated, the Eigenvalues and Eigenvectors are derived from the similarity matrix of the relevés. This differs from PCA, where the species-similarity matrix is used to derive the coordinates of the relevés. As a minor disadvantage, there are no Eigenvectors available to help in the interpretation of the ordination.

The position of the data points is initially defined by a distance or similarity matrix, which may be either metric or nonmetric. If distances are given then the elements d_{ij} are transformed according to:

$$s_{ij} = -\frac{1}{2}d_{ij}^2 \tag{5.2}$$

The elements of S are interpreted as direction cosine and have to be adjusted for range. The new matrix A has the elements:

$$a_{ij} = s_{ij} - \overline{s}_i - \overline{s}_j + \overline{s} \tag{5.3}$$

where $\overline{s}_i$ and $\overline{s}_j$ are the row and column means of S and $\overline{s}$ is the same of the grand total. Then the Eigenvalues, $\lambda_1, \ldots, \lambda_n$, and the corresponding Eigenvectors, $\beta_1, \ldots, \beta_n$, of A are found. The Eigenvectors are adjusted to the Eigenvalues to satisfy the condition:

$$\beta_{1i}^2 + \ldots + \beta_{pi}^2 = \lambda_i^2 \tag{5.4}$$

These are now the ordination coordinates. The question arises whether and how they deviate from PCA coordinates of the same data set. This is shown in an example using the data set previously presented in Figure 5.5. The cover-abundance scores are first changed into a rank scale and then scalar is transformed according to $x' = x^{0.5}$. For PCA, the correlation matrix of the species is computed; for PCOA it is the matrix of the relevés. The resulting Eigenvalues are as follows:

	PCA %	PCOA %	PCA, variance	PCOA, variance
λ_1	20.62	27.96	24.5	12.2
λ_2	8.07	9.43	9.61	4.11
λ_3	6.07	5.99	7.23	2.61

Clearly, the Eigenvalues differ in size and proportion. The resulting ordinations are shown in Figure 5.6. The two point clouds are superimposed after (heuristic) linear adjustment of the scale (scores of PCOA are multiplied by a factor of 5.57). It can be seen that the overall shape of the point cloud, a horseshoe, really is the same. The individual points, however, are slightly displaced. Since PCA reproduces the geometrical configuration of points,

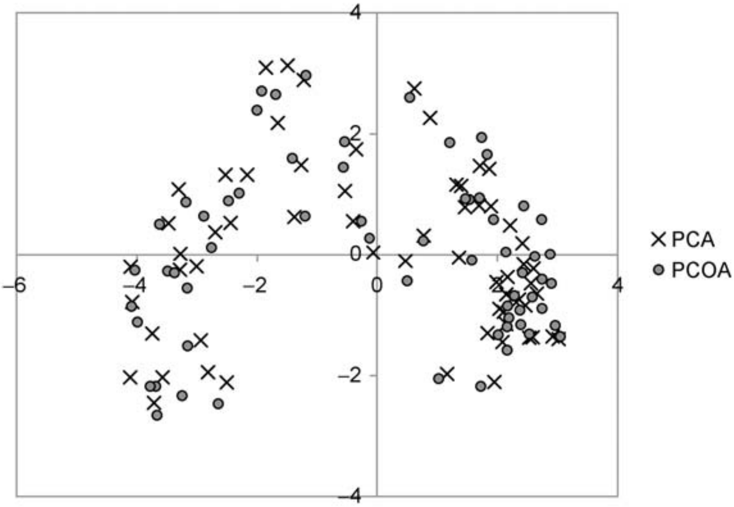


Figure 5.6 Comparison of PCA and PCOA using the 'Schlaenggli' data set of 63 relevés. Data points from PCA (crosses) and from PCOA (triangles) superimposed after adjustment of scale.

there has to be a minor (but for ecological interpretations, unimportant) distortion in the ordination of PCOA.

Why this distortion? Depending on the initial resemblance measure used, matrix *A* (Formula (5.3)) is usually not strictly metric. PCOA will then extract the metric portion from *A* and the corresponding positive Eigenvalues express the explanatory power of the axes. The remaining nonmetric part appears in the form of negative Eigenvalues. This cannot be displayed in an ordination diagram.

5.4 Correspondence analysis (CA)

Correspondence analysis is distinct from PCA and PCOA due to the intrinsic assumptions and the corresponding transformations applied. In CA the data table is assumed to be a contingency table; that is, a table containing counts. Some of the very many alternative names for CA reflect this fact:

- Contingency table analysis (Fisher 1940)

- Analyse factorielle des correpondances (Benzécri 1969)

- Reciprocal averaging (Hill 1973)
- Reciprocal ordering (Orlóci 1978)
- Dual scaling (Nishisato 1980)

The elements of the initial data table, f_{hj}, are frequency counts, which are then relativized by the row and column sums, and from this deviations from expectations, u_{hj}, are derived according to:

$$u_{hj} = \frac{f_{hj}}{\sqrt{f_{h.}f_{.j}}} - \frac{\sqrt{f_{h.}f_{.j}}}{f_{..}} \tag{5.5}$$

The notation used is shown in Table 5.1 (a). Examples (b) and (c) illustrate the typical effects of this transformation. (b) points to the fact that the scores analysed are deviations from expectation and not the raw information. The first element, $f_{11} = 3$, turns out to be a relatively low score and the final $u_{11} = -0.05$ is negative as it is below expectation. It is an element of a row with a fairly high marginal total ($f_{1.} = 4$) and also of a column of high marginal total ($f_{.1} = 4$). The elements $f_{12} = 1$ and $f_{21} = 1$, in contrast, are above expectation. It is important to note that CA will use these derived values and not the original scores!

A typical effect of the adjustment by rows and columns is demonstrated in Table 5.1 (c). What matters is the proportions of the elements of the data vectors. CA causes species 1 and species 2 to reflect the same pattern since the proportion of the scores are the same. Similarly, relevés 1 and 2 are rated identical. In terms of CA, the two relevés and the two species are identical and no usable information can be analysed.

Like in some variants of PCA, the calculations are now based on a matrix of product moments, $S = UU'$, computed for p attributes with a characteristic element:

$$s_{hi} = \sum_{j=1}^{s} u_{hj}u_{ij} \tag{5.6}$$

From this similarity matrix the non-zero Eigenvalues, $\lambda_1, \ldots, \lambda_t$, and the associated Eigenvectors, $\alpha_1, \ldots, \alpha_t$, are extracted. The Eigenvalues have the form of correlation coefficients: the m^{th} Eigenvalue is the square of the m^{th} canonical correlation. The Eigenvector matrix A, after the adjustment shown below, gives ordination scores for the attributes.

Table 5.1 (a) Notation used in correspondence analysis. (b) An illustrative numerical example. (c) An example with no information content in terms of CA. Frequency tables are in the left row, deviations from expectation in the right row.

A

	rel_1	rel_2	
sp_1	f_{11}	f_{12}	$f_{1.}$
sp_2	f_{21}	f_{22}	$f_{2.}$
	$f_{.1}$	$f_{.2}$	$f_{..}$

	rel_1	rel_2	
sp_1	u_{11}	u_{12}	$u_{1.}$
sp_2	u_{21}	u_{22}	$u_{2.}$
	$u_{.1}$	$u_{.2}$	$u_{..}$

B

	rel_1	rel_2	
sp_1	3	1	4
sp_2	1	0	1
	4	1	5

	rel_1	rel_2	
sp_1	–0.05	0.10	0.05
sp_2	0.10	–0.20	–0.10
	0.05	–0.10	–0.05

C

	rel_1	rel_2	
sp_1	3	6	9
sp_2	1	2	3
	4	8	12

	rel_1	rel_2	
sp_1	0.00	0.00	0.00
sp_2	0.00	0.00	0.00
	0.00	0.00	0.00

As explained in Legendre & Legendre (1998, p. 456), there are different ways to scale the scores. For ecological applications it is most appropriate to choose an adjustment that allows the joint plot of row (relevé) and column (species) scores. From the Eigenvectors, the species scores X are derived directly by weighting with the square root of the inverse of the marginal totals:

$$x_{hm} = \frac{(\alpha_{hm} - \overline{\alpha}_m)}{\left[\sum_{h=1}^{p} (\alpha_{hm} - \overline{\alpha}_m)^2\right]^{1/2}} \sqrt{\frac{f_{..}}{f_{h.}}}. \tag{5.7}$$

This formula also involves standardization of the Eigenvectors to fulfil the following conditions:

$$\sum_{h=1}^{p} \alpha_{hm}^2 = 1 \qquad \text{and} \qquad \sum_{h=1}^{p} \sqrt{f_{h.}}\alpha_{hm} = 0. \tag{5.8}$$

To compute the relevé matrix Y, one could transpose the data matrix F and repeat all the computations. If so, one would observe that the resulting Eigenvalues were exactly the same. There is a direct way to derive the relevé scores from the species scores:

$$y_{jm} = \sum_{h=1}^{p} \frac{f_{hj} x_{hm}}{f_{.j} R_m} \tag{5.9}$$

where R_m is the m^{th} canonical correlation. If all the calculations are done on the transposed data matrix then the results will be entirely identical. When analysing large data sets, carrying out the analysis on the smaller similarity matrix will save computation time.

The difference between PCA (or PCOA) and CA is considerable in terms of the content and shape of the point cloud. Using the same data set, a CA and a PCA ordination are computed and displayed in Figure 5.7 for comparison. Superimposed is the same classification of relevés. As in many other cases, it can be observed that the gradient displayed in CA is v-shaped, whereas the gradient in PCA is u-shaped. Also, the two-dimensional resolution of the classification (the distinction of groups) is usually somewhat better in PCA than in CA. The order of the relevés along the main gradient is roughly the same and I therefore conclude that both of the methods reveal the underlying pattern.

CA has some unpleasant properties to be kept in mind when using it. First, it is more sensitive to outliers (see Section 11.3) than other ordination

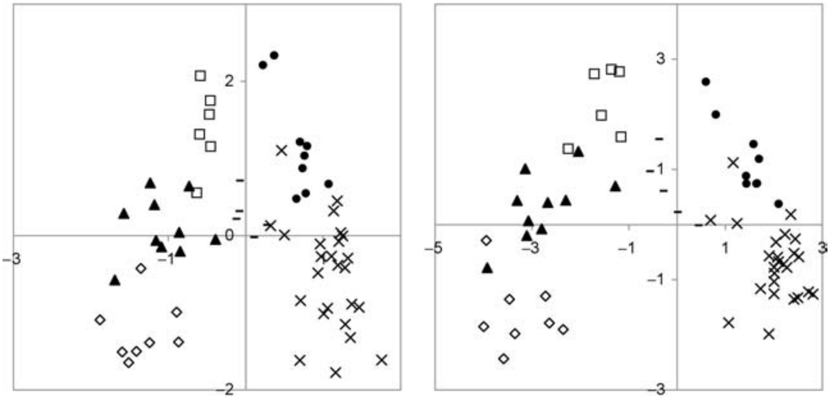


Figure 5.7 Comparison of CA and PCA. Data points are identically classified. 'Schlaenggli' data set used (Appendix B).

methods. Unlike in PCA and PCOA, the range of the coordinates increases with higher dimension (and lower Eigenvalue). It is good practice to restrict adjustment of scales of the ordination axes to the ones used for plotting.

5.5 The horseshoe or arch effect

5.5.1 Origin and remedies

The fact that relevés originating from an environmental gradient are curve-shaped (e.g. Figure 2.3) led to much confusion in the past. In many papers and textbooks it was rated a deficiency of the ordination method used, and much emphasis was put on researching 'better' methods. However, the origin of the problem can easily be found when inspecting the original data. When biological parameters are used as axes, such as species performance, these parameters usually correlate nonlinearly. A plot of performance against an environmental variable typically results in a bell-shaped (Gaussian) curve. Using two such variables as axes in a two-dimensional graph yields the arch or horseshoe (Figure 5.8). As a result, the extreme points of a gradient are located in close neighbourhood. This is because at the extremes of the gradient both variables react negatively to site conditions. Introducing a third variable may correct this, but still not in the desired manner: the configurations then usually look like spirals in three dimensions. In fact, there is no remedy to this!

Not all methods deliver the same shape for a gradient when generating a point cloud. There are two reasons for this. First, the data may intrinsically be transformed, which alters the similarity space. An example of this is *correspondence analysis*. Other methods directly alter the geometry of the

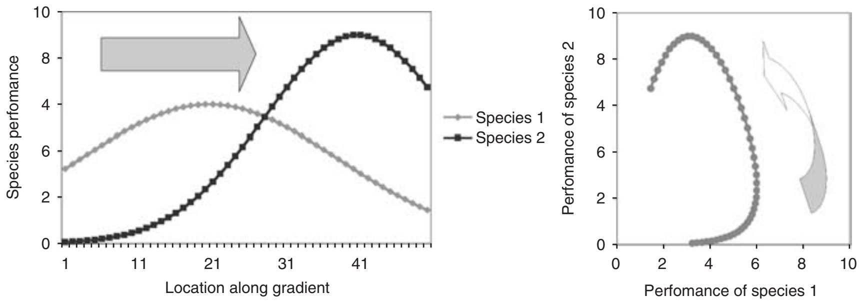


Figure 5.8 Performance of two species along a hypothetical environmental gradient (left). A species-by-species plot delivers an arch-shaped figure (right).

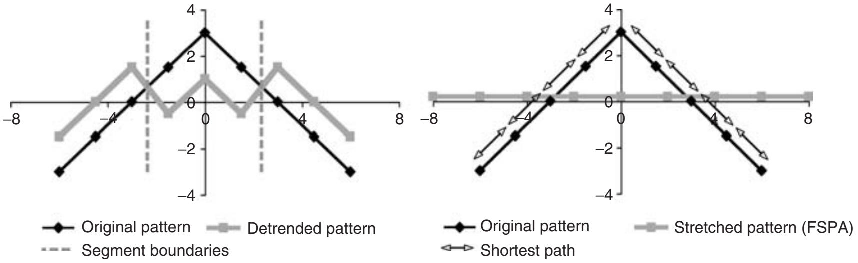


Figure 5.9 The principle of detrending by segments (left, according to Legendre & Legendre 1998, p. 460) and flexible shortest-path adjustment FSPA (Wildi & Orlóci 1996).

ordination. The best known example is *detrended correspondence analysis (DCA)*, first described by Hill (1979a). The principle of detrending by segments is shown in Figure 5.9, left-hand side, a simplified version of the the sketch in Legendre & Legendre (1998, p. 460). The original figure is a triangle (dark line). This is divided into three segments (vertical dashed lines). The segments are then shifted by centring the y-coordinates. The example in Figure 5.9 illustrates how this may lead to serious local distortions. The application is best justified in the presence of a classical horseshoe and in the absence of too much statistical noise. Detrending by polynomials is a smooth version of this. But it also carries the risk of uncontrolled alteration of the similarity pattern.

An alternative to detrending is flexible shortest-path adjustment, FSPA (Wildi & Orlóci 1996), in which the large distances in the point cloud are erased and recalculated via intermediate data points. The resulting distance space is nonmetric. To generate an ordination, principal coordinates analysis is used (Section 5.3), retrieving metric ordination axes. At first glance this looks like the ideal solution. However, the effect is difficult to control if there is some noise in the data and it should be used with great care.

FSPA is a nice example of the hidden origin of many methods in data analysis. It was used in a paper in *Ecology* by Bradfield & Kenkel (1987) citing Floyd (1962), but like many other straightforward ideas, it has since been reinvented for different applications. This can be seen in two papers appearing in *Science* (Tenenbaum *et al.* 2000, Roweis & Saul 2000), where more nice examples illustrate the idea.

The two methods shown so far both operate on an already existing ordination and attempt to improve it. Nonmetric multidimensional scaling (NMDS) is an ordination method by itself, but also alters an already existing ordination by improving it through iteration (Legendre & Legendre 1998). The new configuration of data points is optimized for a limited number of dimensions, perhaps two or three. Not all computer programs do this the same way and the results may be difficult to reproduce. Apparently the method was invented around 1960 (Shepard 1962) and became popular through a paper by Clarke (1993). The recent availability of tremendous computing power has also added to its popularity (see for example Belden & Pallardy 2009, Rogers *et al.* 2009, Stromberg *et al.* 2009).

5.5.2 Comparing DCA, FSPA and NMDS

Comparisons presented in this chapter concern ‘adaptive’ methods aimed at improving on already existing ordinations, mainly correcting the horseshoe. This imposes at least two restrictions on comparison. First, resulting ordinations have to be opposed to the initial configurations: for example, detrended correspondence analysis (DCA) versus ordinary correspondence analysis (CA), nonmetric multidimensional scaling (NMDS) versus principal components analysis (PCA) and flexible shortest-path adjustment (FSPA) versus its underlying principal coordinates analysis (PCOA) ordination. Second, the three methods are flexible by nature, offering options for conducting the process by various means. I am presenting just one among an almost infinite number of solutions, and readers should be aware that running computer programs differently may alter the results.

Another issue is measuring quality; that is, performance of ordinations. Generally performance of an ordination is high when the proportion of variance explained by environmental factors is high. That is what constrained ordination really measures (Section 7.5), but the results primarily express a property of data used; therefore the comparison eventually becomes an evaluation of data rather than of method. Ordinations can always be compared by stress functions, as shown in Section 7.3.1. However, these consider regular patterns and statistical noise simultaneously, whereas in practice priority is usually given to revealing striking patterns. That is what I do below, taking a small data set (‘Schlaenggli’) exhibiting an obvious pattern (a gradient) and revealing another – a group pattern – with the aim of displaying both for visual inspection. Using minimum-variance clustering analysis the number of relevé groups is set to three to facilitate distinction of symbols.

The first example compares CA and DCA (Figure 5.10). In DCA 26 segments are used and 4 iteration cycles applied. Only relevé data points are displayed. Both ordinations nicely resolve the group pattern and DCA succeeds in stretching the horseshoe. Hence, while maintaining the order along the x-axis, the y-axis is compressed by DCA.

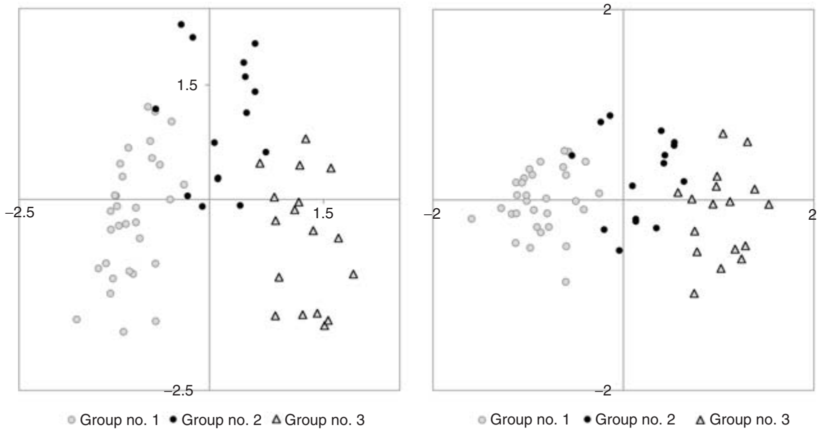


Figure 5.10 Comparison of CA (left) and DCA (right). Data point groups are from minimum-variance cluster analysis.

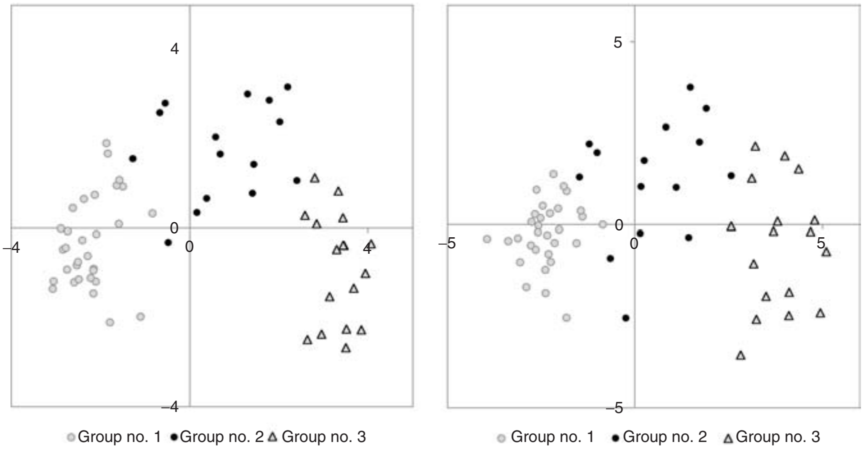


Figure 5.11 Comparison of PCA (left) and NMDS (right). Data point groups are from minimum-variance cluster analysis.

The second example is about NMDS, with the initial configuration being PCA derived from a correlation matrix. The NMDS ordination is optimized for two dimensions (Figure 5.11). The strong horseshoe from PCA ordination is still visible, as is the group structure. While NMDS does not alter the ordination too much, it must be noted that replacing the correlation coefficient by the frequently used Bray–Curtis index (see for example Gauch 1982) would change the result considerably.

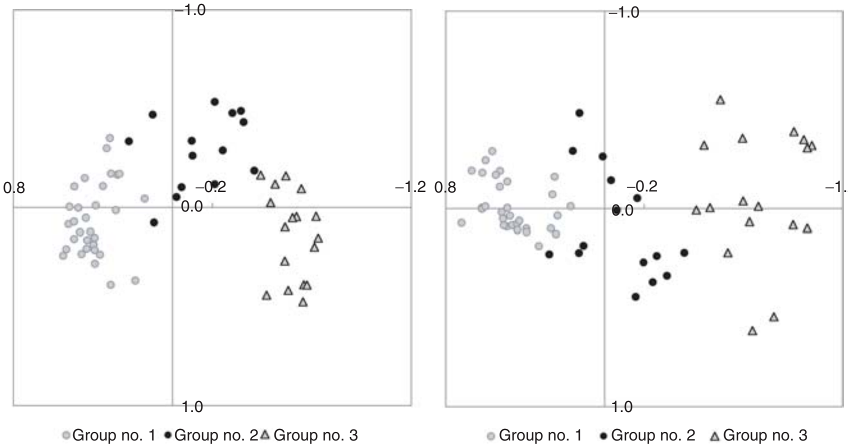


Figure 5.12 Comparison of PCOA (left) and flexible shortest-path adjustment (right). Data point groups are from minimum-variance cluster analysis.

The third example compares PCOA ordination with its analogue processed by FSPA. Resemblance is the correlation coefficient (Section 4.4), of which FSPA takes the one-complement, a distance measure. I choose 47.8% of the total of 1953 distances to be recalculated in order to change the pattern remarkably. As can be seen in Figure 5.12, FSPA stretches the horseshoe while retaining the group structure. Because this predominantly affects the first axis, the first Eigenvalue accounts for an astonishing $\lambda_1 = 53.14\%$ of total variance, and the two dimensions for 62.17%.

It does not come as a surprise that all methods produce usable results. A main issue in all examples is the flexibility with which initial configurations, alternative pathways and the number of iterations can be chosen. This adds uncertainty to the methods not existent in PCA, PCOA and CA. While the need for stretching of a horseshoe is debatable, it becomes clear that the quality of the data used is far more important than the selection of the best ordination method.

5.6 Ranking by orthogonal components

5.6.1 Method

The term ‘ranking’ generally implies an evaluation of either sampling units or attributes. A rank order is established based on some measurable criteria. This serves the data reduction in an efficient way.

The ranking procedure proposed here is based on independent components of the sum of squares (Orlóci 1973, 1978) and is therefore closely related to PCA. The variables chosen should explain as much of the total variance as possible and should be independent (not correlated). The main difference compared to PCA is that the variables are chosen from among the original attributes (or sampling units) and not generated by linear combination. This makes ranking somewhat less efficient than in PCA, but also easier to interpret. The algorithm presented below is for ranking attributes (Wildi & Orlóci 1996):

1 The data, X, is centred within the attributes in order to obtain a new matrix, A, with elements:

$$A_{hj} = \frac{X_{hj} - \overline{X}_h}{Q_h} \tag{5.10}$$

where $\overline{X}_h$ and Q_h are the mean and a factor of adjustment, respectively, in attribute h. X_{hj} is the value of attribute h in relevé j. For Q_h I refer to Table 3.1, where it can be seen that this choice affects the type of the similarity coefficient, S.

2 Cross-products, $S = AA'$, are computed. A characteristic element is given by:

$$S_{hi} = \sum_{j=1}^{n} A_{hj} A_{ij} \tag{5.11}$$

where n indicates the number of relevés.

3 Dispersions and highest values are calculated:

$$SS = max\left(\sum_{h=1}^{p} \frac{S_{hi}^2}{S_{hh}}\right) \qquad h = 1, \ldots, p \tag{5.12}$$

where p is the number of attributes. The quantity, SS, is a measure of redundancy in the sample of p attributes with respect to attribute m. Rank 1 is declared for attribute m associated with SS.

4 Residuals are computed:

$$S_{hi} := S_{hi} - Y_{hm}Y_{im} \text{ for any } h, i = 1, \ldots, p \tag{5.13}$$

in which:

$$Y_{hm} = \frac{S_{hm}}{\sqrt{S_{mm}}} \text{ and } Y_{im} = \frac{S_{im}}{\sqrt{S_{mm}}} \tag{5.14}$$

5 Computation of a new value for SS from the elements of the residual, S, and declaration of rank 2 for the corresponding attributes. Then repeat steps 3 and 4 as many times as necessary until all attributes are ranked.

An interactive version of this has been proposed by Wildi (1984), concerning step 3. Instead of selecting the variable with the highest value for SS, the choice is left to the user. This reduces the efficiency of the procedure but allows it to omit attributes that do not seem feasible for the application in mind, such as species that are difficult to identify.

The RANK algorithm is most efficient when applied to data sets with a very high number of attributes. This can be seen from the following example, which demonstrates the method of function and the interpretation.

5.6.2 A numerical example

This example shall demonstrate the high efficiency of the RANK method. It is shown in Table 5.2, where the results are also summarized. The procedure starts by calculating the resemblance matrix of species, R. This implies

Table 5.2 Data set for illustrating the RANK algorithm.

relevé	1	2	3	4	rank no.	expl. variance, %
species 1	2	2	1		4.	0.0
species 2	2	1	1		2.	17.0
species 3			1	1	1.	78.5
species 4			2	1	3.	4.5

standardization of the vectors (Formula 5.10) and the computation of the cross product (Formula 5.11). We get:

$$\mathbf{R} = \begin{pmatrix} 1.0 & 0.85 & -0.91 & -0.64 \\ 0.85 & 1.0 & -0.71 & -0.34 \\ -0.91 & -0.71 & 1.0 & 0.91 \\ -0.64 & -0.43 & 0.91 & 1.0 \end{pmatrix}$$

As can be expected from the data, species 1 and 2 as well as 3 and 4 are highly correlated ($r = 0.85$ and 0.91 respectively). The dispersions (variances) explained by the individual attributes are found in the respective rows or columns of the correlation matrix R (see Formula 5.12). The correlation coefficient simplifies matters as S_{hh}, the diagonal values, are always equal to 1. The variances the attributes account for are:

$$SS_1 = \frac{1}{1}\left[(1.0)^2 + (0.85)^2 + (-0.91)^2 + (-0.64)^2\right] = 2.95$$

$$SS_2 = \frac{1}{1}\left[(0.85)^2 + (1.0)^2 + (-0.71)^2 + (-0.43)^2\right] = 2.41$$

$$SS_3 = \frac{1}{1}\left[(-0.91)^2 + (-0.71)^2 + (1.0)^2 + (0.91)^2\right] = 3.14$$

$$SS_4 = \frac{1}{1}\left[(-0.64)^2 + (-0.43)^2 + (0.91)^2 + (1.0)^2\right] = 2.41$$

Species number 3 has the highest explanatory power and will get rank no. 1. It is important to note that the other species achieve high values as well. Taking species number 1 instead of number 3, for example, would reduce the efficiency only moderately. This situation is typical for high-dimensional vegetation data.

The correlation matrix is now reduced by the fraction of variance explained by species 3 according to Formula 5.12:

$$r'_{11} = 1.0 - (-0.91 * -0.91) = 0.17$$

$$r'_{12} = 0.85 - (-0.71 * -0.91) = 0.20$$

$$r'_{13} = -0.91 - (1.0 * -0.91) = 0.0$$

$$r'_{14} = -0.64 - (-0.91 * 0.91) = 0.19$$

In the new, reduced matrix R' the rows and columns related to species 3 are now all zero:

$$\mathbf{R}' = \begin{pmatrix} 0.17 & 0.20 & 0.0 & 0.19 \\ 0.20 & 0.50 & 0.0 & 0.21 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.19 & 0.21 & 0.0 & 0.17 \end{pmatrix}$$

The procedure according to Formula 5.12 is now applied to matrix R'. As can be seen in Table 5.2 the variance explained decreases rapidly, confirming the efficiency of the method. It even turns out that species number 1 no longer contributes to the total variance, indicating that the dimension of the total resemblance matrix is equal to 3 only.

5.6.3 A sampling design based on RANK (example)

The method proved to be extremely useful in selecting typical plots for permanent observation. As these plots account for maximum covariation in the data, they are considered ‘typical’. In order to design an efficient plan for permanent plot research in an area (Wildi 1990) the following steps are taken:

1 Complete an initial investigation of the area. The sampling intensity should be sufficiently high to give an accurate account of types and gradients.

2 Select a low number of representative plots using RANK. Survey and analyse these periodically.

3 As soon as a marked trend occurs in the selected plots, re-survey the entire sample.

This example aims to demonstrate the efficiency of the method. The criterion for efficiency is the proportion of variance explained by the plots chosen for permanent survey. This depends on the correlations occurring in any one data set. The ‘Schlaenggli’ data set is a typical example of a moderate-sized sample.

The result of the analysis is shown in Table 5.3. It is based on the correlation of 119 plant species occurring within 63 sampling units. There are

Table 5.3 Ranking relevés of the 'Schlaenggli' data set.

Rank no.	Plot no.	Explained variance (%)	Cumulative variance (%)	pH peat
1	520	26.23	26.23	4.6
2	560	10.61	36.84	6.2
3	527	5.55	42.39	4.6
4	546	4.10	46.49	5.5
5	553	4.05	50.54	6.2
...	...	...	...	...
10	557	1.84	62.61	6.3
...	...	...	...	...
24	511	0.95	80.08	4.9

just 5 out of the 63 sampling units needed to account for 50% of the total variation. As the method intends to maximize co-variation, the subsample of 5 is considered a typical representation of the entire sample.

Increasing the subsample to 10, as shown in Table 5.3, increases the explained variance to 62%. Obviously, the contribution of any further plot does not add much to the efficiency of the survey. 24 plots are needed for 80% of the variance explained and all 63 for 100%.

As the plots are as independent as possible, RANK takes these from contrasting locations within the investigation area. In Figure 5.13 the five plots listed in Table 5.3 are shown within the sampling plan. The fact that they represent contrasting plots directly translates to their spatial location: They are dispersed all over the sampling area. Due to spatial autocorrelation (Section 7.3.3) it is rather unlikely that neighbouring (and hence similar) plots occur in the first few ranks. The independence of the sampling units can also be observed in the pH values. The orthogonality of ranks is reflected in large pH steps, from 4.6 up to 6.2 and down again to 4.6 (but probably with other factors causing a difference to rank 1!). I emphasize that the analysis is totally based on species composition, and pH just shows the effect in terms of measured site conditions.

The algorithm can also be used to select a powerful set of indicator species. In general, species linear correlations are lower than relevé correlations (Section 3.2). As in the 'Schlaenggli' data set, it frequently happens

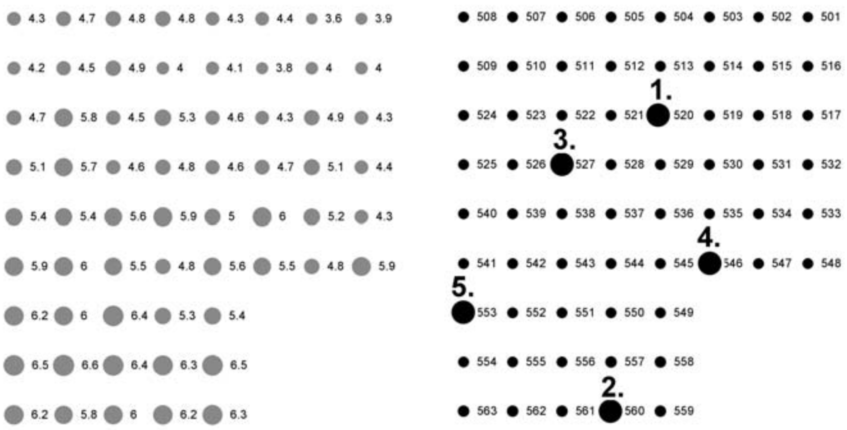


Figure 5.13 Relevés chosen by RANK for permanent investigation. Left: pH measured in peat. Right: plot numbers and first five ranks of plots.

Table 5.4 Ranking species of the ‘Schlaenggli’ data set.

Rank no.		Species	Explained variance (%)	Cumulative variance (%)
1	45	*Carex pulicaris*	15.43	15.43
2	4	*Oxycoccus quadripetalus*	5.08	20.51
3	25	*Drosera rotundifolia*	4.98	25.48
4	90	*Bellidiastrum michelii*	4.25	29.73
5	139	*Orchis latifolia*	2.80	32.53
6	96	*Rhytidiadelphus squarrosus*	2.76	35.29
7	112	*Cirsium oleraceum*	2.64	37.93
8	119	*Rhinanthus minor*	2.56	40.48
9	152	*Ctenidum molluscum*	2.49	42.97
10	95	*Hylocomium splendens*	2.24	47.52
...		...	...	...
20	57	*Sphagnum subsecundum*	1.58	64.65
...		...	...	...
42	145	*Acer pseudoplatanus*	0.78	90.30

that the number of species exceeds the number of relevés. The species ranked first will therefore account for less variance than the relevé ranked first. The result of this analysis is listed in Table 5.4. The 10 first ranks account for almost 50% of total variance; 42 out of 119 are needed for 90%; 100% is reached with 63 species. This is the dimensionality of the similarity space as given by the number of relevés.

6 Classification

6.1 Group structures

'The aim of classification is to obtain groups of objects (samples, species) that are internally homogeneous and distinct from other groups' (Lepš & Šmilauer 2003). Working with a small number of groups rather than a large number of relevés and species is the main practical advantage of classification, and because species combinations tend to re-occur at different locations, classification is justified from the theoretical point of view as well. But a group structure is also a type of pattern and therefore classification can be considered a tool for pattern recognition. As shown in Figure 6.1, however, this is not always the case. The example on the right illustrates a case where two groups are formed within a perfectly continuous point cloud. Additional investigation is needed to distinguish this from the case in the left graph where groups confirm an obvious pattern and therefore are considered *natural*, in contrast to the *artificial* shown on the right. It can be seen from the

Data Analysis in Vegetation Ecology Otto Wildi

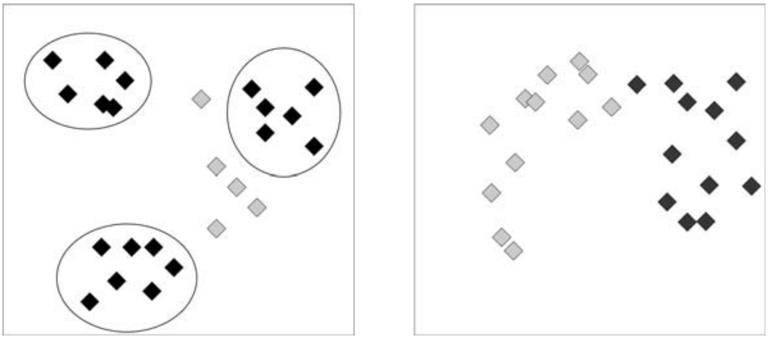


Figure 6.1 Two-dimensional group structures. Left: natural groups (within circles) and intermediate points. Right: continuous structure with artificial division into two groups.

intermediate data points in the left graph that this distinction is not trivial. Just a few data points may connect two natural groups, which then become artificial.

Classification also reduces the dimensionality of data, just as ordination does (Chapter 5). The maximum number of dimensions of an ordination, m, is the smaller of either the number of relevés n or the number of species p; that is, $m = min(n, p)$. When ordinating relevé groups and species groups, the number of dimensions will not exceed the smaller number of either the relevé groups or the species groups. Depending on data, it may be even smaller as in Figure 6.1, where the entire point cloud is presented in two dimensions only. Forming groups when only one or two dimensions exist is a simple task to be carried out visually. Mathematical methods forming groups are needed when the number of dimensions is high, a property typical with vegetation data.

When a species or a relevé is assigned to one group only, then the pattern is discrete. However, classification also involves a continuous concept known as fuzzy classification (Roberts 1986). In fuzzy classification all relevés and species have a degree of belonging to all groups, measured on a continuous scale from 0 to 1, in analogue to ordination where correlations with the axes are distinguished. It happens that classification and ordination converge.

The issue of finding the best – or even just a good – group structure is by no means trivial. First, one has to find an appropriate number of groups, m, in the range:

$$1 \leq m \leq n$$

under the assumption that a group consists of at least one data point (i.e. sampling unit) and n is the sample size. Second, a clustering technique has to be chosen from among the many available software packages to group the sample accordingly.

Why are there so many different methods for clustering and not just one? The reason is that no single method works perfectly, even though the ultimate solution exists: it is the enumeration of all possible assignments of sampling units to groups followed by the selection of the 'best' one. As noted by Anderberg (1973), the number of combinations $\mathcal{S}_n^{(m)}$ is:

$$\mathcal{S}_n^{(m)} = \frac{1}{m!} \sum k = 0^m (-1)^{m-k} \binom{m}{k} k^n \tag{6.1}$$

Even for the very moderate task of assigning 25 sampling units to 5 groups this yields:

$$\mathcal{S}_n^{(m)} = 2\ 436\ 684\ 974\ 110\ 751$$

combinations. This number is so high that the procedure is out of reach of today's computers. The many classification methods invented simplify the task of forming groups; nevertheless they are all restricted by specific strategies imposing internal assumptions. A possible distinction of major strategies is this:

Heuristic	*vs.*	Formal
Agglomerative	*vs.*	Divisive
Hierarchical	*vs.*	Nonhierarchical
Deterministical	*vs.*	Stochastical

A heuristic algorithm implies two elements. First, assumptions are made concerning the initial group structure. Second, an initial configuration is improved through a formal and iterative reallocation procedure. Anderberg (1973) discusses these methods in the context of nonhierarchical clustering: 'Such algorithms begin with an initial point and then generate a sequence of moves from one point to another, each giving an improved value of the objective function, until a local optimum is found' (p. 156). It is worthwhile to note that the optima found depend on the initial assumptions and even on the order in which the data are processed. For very large data sets, processing the sampling units sequentially may be the only feasible method, considering the constraints of computing power (Chapter 11). However, heuristic (and probably all divisive) methods should be avoided in favour of formal agglomerative (and among these, hierarchical) methods.

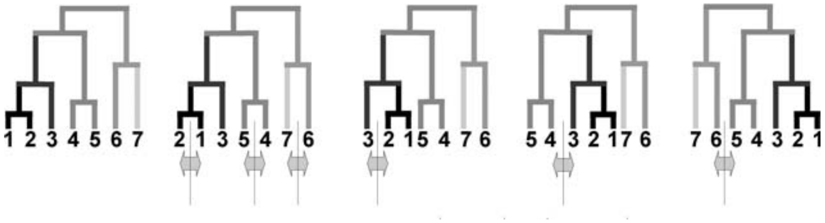


Figure 6.2 A dendrogram as output from an agglomerative hierarchical clustering method. All branches may be in spin and therefore vicinity doesn't signify similarity.

For ecological applications it is often sufficient to distinguish groups without considering hierarchy. However, hierarchy has the advantage that it defines similarity relationships between the groups, which allows us to change the number of resulting groups by altering the hierarchical level. Hierarchy is of course needed when an analysis is based on hierarchy theory (Allen & Starr 1982).

This chapter focuses on agglomerative clustering, with some comments on other approaches. Agglomerative clustering generates a hierarchy, which is usually displayed in the form of dendrograms; that is, resemblance trees depicting the similarity of individuals and groups, as shown in Figure 6.2. On the horizontal axis, sampling units 1 through 7 are lined up, connected by arches. The height of any arch measures the dissimilarity (distance) between the corresponding sampling units – or group of sampling units. However, the order of the sampling units and hence the arches is not given by the algorithms. The configuration is allowed to spin around all vertical axes (Figure 6.2). Thus, the vicinity of data points along the x-axis has no specific meaning and cannot be used for interpretation.

6.2 Linkage clustering

The process of agglomerative hierarchic clustering is demonstrated for three methods: single-, complete- and average-linkage clustering. Whereas single- and complete-linkage clustering are unambiguous terms, average-linkage clustering is used for different methods. The one presented below is also called centroid clustering, while others are listed in Section 6.4.

All three methods use the same definition for the comparison of any two sampling units. The resemblances (distance or similarity) are always taken from the resemblance matrix of the sample, but methods differ in the way they consider larger groups (Figure 6.3, left side). Single-linkage always uses the distance (or similarity) of the closest members of any two groups.

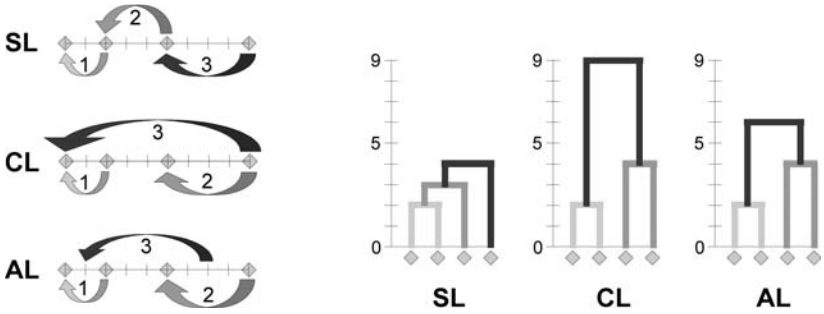


Figure 6.3 Comparing single- (SL), average- (AL) and complete-linkage (CL) clustering. Left: group definitions applied in a one-dimensional example. Right: results displayed as dendrograms.

Complete-linkage refers to the two most distant members of any two groups, thus distance between groups is much larger. Average-linkage measures the similarity between the centroids of the groups; in the one-dimensional example of Figure 6.3 this is the centre of two (or more) sampling units involved. The definition implies that the similarity between groups is generally greater in complete-linkage than in single-linkage, and intermediate in average-linkage.

The example in Figure 6.3 is one-dimensional. The dissimilarity of any two points is therefore just their distance on a straight line. When arranging the sampling units from left to right, the distance matrix is:

$$\mathbf{D} = \begin{array}{cccc} 0 & & & \\ 2 & 0 & & \\ 5 & 3 & 0 & \\ 9 & 7 & 4 & 0 \end{array} \tag{6.2}$$

In the first step (numbered arrows in Figure 6.3) all methods do the same. They find the first two points to be the most similar, separated by two units. This yields a first arch in the dendrogram, which is two units in height. In the second step the next closest neighbours are searched for. In single-linkage this is the third point with a distance of three units from the group formed before. For complete-linkage the third point is five units apart from the new group. The next fusion is therefore formed by points three and four being only four units apart. The corresponding arch has height four. The same holds for average-linkage. However, there are two solutions as the distance between the first two points and the third point is also four units. In the third and final step, single-linkage adds the fourth point to the previous cluster. This is done at height four of the arch, according to the distance of the third and the fourth points. In complete-linkage, the two two-member groups are fused at arch level nine, the distance between the most distant points. The same is done in average-linkage, but the height of the arch reflects the distance between the centres of the involved groups.

Although this is a tiny example, the shape of the resulting dendrograms is typical. In single-linkage, the chaining effect can be seen. The dendrogram formed by complete-linkage is more balanced and typically much higher. The average-linkage dendrogram has intermediate shape.

6.3 Minimum-variance clustering

Single- and complete-linkage clustering consider single data points in their definition of group similarity. *Minimum-variance clustering* (Ward 1963), also called sum-of-squares clustering (see Orlóci (1967)), is based on the relationship of all members of a group (Legendre & Legendre 1998). The objective of minimum-variance clustering is to unify groups such that the increase within group variance is minimized. Since Euclidean space is used, variance can be illustrated, as shown in Figure 6.4. The variance within any group g can be derived from the full data set as follows:

$$Q_g = \sum_{i=1}^{p} \sum_{j=1}^{n_g} (x_{ij} - \overline{x}_i)^2 \tag{6.3}$$

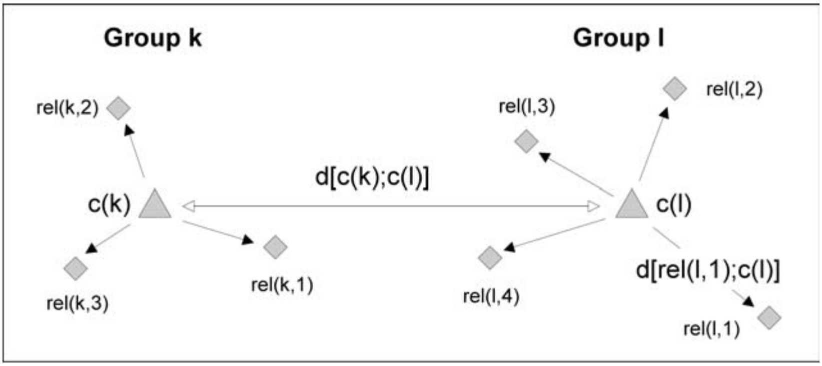


Figure 6.4 Variance within and between groups in minimum-variance clustering. The variance is the sum of the squared distances (arrows). $c(j)$ is the centroid of group j, $rel(j, i)$ relevé score i in group j.

where x_{ij} is the score of species i in relevé j, n_g is the size of group g and p is the number of species. In Orlóci (1978), p. 205, we find that this calculation can be simplified if there is a matrix of squared Euclidean distances, D^2, available:

$$Q_g = \frac{1}{n_g} \sum_{i<j} d_{ij}^2 \tag{6.4}$$

To explain the principle I use the example from Figure 6.3. The distance matrix is displayed in Formula 6.2. When taking the squared distances the matrix becomes:

$$\mathbf{D^2} = \begin{array}{cccc} 0 & & & \\ 4 & 0 & & \\ 25 & 9 & 0 & \\ 81 & 49 & 16 & 0 \end{array} \tag{6.5}$$

It is important to note that the levels at which fusions take place in the dendrogram are the increases in group variance, and not the total group variance. When two groups a and b join, the new variance, $Q(a, b)$, is equal to the total variance of the group minus the variance of the contributors, $Q(a)$ and $Q(b)$:

$$Q(a, b) = Q(a + b) - Q(a) - Q(b) \tag{6.6}$$

In the present example, the first group is built by data points 1 and 2. The variance explained is:

$$Q(1, 2) = \frac{1}{2}4 = 2 \tag{6.7}$$

and for the next fusion it is:

$$Q(3, 4) = \frac{1}{2}16 = 8 \tag{6.8}$$

These two internal variances have to be deducted from the total variance in the final step:

$$Q(5, 6) = \frac{1}{2 + 2}(4 + 25 + 9 + 81 + 49 + 16) - 2 - 8 = 36 \tag{6.9}$$

Clearly, a dendrogram with fusion levels {2; 8; 36} looks different from those seen in Figure 6.3. The differences in appearance get even more pronounced in data sets of larger size.

Minimum-variance clustering is capable of distinguishing groups of different extent and density; that is, groups with high internal variance versus groups with low variance. This is a situation often encountered in vegetation data. As pointed out by Legendre & Legendre (1998), it is often also considered one of many variants of centroid clustering and yet another intermediate solution between the extremes represented by single- and complete-linkage clustering.

6.4 Average-linkage clustering: UPGMA, WPGMA, UPGMC and WPGMC

In many fields of science an intermediate solution to clustering is preferred as standard. The term ‘average-linkage clustering’ has been used for many different methods (e.g. by Sneath & Sokal 1973), such as the abbreviations given in the title of this section. They have in common that the definition of group similarity is based on all members of a group and not just one as in single- and complete-linkage clustering. It is in this regard that they are related to minimum-variance clustering (Section 6.3).

The four methods are discussed in some detail in Legendre & Legendre (1998), where a small numerical example is given for illustration and comparison. I will continue using the abbreviations and refer to Table 6.1 for the full names. The methods are distinguished by two alternative criteria (Table 6.1). UPGMA and WPGMA use the average resemblance of all group members as a criterion for between-group resemblance. In UPGMC and WPGMC, a group centroid is established: in geometrical terms this is the centre of gravity of any one group. Between-group resemblance is thus the distance or similarity of any two centroids. As with many groups of methods, the results depend on the data analysed and the method used, and they may be identical or differ considerably.

The second criterion of distinction concerns weighting. ‘U’ (UPGMA, UPGMC) signifies ‘unweighted’, which, however, may be somewhat misleading. When computing average resemblance as well as centroids, group sizes are taken into account. ‘Unweighted’ means that the weight of the original set of resemblances is retained. ‘W’ (WPGMA, WPGMC) signifies ‘weighted’. This means that groups of different size get the same weight when fused. In this case, the weight is the inverse of group size.

Table 6.1 Properties of four popular clustering methods (adapted from Legendre & Legendre (1998)).

Properties	Consider the average similarities or distances of all members of a cluster as candidates for further fusions	Consider the centroid of all members of a cluster as candidates for further fusions
Give equal weight to the original resemblances (weight of groups proportional to group size)	**UPGMA** (unweighted arithmetic average clustering)	**UPGMC** (unweighted centroid clustering)
Give equal weight to any two branches of the dendrogram (weight of groups identical irrespective of size)	**WPGMA** (weighted arithmetic average clustering)	**WPGMC** (weighted centroid clustering)

The relationship between the four methods is shown in Table 6.1, an extension of Table 8.2 in Legendre & Legendre (1998). Even though these methods seem to be popular (partly due to the appealing abbreviations), it must be noted that they all represent special cases rather than a justified standard. Furthermore, especially when using centroid clustering, reversals may occur in the dendrograms: subsequent fusions may take place at lower levels than the previous. Dendrograms of this kind are both difficult to draw and difficult to interpret.

6.5 Forming groups

A dendrogram offers unique flexibility in reducing the number of members considered to constitute a population. When cutting horizontally, it will divide the sample into groups. The dissimilarity of the resulting groups will be as large as the level at which cutting takes place, as is schematically shown in Figure 6.5. Cutting the dendrogram just below the uppermost arc will divide the sample into two groups. When moving down to the next arc, one new group is formed (unless two or more arcs are found to be at exactly the same level). The procedure ends when all groups consist of one sampling unit only.

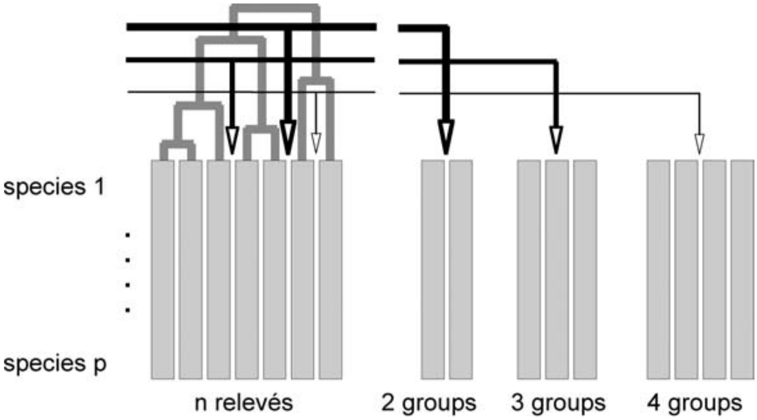


Figure 6.5 Cutting dendrograms at different levels of dissimilarity.

But how many groups should be formed, and how large and how homogeneous should they be? Considering the huge number of potential solutions to this problem (Formula 6.1), I suggest that in most cases the decision has to be based on the purpose of the classification. This means that a criterion has to be defined which is independent of the data yielding the classification. As an example, a site factor can be used for testing the explanatory power of classified vegetation samples. Only in rare cases will the structure found in clustering yield a straightforward solution, as illustrated in Figure 6.1, left side.

When group number and size are chosen there is frequently an opportunity to test these for significance. In Section 7.2 the use of analysis of variance for measuring the predictive power of a classification will be raised. The test criterion, the F-value, helps us to find guidelines for group number and size. It is defined as:

$$F = \frac{Var_{between-groups}}{Var_{within-groups}} \tag{6.10}$$

The significance of any one F-value can be checked in the F-table of a statistical textbook. The F-value has two degrees of freedom, df1 and df2, where:

$$\begin{aligned} df1 &= m - 1 \quad \text{and} \\ df2 &= n - m \end{aligned} \tag{6.11}$$

in which m is the number of groups and n is the sample size. Upon inspection of F-tables it becomes clear that n must be sufficiently large: much larger than m, the number of groups. However, if the number of groups is too low, the predicting power of the classification is also poor. Two groups (df1 = 1) are unfavourable in almost all cases, a number around five (df1 = 4) promises better results. Furthermore, from the point of view of the analysis of variance, the number of sampling units per group should also not drop below around five.

6.6 Structured synoptic tables

6.6.1 The aim of ordering tables

In early times of plant ecology the predominant method of data analysis consisted of the rearrangement of synoptic tables: rows and columns were shifted to achieve an order for relevés and species reflecting the similarity pattern of the sample. An early description of the method was published by Ellenberg (1956) (see also Mueller-Dombois & Ellenberg 1974). It basically implemented some rules for simultaneous ordering of rows and columns. When multivariate clustering became operational for large data sets, several approaches were developed for substituting the manual process. Examples are the computer programs TABORD (Maarel *et al.* 1978) and TWINSPAN (Hill 1979b). While TABORD is heuristic and finds a solution through iteration, TWINSPAN includes in its first steps divisive clustering applied to the result of correspondence analysis. Legendre & Legendre (1998) mention that this regularly leads to misclassifications as the similarity space is not considered in all dimensions simultaneously.

A second series of approaches appeared some 10 years later. They include the method described below (Wildi 1989), a strategy by Podani & Feoli (1991), the program ESPRESSO by Bruelheide & Flintrop (1994) and probably others. The method shown below is entirely based on multivariate analysis. Podani & Feoli (1991) extend this by implementing additional tests of success and reallocations. ESPRESSO, on the other hand, is again an entirely heuristic, iterative procedure.

As will be shown, ordered vegetation tables are more than just clustered two-dimensional arrangements. They usually exhibit a main gradient found in the sample, representing a large-scale pattern, and groups (or subgroups) at smaller scale. High presence scores form the diagonal of the table, so that for each relevé group the characteristic species occurrence can immediately be found. And finally, species with lower predictive power are moved down to the bottom part of the list. This explains why more than one method is needed to achieve such a result. The method I describe also serves as an example

Table 6.2 Steps involved in sorting synoptic tables by multivariate methods according to Wildi (1989).

Step	Transformation	Method	Effect
1 Clustering relevés	$x' = x^{0.2}$	Sum-of-squares clustering	Forming relevé groups
2 Ordinating relevés and species	$x' = x^{0.2}$	Correspondence analysis	Determining major trend for order within groups
3 Clustering species	$x' = x^{0.1}$	Complete linkage based on Euclidean distance	Forming species groups
4 Analysis of concentration	presence–absence	Analysis of concentration	Determining major trend among relevé/species groups
5 Species ranking	$x' = x^{0.1}$	Analysis of variance (F-values)	Separating species with high vs. low resolving power
6 Printing	none	Apply ordering criteria from steps 1–5	Sorting, as suggested by Ellenberg (1956)

of the application of different multivariate methods in combination, rather than just representing the best solution for generating ‘nice’ tables. Ordering tables is not the ultimate tool for analysing and interpreting environmental data and some more limitations are discussed in Section 11.5.

6.6.2 Steps involved

The steps described below follow the suggestions published in Wildi (1989), with some minor changes. They are summarized in Table 6.2.

Step 1. The procedure starts with clustering of relevés. If the data are percentages, like the data set of Ellenberg (1956) shown below, then the scores should be transformed to closer reflect presence–absence. I suggest the use of $x' = x^{0.2}$ or so. The relevé vectors are then normalized to compensate for differences in species richness. Using minimum-variance clustering (based on the centred scalar product) allows for groups greatly differing in size.

Step 2. Before proceeding to the analysis of species, the relevés are subjected to ordination to find the dominating gradient in the sample. This will be used later to rearrange relevés and species within the groups. While the scalar transformation remains the same, correspondence analysis is often a good choice for identifying the dominating trend of relevés. Simultaneously, it yields the same for the species.

Step 3. For clustering species, the methods used differ (see Figure 3.4), due to the fact that correlation of species is generally rather low and nonlinear. To enhance joint occurrence, the scores are transformed close to presence–absence, using $x' = x^{0.1}$. I suggest normalizing (but not standardizing!) the species vectors. Since groups consisting of frequent but also rare species are generally not welcome, a resemblance measure has to be used that does not centre the vectors. In Table 6.2 Euclidean distance is proposed. With complete-linkage clustering, more evenly sized groups can be expected.

Step 4. Because the relevé and species groups are generated form dendrograms, their order is arbitrary (see Figure 6.2). Blocks of high-score density will be dispersed all over the table. This is shown in Figure 6.6 (a). Analysis of concentration (Section 7.4) now counts the frequency of non-zero scores in each block, thereby forming a contingency table. This table, after adjustments, is analysed as in correspondence analysis. According to the first ordination axis, the procedure yields a gradiental order for both relevé groups and species groups.

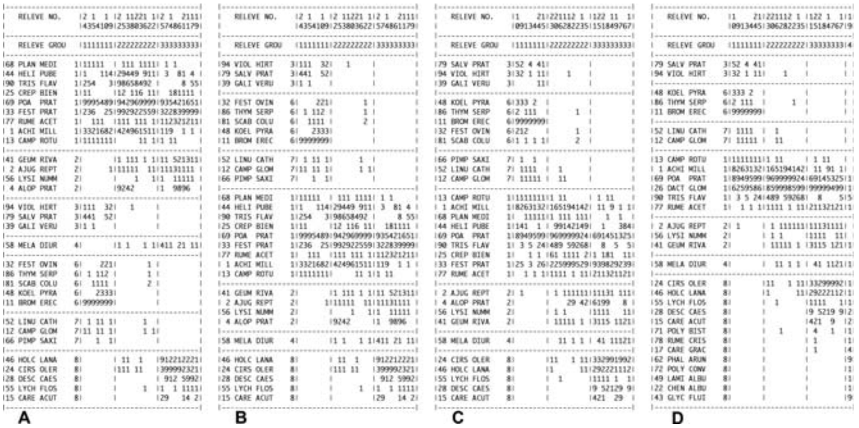
Step 5. The species are ranked by Jancey's ranking (Jancey 1979). For formal reasons transformation should be the same as in the clustering of species. The species with the highest F-values are then taken for the upper, discriminating part of the table. It is a matter of taste how many are taken. Significance levels are not valid in the statistical sense. In any case, species with low F-values (approaching 1.0) should of course be suppressed and moved to the bottom part of the table.

Step 6. When all the row and column vectors are arranged as seen in the previous steps, the table can be printed. Since the sign of ordination axes is arbitrary, it can happen that the diagonal of high scores points from the upper right to the lower left. This can be changed by reversing the order of the relevés.

6.6.3 Example: Ordering Ellenberg's data

For the purpose of illustration the data set of Ellenberg (1956) (also published in Mueller-Dombois & Ellenberg 1974) is used in Figure 6.6 to demonstrate the effect of different methods. The number of relevé groups chosen is three, the same as chosen by Ellenberg. In many cases, taking the square root of the number of relevés turned out to yield pleasing results (i.e. about 15 for a table of size 200, 30 for a table of size 1000, etc.). The number of species groups in this example is eight. Because the correlation among species is generally low, the groups should be small in size (i.e. three to six species). Under many circumstances, dividing the number of species by about four may be a good choice for the number of species groups. The number of species displayed is 30 out of 94.

In Figure 6.6 (a) the result of clustering is shown. As can be expected, the blocks of high-frequency non-zero scores are dispersed. Figure 6.6



```
|---------------------------------------------|
|   RELEVE NO.     |2 1  1 |2 11221 1|2 1  2111|
|                  |4354109|253803622|574861179|
|---------------------------------------------|
|   RELEVE GROU    |1111111|222222222|333333333|
|---------------------------------------------|
|68 PLAN MEDI    1|11111  | 111 1111| 1 1     |
|44 HELI PUBE    1|1   114|29449 911| 3  81 4 |
|90 TRIS FLAV    1|254   3|98658492 |     8 55|
|25 CREP BIEN    1|11     |12 116 11|  181111 |
|69 POA  PRAT    1|9995489|942969999|935421651|
|33 FEST PRAT    1|236  25|992922559|322839999|
|77 RUME ACET    1|  111  |111 111 1|112321211|
| 1 ACHI MILL    1|3321682|424961511|119  1 1 |
|13 CAMP ROTU    1|1111111|     11 1|1 11     |
|---------------------------------------------|
|41 GEUM RIVA    2|       |1 111 1 1|11 521311|
| 2 AJUG REPT    2|      1|11111  11|11131111 |
|56 LYSI NUMM    2|       |    1   1|1  11111 |
| 4 ALOP PRAT    2|       |9242     |1  9896  |
|---------------------------------------------|
|94 VIOL HIRT    3|111  32|   1     |         |
|79 SALV PRAT    3|441  52|         |         |
|39 GALI VERU    3|1 1    |         |         |
|---------------------------------------------|
|58 MELA DIUR    4|       |1 1   1 1|411 21 11|
|---------------------------------------------|
|32 FEST OVIN    6|    221|       1 |         |
|86 THYM SERP    6| 1 112 |       1 |         |
|81 SCAB COLU    6|  1111 |       2 |         |
|48 KOEL PYRA    6|   2333|         |         |
|11 BROM EREC    6|9999999|         |         |
|---------------------------------------------|
|52 LINU CATH    7| 1 11 1|      1  |         |
|12 CAMP GLOM    7|11 11 1|      1 1|         |
|66 PIMP SAXI    7|   1  1|         |         |
|---------------------------------------------|
|46 HOLC LANA    8|       | 11  1   |912212221|
|24 CIRS OLER    8|       |111 11   |399992321|
|28 DESC CAES    8|       |         | 912 5992|
|55 LYCH FLOS    8|       |      1  |1  1 1111|
|15 CARE ACUT    8|       |         |29   14 2|
|---------------------------------------------|
```

A

```
|---------------------------------------------|
|   RELEVE NO.     |2 1  1 |2 11221 1|2 1  2111|
|                  |4354109|253803622|574861179|
|---------------------------------------------|
|   RELEVE GROU    |1111111|222222222|333333333|
|---------------------------------------------|
|94 VIOL HIRT    3|111  32|   1     |         |
|79 SALV PRAT    3|441  52|         |         |
|39 GALI VERU    3|1 1    |         |         |
|---------------------------------------------|
|32 FEST OVIN    6|    221|       1 |         |
|86 THYM SERP    6| 1 112 |       1 |         |
|81 SCAB COLU    6|  1111 |       2 |         |
|48 KOEL PYRA    6|   2333|         |         |
|11 BROM EREC    6|9999999|         |         |
|---------------------------------------------|
|52 LINU CATH    7| 1 11 1|      1  |         |
|12 CAMP GLOM    7|11 11 1|      1 1|         |
|66 PIMP SAXI    7|   1  1|         |         |
|---------------------------------------------|
|68 PLAN MEDI    1|11111  | 111 1111| 1 1     |
|44 HELI PUBE    1|1   114|29449 911| 3  81 4 |
|90 TRIS FLAV    1|254   3|98658492 |     8 55|
|25 CREP BIEN    1|11     |12 116 11|  181111 |
|69 POA  PRAT    1|9995489|942969999|935421651|
|33 FEST PRAT    1|236  25|992922559|322839999|
|77 RUME ACET    1|  111  |111 111 1|112321211|
| 1 ACHI MILL    1|3321682|424961511|119  1 1 |
|13 CAMP ROTU    1|1111111|     11 1|1 11     |
|---------------------------------------------|
|41 GEUM RIVA    2|       |1 111 1 1|11 521311|
| 2 AJUG REPT    2|      1|11111  11|11131111 |
|56 LYSI NUMM    2|       |    1   1|1  11111 |
| 4 ALOP PRAT    2|       |9242     |1  9896  |
|---------------------------------------------|
|58 MELA DIUR    4|       |1 1   1 1|411 21 11|
|---------------------------------------------|
|46 HOLC LANA    8|       | 11  1   |912212221|
|24 CIRS OLER    8|       |111 11   |399992321|
|28 DESC CAES    8|       |         | 912 5992|
|55 LYCH FLOS    8|       |      1  |1  1 1111|
|15 CARE ACUT    8|       |         |29   14 2|
|---------------------------------------------|
```

B

```
|---------------------------------------------|
|   RELEVE NO.     |1    21|221112 1 |122 11  1|
|                  |0913445|306282235|151849767|
|---------------------------------------------|
|   RELEVE GROU    |1111111|222222222|333333333|
|---------------------------------------------|
|79 SALV PRAT    3|52 4 41|         |         |
|94 VIOL HIRT    3|32 1 11|    1    |         |
|39 GALI VERU    3|     11|         |         |
|---------------------------------------------|
|48 KOEL PYRA    6|333 2  |         |         |
|86 THYM SERP    6|2 111  |      1  |         |
|11 BROM EREC    6|9999999|         |         |
|32 FEST OVIN    6|212    |      1  |         |
|81 SCAB COLU    6|1 1 1 1|      2  |         |
|---------------------------------------------|
|66 PIMP SAXI    7| 1  1  |         |         |
|52 LINU CATH    7| 1111  | 1       |         |
|12 CAMP GLOM    7| 11111 | 11      |         |
|---------------------------------------------|
|13 CAMP ROTU    1|1111111|1 11     | 1 11    |
| 1 ACHI MILL    1|8263132|165194142| 11 9 1 1|
|68 PLAN MEDI    1|  11111|1 111 111|   1  1  |
|44 HELI PUBE    1|141  1 | 99142149| 1    384|
|69 POA  PRAT    1|8949599|969999924|691451325|
|90 TRIS FLAV    1| 3 5 24|489 59268|  8  5  5|
|25 CREP BIEN    1|   1 1 |61 1111 2|1 181  11|
|33 FEST PRAT    1|25 3 26|225999529|939829239|
|77 RUME ACET    1|  1 1 1|1111 1 11|211321121|
|---------------------------------------------|
| 2 AJUG REPT    2| 1     | 1 111111|11131 111|
| 4 ALOP PRAT    2|       |    29 42|6199   8 |
|56 LYSI NUMM    2|       | 1 1     |1111   11|
|41 GEUM RIVA    2|       | 11111 1 |3115 1121|
|---------------------------------------------|
|58 MELA DIUR    4|       |  11 1 1 | 41 11121|
|---------------------------------------------|
|24 CIRS OLER    8|       |11   1 11|332991992|
|46 HOLC LANA    8|       |1      11|292221112|
|55 LYCH FLOS    8|       | 1       |1111 1  1|
|28 DESC CAES    8|       |         |9 52129 9|
|15 CARE ACUT    8|       |         |421   29 |
|---------------------------------------------|
```

C

```
|-----------------------------------------------
|   RELEVE NO.     |1    21|221112 1 |122 1  1|1
|                  |0913445|306282235|15184767|9
|-----------------------------------------------
|   RELEVE GROU    |1111111|222222222|33333333|4
|-----------------------------------------------
|79 SALV PRAT    3|52 4 41|         |        | 
|94 VIOL HIRT    3|32 1 11|    1    |        | 
|-----------------------------------------------
|48 KOEL PYRA    6|333 2  |         |        | 
|86 THYM SERP    6|2 111  |      1  |        | 
|11 BROM EREC    6|9999999|         |        | 
|-----------------------------------------------
|52 LINU CATH    7| 1111  | 1       |        | 
|12 CAMP GLOM    7| 11111 | 11      |        | 
|-----------------------------------------------
|13 CAMP ROTU    1|1111111|1 11     | 1 11   | 
| 1 ACHI MILL    1|8263132|165194142| 11 91 1| 
|69 POA  PRAT    1|8949599|969999924|69145325|1
|26 DACT GLOM    1|6259586|859998599|99999499|1
|90 TRIS FLAV    1| 3 5 24|489 59268|  8    5|5
|77 RUME ACET    1|  1 1 1|1111 1 11|21132121|1
|-----------------------------------------------
| 2 AJUG REPT    2| 1     | 1 111111|11131111| 
|56 LYSI NUMM    2|       | 1 1     |1111  11| 
|41 GEUM RIVA    2|       | 11111 1 |3115 121|1
|-----------------------------------------------
|58 MELA DIUR    4|       |  11 1 1 | 41 1121|1
|-----------------------------------------------
|24 CIRS OLER    8|       |11   1 11|33299992|1
|46 HOLC LANA    8|       |1      11|29222112|1
|55 LYCH FLOS    8|       | 1       |1111   1|1
|28 DESC CAES    8|       |         |9 5219 9|2
|15 CARE ACUT    8|       |         |421  9  |2
|71 POLY BIST    8|       |  1      | 4   1  |1
|78 RUME CRIS    8|       |         | 1      |1
|17 CARE GRAC    8|       |         | 1      |4
|62 PHAL ARUN    8|       |         |        |9
|72 POLY CONV    8|       |         |        |1
|49 LAMI ALBU    8|       |         |        |1
|22 CHEN ALBU    8|       |         |        |1
|43 GLYC FLUI    8|       |         |        |9
|-----------------------------------------------
```

D

Figure 6.6 Structuring the meadow data set of Ellenberg (Mueller-Dombois & Ellenberg 1974). (a) Ordering based on cluster analysis only. (b) Blocks rearranged by analysis of concentration. (c) Within-group order changed according to correspondence analysis (complete ordering). (d) Ordering with four instead of three relevés groups.

(b) exhibits the same classification, but the blocks are rearranged by analysis of concentration. In (c), within-group order is changed according to correspondence analysis. In some cases, small within-group gradients can be identified. This slightly improves the appearance of the diagonal structure of the table, but does not improve on the result of classification. In (d), the number of relevé groups is increased to four. This example demonstrates that the relevé groups determine the list of differentiating species. The selection of differentiating species has changed completely. The last group consists of one relevé only, and species occurring only there achieve high F-values. It has to be noted that such a solution with a single relevé in a group is not practical as the variance within the group is not defined. Clearly, relevé number 19 is an outlier (see Section 11.3) and should be removed from the set prior to analysis (Wildi 1989).

7
Joining ecological patterns

7.1 Pattern and ecological response

The identification of patterns is a first step in finding rules governing systems. Patterns can be found in biotic space, but also in environmental, spatial and temporal (Section 2.3.2). As a next step in the investigation of ecosystems, spaces are compared in search of common patterns. Clarke (1993), introducing this strategy to analysis of benthic communities, puts it as follows: 'Having allowed the community data to "tell its own story", its relationship to matching environmental data is examined by superimposing the values of each abiotic variable separately onto the biotic ordination.' Hence, when talking of 'joining' in this chapter I address various methods of comparison, such as superimposing, correlating, variance partitioning and so on. The simplest case of comparison is the univariate, where a vector expressing performance of a species is correlated with an environmental

Data Analysis in Vegetation Ecology Otto Wildi

factor such as precipitation. However, even if a strong correlation is found, this will not allow straightforward conclusions, because plant species interact and the change in performance may not be directly related to an environmental factor, but caused by competition or facilitation from another species. The same holds among environmental factors: water uptake of plants, for example, depends on temperature, and an ecologically sound interpretation requires a simultaneous view of these two factors. Hence, rather than just pairwise correlating variables, the focus of this chapter is on the analysis of relationships between *spaces* – the biological, the ecological, the physical and the temporal (according to Section 2.3.2). There are methods relating single variables of one space to another space, and others correlating two or more entire spaces. An example for the comparison of two multivariate spaces is the analysis of contingency, where data from different spaces are classified and the resulting frequencies are compared in contingency tables. These are further analysed, agreement measured, tested and interpreted. And then there is constrained ordination, where variance is partitioned and that shared by two spaces is used for joint ordination, of vegetation and environmental data, for example. The spaces are usually vegetation and site data. Examples are shown in Section 7.5.

What does it mean if one finds common properties in patterns? It is a good reason to hypothesize that there exists a response in either direction; that is, an interaction. Typically, vegetation responds to environmental conditions, which play the role of *independent* – and vegetation the *dependent* – variables. In the ecological reality, the opposite can happen too. In a ruderal environment species composition can be a result of the seed bank and not so much of present site conditions. As time progresses, it is expected that a correlation between species composition and site factors will start to emerge, not only developing dependence, but also generating a temporal pattern. Although time processes are treated separately in Chapter 9, much of what is presented here applies to these as well.

Often occurring in spatio-temporal systems is *autocorrelation*. This is the phenomenon that any two observational vectors are more similar than expected if they occur in close neighbourhood, be it spatial or temporal. Thus, autocorrelation is the phenomenon of dependence across small distances. In vegetation data spatial autocorrelation is often a result of plant dispersal and for this reason it is of great ecological significance. As will be seen in the sequel, there are methods to identify true autocorrelation, whilst most other methods of correlation are hampered by this.

7.2 Analysis of variance

In this section I present the basic idea of analysis of variance as a tool for measuring the strength of group pattern. In the statistical context, this is hypothesis testing (by statistical inference). Alternatively, it can serve as a purely descriptive tool for quantitatively measuring distinctness of patterns. In this case, the result cannot be considered for statistical testing, because it is derived from the same body of data used for forming groups. But it is a means of ordering variables according to predictive power: a kind of ranking which in turn serves dimensionality reduction when omitting unimportant variables.

7.2.1 Variance testing

Forming groups in classification is not trivial (Section 6.5). In order to test a classification some external, independent criteria are needed. Since we are interested in the interactions between vegetation and site, we will first seek a group structure in vegetation and ask if this also exists in one or several site factors.

Analysis of variance is used to test this structure for significance. The underlying idea in this is in partitioning total variance:

$$V_T = V_W + V_B \tag{7.1}$$

The principle of total (V_T), between-groups (V_B) and within-groups (V_W) deviation from means is illustrated in Figure 7.1. This one-dimensional case illustrates that the proportion of the variance (the squared distances) between the groups and the same within the groups measures the crispness of the classification. This is what the F-value measures (see Section 6.5):

$$F = \frac{V_B}{V_W} \tag{7.2}$$

Provided the classification is given, this can directly be calculated from the raw data, X. The following notation is used in the equations below:

n	sample size, total	
m	no. of groups	
p	no. of variables	
i	current sampling unit	$i = 1, \ldots, n$

j	current variable	$j = 1, \ldots, p$
k	current group	$k = 1, \ldots, s$
l	current sampling unit in group k	
s	current group size of group k	

The application shown serves the comparison of the p variables for their resolving power in explaining vegetation groups. The total variance of current variable j is obtained by:

$$V_{T,j} = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2 \tag{7.3}$$

Variance within groups does literally the same, but it refers to the group means (the centroids) instead of the total. For current sampling unit l in group k this is:

$$V_{W,j} = \sum_{k=1}^{m} \sum_{l=1}^{s} (x_{lj} - \overline{x}_{kj})^2 \tag{7.4}$$

For variance between groups, only the group means are compared to the total:

$$V_{B,j} = \sum_{k=1}^{m} s_k (\overline{x}_{kj} - \overline{x}_j)^2 \tag{7.5}$$

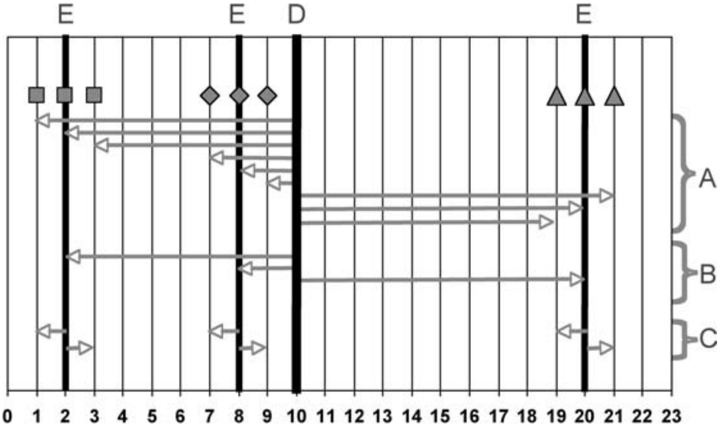


Figure 7.1 Distinctness of group structure. Symbols are sampling units belonging to three groups. (a) Total deviation from sample mean ($n = 9$). (b) Total deviation of group means from sample mean ($m = 3$). (c) Within-group deviations from group mean. (d) Sample mean. (e) Group means.

In order to consider all sampling units, the deviations of group means from the total mean have to be multiplied by group size, s_k. For the purpose of testing, the degrees of freedom have to be determined. As shown in Section 6.5, these are:

$$df1 = m - 1 \qquad \text{and} \qquad df2 = n - m \tag{7.6}$$

To demonstrate the application of this test, an example is shown in Table 7.1.

These are the ‘nzzm5’ data, an artificial data set of 11 relevés, 21 species and 3 site factors (pH, slope, altitude) describing forest stands on the Swiss Plateau (Appendix B). The data table is classified using the method explained in Section 6.6. The number of relevé groups chosen is three; the number of species groups (not relevant in this application) is four. For the three site factors we obtain:

pH	F=	23.6980	**
Slope	F=	3.7161	*
Altitude (m)	F=	0.2234	

Using Formula 7.6 we get $Df1 = 2$ and $Df2 = 8$. Inspecting the F-distribution table in any standard statistical textbook yields the significance levels above. The pH value qualifies as the best predictor for species composition as it reflects the group structure with a probability error of $p = 0.01$. For slope, significance is given only at $p = 0.05$, while altitude does not mirror this pattern.

7.2.2 Variance ranking

The idea of ranking species based on a variance criterion has been proposed by Jancey (1979). Technically it is identical to the F-testing shown in the previous section. The variables tested, the species, are the same as used for the classification of the relevés. Therefore, the F-value cannot serve as a test criterion: it merely represents a measure for the relative resolving power of the species in distinguishing the relevé groups. Jancey (1979) suggests sorting the species list according to the F-values in decreasing order. Such a list is presented in Table 7.2. Species with sufficiently high F-values can be used to build keys for vegetation mapping in the field.

Jancey (1979) also mentions that F-values can be computed for a restricted set of groups. In this case, the analysis yields the resolving power of species based on the selected relevé groups only. In fact, any new classification of

Table 7.1 The structured data set ‘nzzm5’. Based on this classification, pH, slope and altitude are being subjected to analysis of variance (see main text).

RELEVE NO.		49	25	2	6	39	18	50	9	4	10	27
GROUP NO.		1	1	1	1	3	3	3	3	2	2	2
1 pH		4.8	5.0	4.4	4.8	5.2	6.0	5.8	5.6	6.2	6.5	6.0
2 Altitude		550	480	450	420	500	400	520	580	500	560	450
3 Slope, deg.		15.0	18.0	10.0	4.5	12.5	2.5	3.0	6.0	0.5	4.0	2.5
8 *Vaccinium myrtillus*	3	+		2	1							
6 *Sambucus racemosa*	3	+		2	1							
21 *Polytrichum formosum*	3	3	+	+	1							
13 *Veronica officinalis*	3	+	+	1	1							
12 *Luzula nemorosa*	3	2	1	2	+							
2 *Quercus petraea*	4	4	3	1	2	+	+		1			+
5 *Lonicera xylosteum*	4		+	+		+				1	+	1
3 *Acer pseudoplatanus*	1					+	+	1	1	1	2	4
4 *Fraxinus excelsior*	1						+	1	+	3	3	2
7 *Sambucus nigra*	2						+			+	2	1
18 *Arum maculatum*	2									1		+
19 *Ranunculus ficaria*	2									1	2	4
16 *Primula elatior*	2									+	2	2
17 *Allium ursinum*	2									4	2	+
1 *Fagus silvatica*	999	2	4	3	4	5	5	+	4	+	2	1
14 *Galium odoratum*	999	+	+	+		+	2		1	+	+	1
9 *Carex silvatica*	999	+		+	1	+	2		+	+		2
10 *Oxalis acetosella*	999		+	+	+	3	1		1	1	2	
11 *Viola silvestris*	999	+		+	+	+		+	1	+		+
20 *Eurhynchium striatum*	999	+	+	1		+	+		1	+		+
15 *Lamium galeobdolon*	999			+	+	+			+	2	1	2

relevés will alter the F-values and accordingly the ranked list. An application of this is demonstrated below.

7.2.3 How to weight cover abundance (example)

Statistical tests can serve as a means of assessing uncertainty in results. Choosing the ‘best’ methods and options for clustering is an example of

Table 7.2 Variance ranking of species based on the classification of relevés shown in Table 7.1.

Rank no.		Species	F-value
1	13	*Veronica officinalis*	2120.9
2	19	*Ranunculus ficaria*	2060.1
3	12	*Luzula nemorosa*	1307.8
4	16	*Primula elatior*	1127.0
5	3	*Acer pseudoplatanus*	1122.5
6	21	*Polytrichum formosum*	750.16
7	17	*Allium ursinum*	664.65
8	4	*Fraxinus excelsior*	12.243
9	7	*Sambucus nigra*	10.798
10	8	*Vaccinium myrtillus*	7.5749
11	6	*Sambucus racemosa*	7.5749
12	18	*Arum maculatum*	5.7973
13	2	*Quercus petraea*	2.9624
14	5	*Lonicera xylosteum*	2.5385
15	1	*Fagus silvatica*	1.7683
16	15	*Lamium galeobdolon*	1.5769
17	14	*Galium odoratum*	0.3515
18	20	*Eurhynchium striatum*	0.0398
19	11	*Viola silvestris*	0.0348
20	10	*Oxalis acetosella*	0.0290
21	9	*Carex silvatica*	0.0175

this as the performance of classifications can only be measured a posteriori based on independent environmental factors. In this example, seven alternative classifications are evaluated by analysis of variance computed for eight different environmental factors. To generate the seven classifications I

Table 7.3 Transformations used in the variance-testing example, Figure 7.2.

Transformation	Scores	Property
none (code)	{0,r,+,1,2,3,4,5}	nonimal type
ranked	{0,1,2,3,4,5,6,7}	rank scale
$x' = x^{0.0625}$	{0,1,1.04,1.07,1.09,1.11,1.12,1.13}	≈1/0
$x' = x^{0.125}$	{0,1,1.09,1.15,1.19,1.22,1.25,1.28}	close to 1/0
$x' = x^{0.25}$	{0,1,1.19,1.32,1.41,1.50,1.57,1.63}	very low weight to cover
$x' = x^{0.5}$	{0,1,1.41,1.73,2.00,2.24,2.45,2.65}	low weight to cover
$x' = x^{1}$	{0,1,2.00,3.00,4.00,5.00,6.00,7.00}	rank scale
$x' = x^{2}$	{0,1,4,9,16,25,36,49}	close to cover %
$x' = x^{4}$	{0,1,16,81,256,625,1296,2401}	low scores suppressed

used the same clustering method and even derived the same number of groups, resulting in an identical number of degrees of freedom. I only altered transformation of cover-abundance data prior to clustering, as previously shown in Table 3.3. In this way the exercise becomes an evaluation of seven different data transformations fitting eight environmental factors via classifications.

In summary, the following standards are imposed on the clustering of the test data set ‘Schlaenggli’ (Appendix B):

1 Cover-abundance data are transformed according to $x' = x^y$ (see Equation 3.1) in seven steps ranging from $y = 0.0625$ to $y = 4$.

2 The resemblance measure used to compare relevés is the correlation coefficient.

3 Minimum-variance clustering is used and there are always six groups formed.

From the 6 groups and 63 relevés involved we get $df1 = 5$ and $df2 = 57$ degrees of freedom for any of the classifications obtained (see Equation 7.6). The transformation applies to the ranks of the cover-abundance code used, which translates the code {0,r,+,1,2,3,4,5} to the raw scores {0,1,2,3,4,5,6,7}. The transformations and resulting scores are listed in Table 7.3.

A selected subset of the results is shown in Figure 7.2, depicting the change of the F-values as transformation of species scores prior to clustering

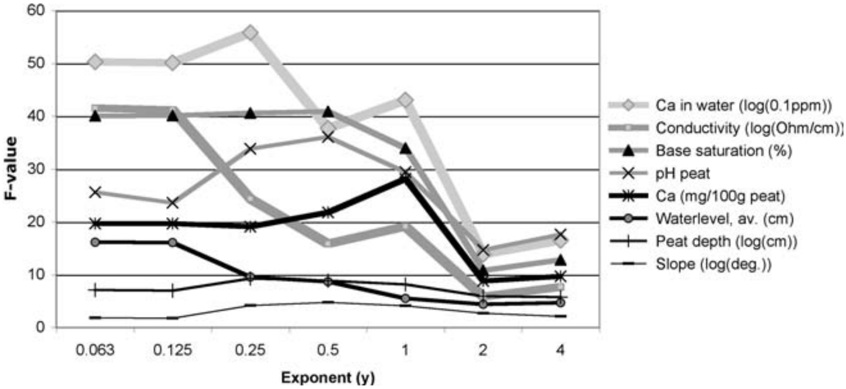


Figure 7.2 F-values of selected site factors. The x-axis values are exponents used for transformations leading to seven different classifications. The y-axis contains the F-values.

occurs from almost qualitative (small exponents used) towards an extremely quantitative view (large exponents used). In any table of F-values we find the following significance thresholds for $df1 = 5$ and $df2 = 60$ degrees of freedom:

$$\begin{array}{ll} p = 0.001 & F = 4.71 \\ p = 0.01 & F = 3.43 \\ p = 0.05 & F = 2.37 \end{array}$$

Almost all of the values in Figure 7.2 are highly significant, a result of the strong gradient pattern inherent in the system. The best coincidence between vegetation and site factors is often achieved when species scores are close to presence–absence (1/0); that is, when y is small. This emphasizes species occurrence – the qualitative view – whereas the quantitative view dominates when y is high; that is, when $y = 2$ or $y = 4$ is chosen. There, small scores remain uninfluential and only dominating species are taken into account. As can be seen in Figure 7.2, this generally leads to low F-values.

The result is somewhat surprising. There is not a ‘best’ solution to classification (and, hence, to transformation) as the environmental variables behave differently. Whereas Ca concentration in peat is best characterized when using the untransformed rank scale ($y = 1.0$), pH in peat reaches maximum when the ranks are square-root transformed ($y = 0.5$), while conductivity (log Ohm/cm water) relates best to presence–absence.

A generally ‘good’ similarity space in terms of the environmental factors is obtained around $y \approx 0.5$ (low weight of cover). A rather unsuitable transformation for this purpose is around $y \approx 2.0$, which should probably be avoided.

7.3 Correlating resemblance matrices

7.3.1 The Mantel test

We have seen so far how vectors are related to multivariate patterns. While this is often helpful, the more general problem is to compare two (or more) multivariate patterns. One solution to this is canonical correlation, a method that relates n variables of one data set to m variables of another. I abstain from this, because in ecology one usually runs into the problem of excessive degrees of freedom: altogether, there are often too many variables involved compared to the number of sampling units: the system becomes over-determined. To compare the patterns of two spaces, a most elegant solution is the correlation of resemblance matrices. In this, all pairwise comparisons of sampling units are evaluated simultaneously. These may concern the biological data space versus the environmental, but also the spatial (i.e. the arrangement of plots in space) and the temporal (i.e. the states in time). This approach has the additional advantage that it is only moderately affected by nonlinearity.

The example presented here may appear too simple because one of the spaces is represented by one vector only (the environmental). But a one-dimensional vector is sufficient to calculate a full resemblance matrix. Let us consider the following situation:

Relevé	1	2	3	4
pH	4.5	4.1	4.2	3.8
Species 1	0	1	1	2
Species 2	3	2	2	1

The pH vector represents the environmental space. The pattern of this is given by the distance matrix, D_e. Since this is one-dimensional, the distance between any two relevés is the difference of the respective pH values:

$$\mathbf{D_e} = \begin{array}{llll} 0 & & & \\ 0.4 & 0 & & \\ 0.3 & 0.1 & 0 & \\ 0.7 & 0.4 & 0.3 & 0 \end{array}$$

The two species vectors represent the floristic space. This pattern is defined by the distance matrix D_f. Using equation 4.1 (Euclidean distance) we get:

$$\mathbf{D_f} = \begin{matrix} 0 & & & \\ 1.41 & 0 & & \\ 1.41 & 0 & 0 & \\ 2.5 & 1.41 & 1.41 & 0 \end{matrix}$$

For the purpose of comparison, the elements of the triangular matrices are now arranged as vectors:

$$\mathbf{D_e}; \mathbf{D_f} = \begin{matrix} 0.4 & 1.41 \\ 0.3 & 1.41 \\ 0.1 & 0 \\ 0.7 & 2.5 \\ 0.4 & 1.41 \\ 0.3 & 1.41 \end{matrix}$$

The two vectors differ in scale, and for comparison they have to be adjusted. Mantel (1967) suggested using the correlation coefficient to measure the fit (see also Legendre & Fortin 1989). As shown in Table 4.3, the product-moment correlation coefficient involves standardization. In the example above we get:

$$\begin{aligned} r(D_e; D_f) &= 0.965 \\ p &= 0.015 \end{aligned} \qquad (7.7)$$

Where does the error probability, p, come from? In statistical textbooks tables of significance levels for correlation coefficients can be found. In this situation, however, they are not valid. The values in the two vectors are not normally distributed; nor do they represent a random sample. As in many other instances in ecology a randomization test is more appropriate. This measures the probability that this result could be obtained by chance. The elements of one vector are rearranged randomly and the calculation of the correlation coefficient is repeated. The result, $p = 0.015$, means that in 15 out of 1000 cases the random order yields an $r \geq 0.965$, sufficient for assuming significance. In the present case there are limitations to this test as the number of elements, six, is rather low. More reliable results are obtained when $n \geq 10$, where the number of off-diagonal elements in the resemblance matrix is $n \geq 45$. Clearly, the Mantel test is a practical means of quickly evaluating a set of site factors for their potential in predicting the similarity pattern based on species composition.

7.3.2 Correlograms: Moran's *I*

Correlograms are mostly used for spatial and temporal analysis. They help identify specific relationships, such as autocorrelation, periodicity and nonlinearity. Basically, they can be applied to any type of ordinal or metric data. Below, I use spatial coordinates as an example. In this (i.e. the physical space) there are different phenomena to be distinguished (Legendre & Legendre 1998):

Spatial dependence. This occurs when there is a gradient present in the investigation area. Along a gradient, plots in close neighbourhood are more similar than distant plots. The site conditions – and most likely also the vegetation – change from one end to the other. Because of the lack of spatial equilibrium such systems are called nonstationary.

Anisotropy. The above-mentioned case of spatial dependence usually differs by direction. Along the main gradient, dependence is strong. Perpendicularly, it is weak or even lacking. If the same dependence exists in all directions, the system is isotropic. Anisotropy is the rule in ecological systems.

Spatial autocorrelation. Even if the system is stationary (i.e. free of an overall gradient), correlation can be observed at small distances. When analysing a system with inherent autocorrelation, the species composition in neighbouring plots is more similar than could be expected from the measured site conditions. Often, it is assumed that this is caused by species propagation: whatever the environmental factors are, it is easier for a species to reach a close plot than to a distant plot.

The principle of measuring autocorrelation at different step lengths is illustrated in Table 7.4. Along a hypothetical gradient, 10 by equal step length, Δs, spaced measurements x_i are taken. This is equivalent to comparing a vector to itself, but with its elements shifted by a distance of $\Delta s = 1$. Consequently, the first and the last element of each vector cannot be used anymore and n reduces to 9, starting with paired values {4.5;4.7} and ending with {4.8,5}. Correlation is $r = 0.74$, indicating that at this distance the

Table 7.4 Autocorrelation in a one-dimensional gradient (Figure 7.3, left). Measurements taken at equal step lengths, Δs, are correlated with neighbours 1–6 steps apart.

Δs	0	1	2	3	4	5	6
vector	x	$y(\Delta s = 1)$	$y(\Delta s = 2)$	$y(\Delta s = 3)$	$y(\Delta s = 4)$	$y(\Delta s = 5)$	$y(\Delta s = 6)$
n=		9	8	7	6	5	4
r=		0.74	0.33	−0.07	−0.70	−0.90	−0.42
x(1)	4.7						
x(2)	4.5	4.7					
x(3)	4.4	4.5	4.7				
x(4)	4.6	4.4	4.5	4.7			
x(5)	4.9	4.6	4.4	4.5	4.7		
x(6)	5.1	4.9	4.6	4.4	4.5	4.7	
x(7)	5.1	5.1	4.9	4.6	4.4	4.5	4.7
x(8)	5.3	5.1	5.1	4.9	4.6	4.4	4.5
x(9)	5	5.3	5.1	5.1	4.9	4.6	4.4
x(10)	4.8	5	5.3	5.1	5.1	4.9	4.6
		4.8	5	5.3	5.1	5.1	4.9
			4.8	5	5.3	5.1	5.1
				4.8	5	5.3	5.1
					4.8	5	5.3
						4.8	5
							4.8

neighbours are correlated. When Δs is increased to two steps, n reduces to 8, with the first pair being {4.4;4.7} and the last {4.8;5.3}. Correlation now also decreases to $r = 0.33$ (a value that, in statistical terms, is no longer significant). The example in Table 7.4 is a typical case in which correlation decreases with increasing step length and eventually even becomes (significantly) negative: whenever one observation in a pairwise comparison is high, its distant counterpart will be low. In Figure 7.3, left side, it can be seen why this happens: here, the vector pH is plotted against an arrangement of plots in space. It turns out that it has a periodic spatial pattern!

Table 7.5 Computed correlogram of the data shown in Table 7.4. Graphical representation shown in Figure 7.3, right.

Dist. Class	from	to	n	Moran's I
1	0.5	1.5	9	0.38
2	1.5	2.5	8	0.11
3	2.5	3.5	7	–0.01
4	3.5	4.5	6	–0.15
5	4.5	5.5	5	–0.21
6	5.5	6.5	4	–0.22
7	6.5	7.5	3	–0.18
8	7.5	8.5	2	–
9	8.5	9.5	1	–

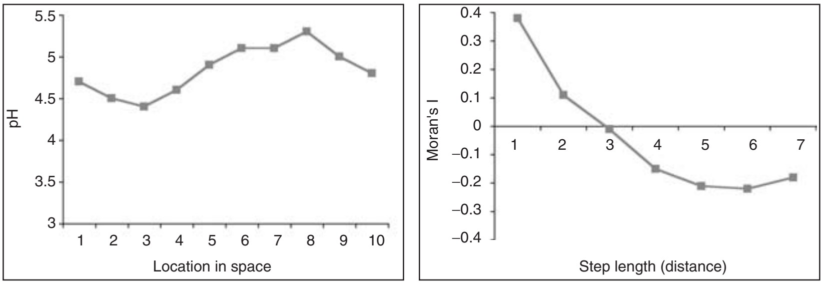


Figure 7.3 Left: spatial arrangement of measurements (pH) according to Table 7.4. Right: corresponding correlogram (data in Table 7.5).

A correlogram is a plot of correlation as a function of step length. Instead of the product-moment correlation coefficient, Moran's I is usually used:

$$I(d) = \frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - \overline{y})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} \qquad \mathit{for} \qquad h \neq i \qquad (7.8)$$

where W is a transformed distance matrix: it considers classes of distance, such as *short*, *medium*, *long*. When an element w_{hi} falls into the class to be analysed, it takes the value of one; otherwise it is zero. In our very small example in Table 7.4, there are possible step lengths between 1 and 9. These are taken as distance classes to compute the correlogram shown in Table 7.5

and graphically in Figure 7.3, right. There are nine distances of length 1 in this data set and the corresponding Moran's I is 0.38 (a significant value according to a permutation test). Step length 2 occurs eight times only, step length 3 yields seven pairs and so on. If distance class is plotted against Moran's I, the correlogram shown in Figure 7.3, right, results. Here, a clear trend of decreasing dependence is found up to step length 6, indicating the presence of spatial dependence. Due to the low number of elements, it cannot be decided if dependence vanishes at long distance. At this point, the Mantel test is more helpful (Section 7.3.1). We get $r = 0.3107$ and $p = 0.1000$: there seems to be a faint trend, but again the data set is too small for a clear answer.

7.3.3 Spatial dependence: Schlaenggli data revisited

This is an example of real-world data illustrating the application of the Mantel test, its directed version and also Moran's I. The investigation area (see for example Figure 8.5 and Appendix B) has the interesting property that it is almost quadratic in shape and trends can therefore be evaluated in different spatial directions. Furthermore, there are many site factors available; some of these correlate with vegetation while others do not (see Section 7.2.3).

From Figure 8.3, upper-left graph, it can be seen that there is a strong floristic gradient in the vertical direction ($\alpha = 90^\circ$). This suggests that the species pattern is likely to be space-dependent. Using the correlation matrix of the relevés and the matrix of Euclidean distances computed from the x- and y-axes in space, the Mantel test yields $r = -0.5204$ and $p = 0.0000$. Hence, spatial dependence is highly significant.

Since this dependence has its origin in a gradient, direction is a major issue. To evaluate different directions, the spatial distances have to be projected on one line. This yields new distances from which Mantel's r or Moran's I is computed. The way distances are projected at an angle of $\alpha = 45^\circ$ is shown in Figure 7.4. In this direction, the Mantel test yields $r = -0.4980$ and $p = 0.000$, a highly significant trend. At an angle of $\alpha = 0^\circ$, where the vertical component of the space is suppressed and only the horizontal expansion is considered, the test yields $r = -0.0884$ and $p = 0.014$. This means that there is still a trend, but much weaker than that in the vertical direction ($\alpha = 90^\circ$).

A full evaluation of all directions in the range of $0^\circ \leq \alpha \leq 180^\circ$ allows identification of the direction in which the gradient is strongest. This is shown in Figure 7.5. The maximum is achieved at $\alpha \approx 75^\circ$, indicating that the main gradient points from the upper-left to the lower-right corner; that

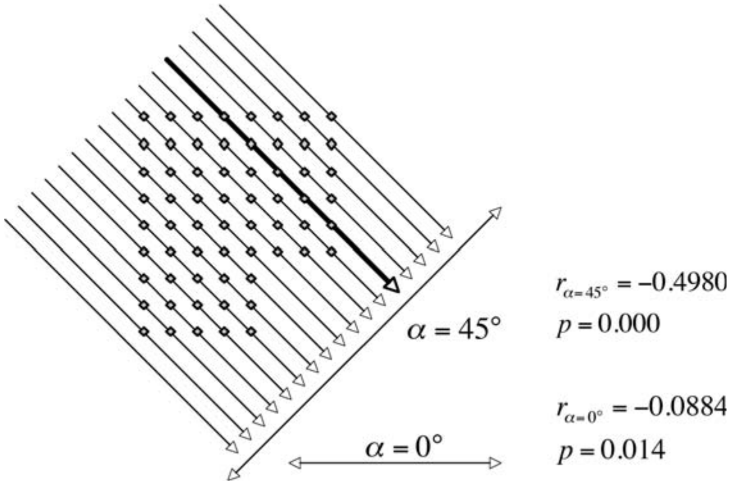
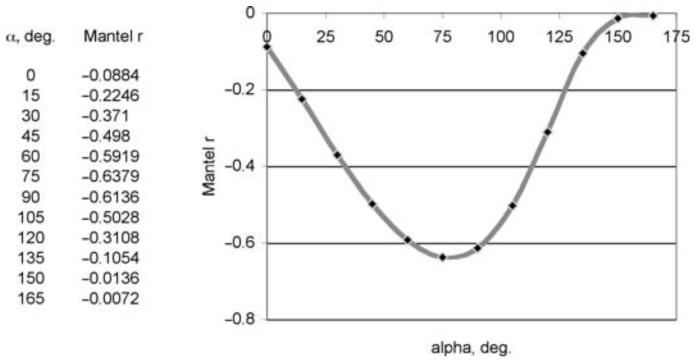

$r_{\alpha=45^\circ} = -0.4980$
$p = 0.000$
$\alpha = 45^\circ$
$r_{\alpha=0^\circ} = -0.0884$
$p = 0.014$
$\alpha = 0^\circ$

Figure 7.4 Projecting distances in one direction. Without considering direction, the neighbouring data points located on the same arrow are separated by 1.414 grid units. When projected at an angle of 45 $^\circ$, they are at the same location.



α, deg.	Mantel r
0	-0.0884
15	-0.2246
30	-0.371
45	-0.498
60	-0.5919
75	-0.6379
90	-0.6136
105	-0.5028
120	-0.3108
135	-0.1054
150	-0.0136
165	-0.0072

Figure 7.5 Evaluating the direction of the floristic gradient. The gradient is strongest at $\alpha \approx 75\,^\circ$.

is, in an almost vertical direction. Perpendicular to this, at $\alpha \approx 165\,^\circ$, the floristic gradient vanishes ($r = -0.0072$).

It can be seen in Figure 7.4 that the sample space, when projected, is one-dimensional only. This is also the case when taking one single site factor instead of spatial axes. In Section 7.2.3 all site factors have been evaluated based on their potential in predicting a specific classification of the relevés. In the present context, this potential is measured independent of classification, based on the full similarity matrices of the relevés. In Table 7.6 the same site factors used in Figure 7.2 are subjected to the Mantel test. For the purpose of comparison, the F-values from Section 7.2.3 are also shown (from the classification based on the transformation $x' = x^{0.25}$). Mantel's r is always negative because the site factors are compared by distance, whereas for the relevés the correlation coefficient (a similarity measure) is used. The results are significant in all cases except the last, where a random variable is used instead of a site measurement.

Moran's I can be computed as well and it will yield a correlogram instead. In the example shown in Figure 7.6 the distance (dissimilarity) matrices are classified. The classes are formed by dividing the longest distance encountered into 10 segments, from which Moran's I is calculated. Figure 7.6 shows correlograms of four different site factors in one graph. Although this helps in the comparison of the curves, interpretation has to be carried out with care: the distance matrices differ among site factors and Moran's I values are based on a slightly different number of data pairs. Generally, the results at distance classes 9 and 10 become unreliable due to insufficient sample size.

The strongest dependence occurs with pH. From distance classes 2–6 the change of Moran's I is almost linear. Only then does the trend level off, indicating nonlinearity occurs at larger differences in pH. The water level yields a similar overall shape of the correlogram, but much less pronounced. Random fluctuation plays a more visible role than in pH. At distance class

Table 7.6 Mantel test of the site factors analysed in Section 7.2.3. F-values are added for comparison.

No.	Site factor	F-value	Mantel's r	Permutation p
1	pH peat	33.87	−0.649	0.000
3	Ca (mg/100g peat)	19.09	−0.634	0.000
9	Base saturation (%)	40.63	−0.737	0.000
12	Waterlevel, av. (cm)	9.58	−0.346	0.000
14	Peat depth (log(cm))	9.27	−0.400	0.000
15	Slope (log(deg.))	4.15	−0.132	0.029
17	Conductivity (log(Ohm/cm))	24.38	−0.597	0.000
18	Ca in water (log(0.1ppm))	55.80	−0.755	0.000
21	random variable	0.57	0.050	0.228

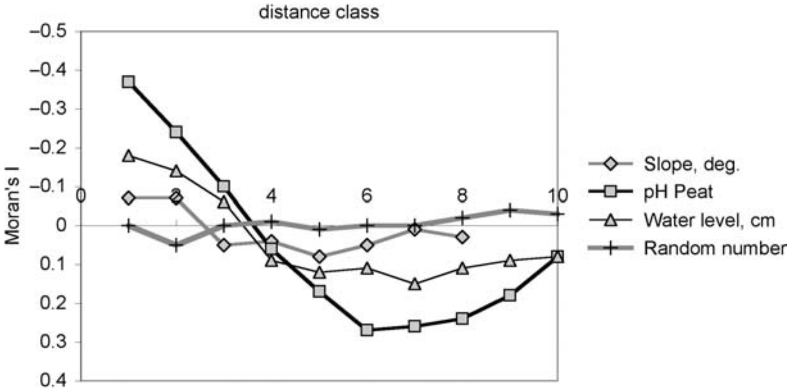


Figure 7.6 Correlograms of site factors. The random variable is an example illustrating the shape when there is no relationship.

1, where small differences are taken into account, Moran's I is still rather reliable. Slope, on the other hand, is an example for weak dependence. The correlogram ends at distance class 8 because n is too small to calculate I at distance classes 9 and 10. The deviations from the zero line hardly exceed what can be expected from a random number, which is also included to illustrate the lack of relationship.

7.4 Contingency tables

All methods of classification and ordination are aimed at analysing data matrices from full samples. However, the basic idea presented in this section is to analyse data in summarized form; that is, after classification of relevés and species. This should reveal interactions between relevé groups and species groups. A contingency table contains counts of occurrence of one or several species within a relevé group, resulting in a comparison of two alternative classifications of the same body of data. During the course of this book we have encountered contingency tables twice, but in different contexts: in Section 4.3 the occurrence of the same species in two different relevés was counted to derive measures of resemblance for relevés; in Section 5.4 correspondence analysis (CA) was introduced as a method of ordination. In CA, the species scores are assumed to be frequencies and the entire vegetation table is considered a contingency table – despite the fact that scores are usually not frequencies, but cover percentage, abundance classes, biomass and so on. In the applications shown below presence–absence scores are counted. The counts are the elements of the contingency table F and the notation used for all elements of the table accords to:

$$\mathbf{F} = \begin{array}{ccccc|c} f_{11} & \dots & f_{1j} & \dots & f_{1n} & f_{1.} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ f_{i1} & \dots & f_{ij} & \dots & f_{in} & f_{i.} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ f_{p1} & \dots & f_{pj} & \dots & f_{pn} & f_{p.} \\ \hline f_{.1} & \dots & f_{.j} & \dots & f_{.n} & f_{..} \end{array} \tag{7.9}$$

where $f_{i.}$ is the sum of the n elements in row i, $f_{.j}$ is the sum of column j with p elements and $f_{..}$ is the grand total. There are n columns and p rows in this matrix.

The method shown below was first proposed by Feoli & Orlóci (1979) under the name 'analysis of concentration' (AOC). 'Concentration' refers to the allocation of non-zero scores within the blocks of relevés and species groups. Concentration is high when the counts are highly concentrated in a few blocks while other blocks are empty. It is low when the scores are dispersed all over the contingency table. The aim of the method is to measure concentration and reveal interactions among and between the classifications of relevés and species.

The method is illustrated using the classified data set shown in Table 7.7. It is the same as that in Table 7.1, but with all species included in the classification. Counting the species scores yields the following contingency table:

$$\mathbf{F} = \begin{array}{ccc|c} 18 & 0 & 0 & 18 \\ 27 & 25 & 21 & 73 \\ 0 & 7 & 6 & 13 \\ 0 & 1 & 14 & 15 \\ \hline 45 & 33 & 41 & 119 \end{array} \tag{7.10}$$

Typically, the sizes of the 12 blocks differ in Table 7.7, while in the final analysis each group is intended to have the same weight. Therefore, an appropriate adjustment is needed. In the first step, the number of relevés and species per group are counted to yield block size, Z:

$$\mathbf{Z} = \begin{array}{ccc|c} 20 & 20 & 15 & 55 \\ 36 & 36 & 27 & 99 \\ 8 & 8 & 6 & 22 \\ 20 & 20 & 15 & 55 \\ \hline 84 & 84 & 63 & 231 \end{array} \tag{7.11}$$

Table 7.7 The structured data set 'nzzm5' (Appendix B). The method used for ordering is the same as in Table 7.1, but without removing species with low F-values.

	RELEVE NO.		49	25	2	6	39	18	50	9	4	10	27
	GROUP NO.		1	1	1	1	3	3	3	3	2	2	2
1	pH		4.8	5	4.4	4.8	5.2	6	5.8	5.6	6.2	6.5	6
2	Altitude		550	480	450	420	500	400	520	580	500	560	450
3	Slope, deg.		15	18	10	4.5	12.5	2.5	3	6	0.5	4	2.5
8	*Vaccinium myrtillus*	3	+		2	1							
6	*Sambucus racemosa*	3	+		2	1							
21	*Polytrichum formosum*	3	3	+	+	1							
13	*Veronica officinalis*	3	+	+	1	1							
12	*Luzula nemorosa*	3	2	1	2	+							
2	*Quercus petraea*	4	4	3	1	2	+	+		1			+
9	*Carex silvatica*	4	+		+	1	+	2		+	+		2
20	*Eurhynchium striatum*	4	+	+	1		+	+		1	+		+
11	*Viola silvestris*	4	+		+	+	+		+	1	+		+
1	*Fagus silvatica*	4	2	4	3	4	5	5	+	4	+	2	1
10	*Oxalis acetosella*	4		+	+	+	3	1		1	1	2	
14	*Galium odoratum*	4	+	+	+		+	2		1	+	+	1
15	*Lamium galeobdolon*	4			+	+	+			+	2	1	2
5	*Lonicera xylosteum*	4		+	+		+				1	+	1
3	*Acer pseudoplatanus*	1					+	+	1	1	1	2	4
4	*Fraxinus excelsior*	1						+	1	+	3	3	2
7	*Sambucus nigra*	2						+			+	2	1
18	*Arum maculatum*	2									1		+
19	*Ranunculus ficaria*	2									1	2	4
16	*Primula elatior*	2									+	2	2
17	*Allium ursinum*	2									4	2	+

Block sizes range from 6 to 36. In the original publication, all frequencies were adjusted to the minimum block size. Orlići & Kenkel (1985) later proposed a method in which the grand total is retained. This is the case when:

$$a_{ij} = \frac{\frac{f_{..} f_{ij}}{n_{ij}}}{\sum_{g=1}^{p} \sum_{h=1}^{q} \frac{f_{gh}}{n_{gh}}} \tag{7.12}$$

In the denominator of this equation, the sum of all frequencies weighted by the inverse of block size is used. In the above example, this value is 5.980. The adjustment of the first element thus yields:

$$a_{11} = \frac{\frac{119*18}{20}}{5.980} = 17.908$$

The matrix of the adjusted frequencies is:

$$\mathbf{A} = \begin{array}{ccc|c} 17.908 & 0.0 & 0.0 & 17.908 \\ 14.923 & 13.818 & 15.476 & 44.217 \\ 0.0 & 17.411 & 19.898 & 37.309 \\ 0.0 & 0.955 & 18.571 & 19.526 \\ \hline 32.831 & 32.184 & 53.945 & 119.0 \end{array} \tag{7.13}$$

It can be seen that this transformation changes the elements considerably in the case of both small (e.g. row 3) and large (e.g. row 2) groups. These adjusted frequencies are now further analysed by correspondence analysis (CA), as shown in Section 5.4. This yields ordination coordinates for relevé and species groups. In the example above, there are two non-zero Eigenvalues (according to the dimension of the data matrix minus 1):

$$\lambda_1 = 0.58834 \qquad \lambda_2 = 0.1221$$

The squared canonical correlations are the square roots of these:

$$R_1 = 0.76703 \qquad R_2 = 0.34947$$

CA is based on deviations from expectation 5.4. ‘Expectation’ is the assumption that all frequencies are evenly dispersed across all blocks, which is the case when there is no structure in the table. The χ^2 by definition sums the squared deviations from expectation and is therefore a measure of concentration. It is obtained from the squared canonical correlations, multiplied by the grand total of F (or A, which is the same):

$$\chi^2 = \chi_1^2 + \ldots + \chi_q^2 = R_1^2 f_{..} + \ldots + R_q^2 f_{..} \tag{7.14}$$

This shows that the χ^2 is in fact the sum of q orthogonal components. In the example we get:

Component i	R_i	χ_i^2	$\lambda_i\%$	df
1	0.76703	70.01	82.8	4
2	0.34947	14.53	17.2	2
Total		84.54	100	6

The degrees of freedom are calculated according to:

$$df = (p-1) + (q-1) - (2i-1) \tag{7.15}$$

(from Orlóci & Kenkel 1985).

In statistical textbooks we find that the significance threshold for $df = 6$ and $p = 0.01$ is $\chi^2 = 16.812$. Hence, the value of $\chi^2 = 84.54$ is highly significant. For comparison of classifications it is quite practical to consider the mean square contingency coefficient, C:

$$C = \frac{\chi^2}{A_{..}(m-1)} \tag{7.16}$$

where m is equal to the smaller of the values n and p. C lies between 0 and 1; in the example above $C = 0.355$.

Just like correspondence analysis, this method yields coordinates. However, the data points now refer to the groups rather than to the individual relevés and species. This simplifies the interpretation considerably (Figure 7.7). In the present example, the ordination confirms what is obvious from the ordered vegetation table: that there is a correspondence between relevé group 1 and species group 3, between relevé group 3 and species group 1, and also between the respective groups number 2. Species group 4 is intermediate and not indicative for any other, being located close to the origin of the coordinate system.

7.5 Constrained ordination

In this category of methods two data matrices are analysed in common; hence the umbrella term *canonical analysis* for all related methods (Legendre & Legendre 1998). Typical results are ordinations with three types

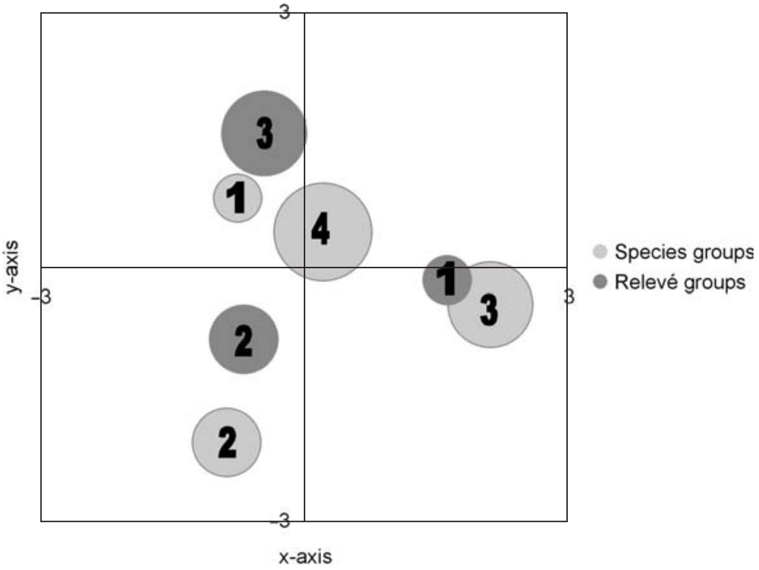


Figure 7.7 Ordination of group structure in the test data set 'nzzm5' as derived from the data in Table 7.7.

of data points: one for sites, a second for environmental variables and a third for species. What distinguishes constrained ordinations from ordinary is the involvement of regression. Regression is partitioning dependent vectors, such as the species scores, into two components: the expected and the deviation from this. In linear regression the expected values are the projections of the scores on the straight regression line. These are the scores used for constrained ordination, highlighting two main issues of constrained ordination:

- The multiple regressions involved divide the total variance into an explained partition (the expected) and an unexplained partition (the residuals). The quotient of explained by total expresses how much variance is common to both data matrices, for example.

- The explanatory variables, such as the environmental vectors, can be subjected to permutation tests. When the elements of the vectors are randomly exchanged, correlation is expected to vanish. Hence, the significance of the environmental factors in contributing to the canonical ordination can be tested.

Randomization tests have to be used with care, as explained by Lepš & Šmilauer (2003) in the context of using the program package CANOCO. All major ordination methods can potentially be extended to a constrained form:

- Redundancy analysis (related to principal component analysis) was first proposed by Rao (1964).
- Canonical correspondence analysis (related to correspondence analysis) was invented by ter Braak (1986).
- Constrained principal coordinates analysis was proposed by Legendre & Anderson (1999) (although they called it 'distance-based redundancy analysis').
- Wagner (2004) devised a spatially constrained correspondence analysis, revealing the effects of spatial dependence and spatial autocorrelation.

In all methods the coordinates generated by programs can be transformed differently and many programs offer options for this. Comparison with the output of different software is further complicated when randomization tests are involved, because random numbers differ between computer programs and sometimes even within the same program run.

The first method shown below is *redundancy analysis* (RDA), the constrained version of principal component analysis (PCA). It differs from the latter in that expected species scores are used instead of the originals. Expectation $\hat{y}_j$ is derived through multiple regression, according to:

$$\hat{y}_j = a_j + b_{j,1}x_{j,1} + b_{j,2}x_{j,2} + \ldots + b_{j,k}x_{j,k} \tag{7.17}$$

The $x_{j,k}$ values are the k environmental factors in each relevé j. The $b_{j,k}$ values are the regression coefficients and a_j is the intercept. Usually the data matrices are centred by relevés such that the intercept vanishes. Denoting the species by relevé matrix Y, a new matrix of expected values, $\hat{Y}$, is obtained, where:

$$Y = \hat{Y} + Y_{res} \tag{7.18}$$

This relationship is of practical relevance as it expresses the fraction of explained (canonical) variance, $\hat{Y}$, compared to the total, Y. Operations explained in Section 5.2 are performed on $\hat{Y}$ and permutation tests are applied on request, as explained above. Hence, RDA is, like PCA, a linear method.

The second method shown below is *canonical correspondence analysis* (CCA), the constrained version of correspondence analysis (CA). It differs from RDA in that the transformation used in CA (Equation 5.5) is applied prior to multiple regression. As in the unconstrained versions, CCA is considered appropriate when species response is unimodal.

To demonstrate the use and interpretation of constrained ordination I compare redundancy analysis and canonical correspondence analysis using data set 'Schlaenggli' (see Appendix B). From the 20 environmental factors I use 5 for the purpose of illustration:

- pH of peat
- Acidity (mval/100g peat)
- Cation-exchange capacity, CEC (mval/100g peat)
- Phosphorous (mg/100g peat)
- Water level (average), cm below surface.

The results are shown in Figure 7.8. Overall performance of the ordinations is as follows:

Method	Constrained variance	Unconstrained variance	Total	Percentage constrained
RDA	7.689	14.446	22.135	34.7%
CCA	0.6302	1.5708	2.1911	28.3%

Performance of both ordinations is rather high, with RDA being slightly better (34.7% explained variance) than CCA (28.3%). In other words, the linear model succeeds in revealing correlation between vegetation and environmental factors despite the strong horseshoe pattern (Section 5.5). Two almost independent factors dominate the system: pH in peat and the depth of the average water table. CEC and acidity are highly correlated, while phosphorous has its maximum in peat bog vegetation on the left, a fact already recognized in the original investigation (Wildi 1977). The overall pattern of RDA is rather similar to the unconstrained version of PCA, as can be seen in comparison with Figure 5.11, whereas CCA is similar to CA

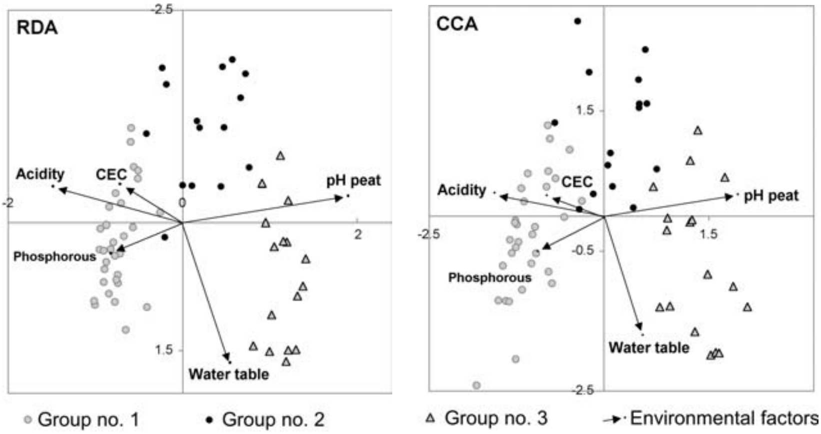


Figure 7.8 Comparison of RDA and CCA, data set 'Schlaenggli' (Appendix B). In both graphs the environmental factor vectors are multiplied by factor two for better resolution.

(Figure 5.10). In conclusion, both methods perform very well and choosing one is probably just a question of taste.

8

Static explanatory modelling

8.1 Predictive or explanatory?

The term ‘modelling’ is used in a wide context and it deserves closer specification. Loehle (1983) suggested a classification of models, finding them to be either *logical*, *theoretical* or *predictive*. Logical and theoretical models have their strength in the universality of validity: the systems described are assumed to be governed by generally valid rules and laws, although the parameters may still come from measurements. The models shown in this chapter operate in the realm of probability and Loehle (1983) would classify these as ‘predictive’. Correlative relationships are being used as tools for ‘forecasting’. Really predicting the future is of course out of their scope: the models merely assess a most probable state – based on past experience. As they are data-driven they also reflect uncertainty: that found in the underlying investigations. In linear regression, uncertainty is the variance not explained by the straight regression line. As for ecology, a more

Data Analysis in Vegetation Ecology Otto Wildi

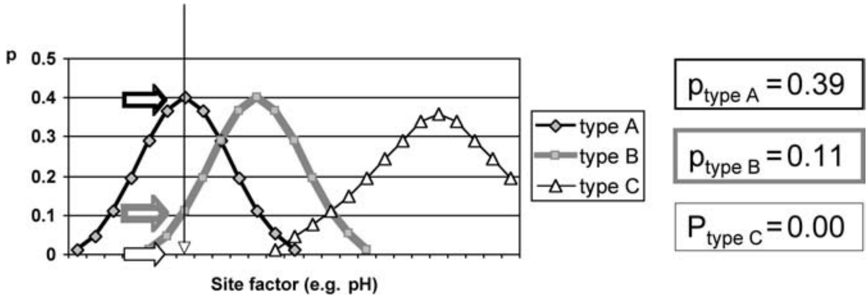


Figure 8.1 Occurrence probability of three hypothetical vegetation types at a given environmental state (pH). Response type is Gaussian.

typical function than the linear is shown in Figure 8.1. When counting the frequency of pH values in different vegetation classes, a frequency distribution results. This can be interpreted as an approximation to a probability function. Whenever a pH value is measured in the real system, the model provides occurrence probabilities for all vegetation types, without referring to specific mechanisms or theories. But as depicted in Figure 8.1, probability functions are often of the Gaussian type, an experience corroborated by a huge body of investigations (Austin 2005). Thus, in Loehle's terms, all these models also involve a logical component. Taking into account all the assumptions and data used, they are *explanatory*, *evidence-based* and to some extent *logical*. Prediction only comes along in the course of interpretation.

This chapter presents models assuming only Gaussian response of species to environmental factors. An alternative class of approach is regression-based, both linear and nonlinear. For an in-depth review of these I refer the reader to the paper of Guisan & Zimmermann (2000), in which general linear models (GLM) and others (e.g. general additive models, GAM) are presented and compared. Comparison of model performance is a big issue and Elith *et al.* (2006) give an impressive example.

8.2 The Bayes probability model

The occurrence of a species population, a community, a vegetation type or a life form can be described by probability functions. A theoretical framework for handling multiple probabilities is given by the Bayesian type of analysis (Fischer 1990), based on a posteriori probabilities. An example of a Bayesian model is given by Brzezecki *et al.* (1993): a simulation of the potential forest

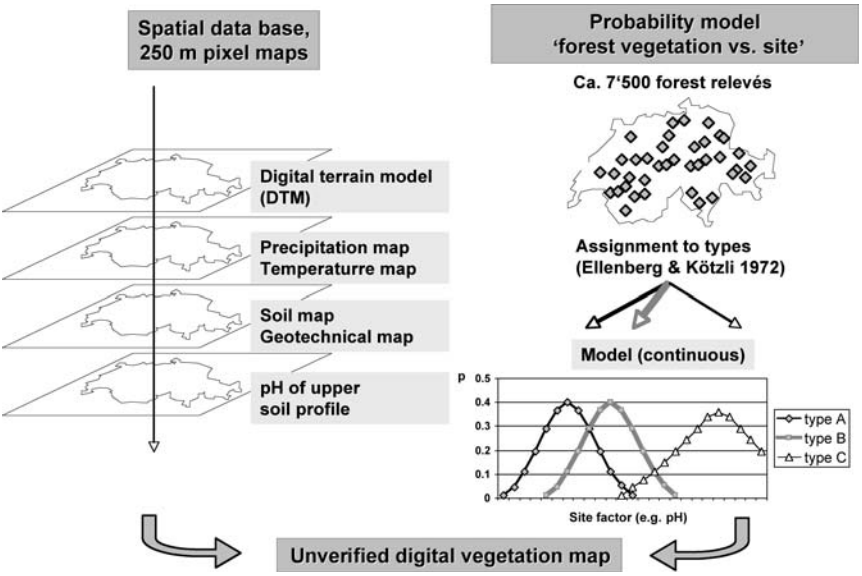

Spatial data base, 250 m pixel maps
Digital terrain model (DTM)
Precipitation map Temperaturre map
Soil map Geotechnical map
pH of upper soil profile
Probability model 'forest vegetation vs. site'
Ca. 7'500 forest relevés
Assignment to types (Ellenberg & Kötzli 1972)
Model (continuous)
p
0.5
0.4
0.3
0.2
0.1
0
type A
type B
type C
Site factor (e.g. pH)
Unverified digital vegetation map

Figure 8.2 Schematic illustration of the construction of a Bayes probability model, used for the simulation of the potential vegetation of Switzerland (Brzezecki *et al.* 1993).

vegetation of Switzerland. The scheme in Figure 8.2 illustrates the steps involved in this kind of modelling; that is, the process of model construction, as described in more detail below. On the right side, upper part, the derivation of this model is sketched. The outset is a sample of the vegetation relevés of Switzerland. These are classified into types. In addition to this, spatial information on site factors is needed, as shown on the left. For each site factor the frequency distribution within each vegetation type is derived. This serves as an approximation to a probability function for the occurrence of any one vegetation type. Processing the joint probabilities will then yield a spatial map of the probability of occurrence of all vegetation types in Switzerland. From this, a crisp map showing the vegetation types with the highest probability within each pixel can be drawn. There are different ways to derive probabilities, depending on whether a continuous or a discrete approach is preferred. The choice depends on the type of data available and also on the characteristics of the probability distribution, as shown below.

In the following sections the Bayes model is explained in detail. In Section 8.3 a small application to wetland vegetation is shown; in Chapter 12 the same is done for forest data.

8.2.1 The discrete model

The advantage of a discrete model lies in the fact that no assumption about the functional relationship between a vegetation type and a site factor is required. The model will properly reflect the real situation, even if the variable involved has an asymetric or bimodal distribution. The site factors are all assumed to be discrete states and their frequency of occurrence is counted. Continuous variables, like elevation, are classified into states, in this case as altitudinal classes. However, the estimation of probabilities requires samples of really large size since all classes have to be of sufficient size. In this regard the continuous model is less demanding. Whenever the discrete model is used, the database for the construction of the model is set up as a contingency table (Section 7.4), where the site-factor classes are the columns and the vegetation types are the rows. In all subsequent explanations the notation shown in Table 8.1 is used.

In reality, vegetation types vary in abundance. The Bayes model considers this through a priori probabilities, weighting the chances the different types have of occurring. An estimate for the occurrence probability of type (Vi) is its relative frequency taken from the survey:

$$p(Vi) = \frac{f_{i.}}{f_{..}} \tag{8.1}$$

where $f_{i.}$ is the frequency of locations at which type (Vi) is found and $f_{..}$ are all the locations studied. Sometimes one intends to apply the model to a different investigation area, which implies that it should be ‘neutral’,

Table 8.1 Notation used in Bayes modelling.

F	contingency table
f_{ij}	an element of F
i	index of the ith vegetation type $(i = 1, \ldots, m)$
j	index of the jth category of the site variable $(j = 1, \ldots, n)$
$f_{i.}$	sum of row i
$f_{.j}$	sum of column j
$f_{..}$	sum of all elements in F (grand total)
x_u	uth site factor $(u = 1, \ldots, \phi)$

giving each type the same chance of occurring. In this case, the a priori probabilities are set to the same (arbitrary) value.

The univariate conditional probability of occurrence of any site class, x_{uj}, at a given vegetation type, Vi, is then:

$$p(x_{uj}|Vi) = \frac{f_{ij}}{f_{i.}} \tag{8.2}$$

In most cases, however, there are several site factors used and the multivariate conditional probability of occurrence of a site vector (a combination of a number of site-factor classes) at a given vegetation type, Vi, is required:

$$p(x|Vi) = \prod_{u=1}^{\pi} p(x_u|Vi), \text{ for all site factors} \qquad x_u, u = 1, \ldots, \phi \tag{8.3}$$

Using the Bayes formula, the probability of vegetation types, Vi, being identified when the site vector, x, is observed is:

$$p(Vi|x) = \frac{p(x|Vi)p(Vi)}{\sum_{j=1}^{n} p(x|Vj)p(Vj)} \tag{8.4}$$

For any site vector, x_u, the method in fact yields a vector of probabilities of occurrence for all vegetation types, V. The term $p(Vi)$ denotes the a priori probabilities. It can easily be seen that they remain uninfluential if they are alike.

8.2.2 The continuous model

In the continuous case, the frequency distribution is replaced by a continuous function such as the Gaussian, an assumption inherent in the model. Here two parameters, μ and σ, have to be estimated. The univariate conditional probability of occurrence of any site condition under a given vegetation type, Vi, is given by:

$$p(x_u|Vi) = \frac{1}{\sqrt{2\pi}} e^{-(x'_u)^2/2.} \tag{8.5}$$

where x'_u is a standardized vector according to:

$$x'_u = (x_u - \mu_{ui})\sigma_{ui} \tag{8.6}$$

where μ_{ui} is the mean and σ_{ui} is the standard deviation of the uth site factor within vegetation type Vi. The computation of $p(Vi|x)$ is the same as in the discrete case.

A Bayes probability model is not restricted to the prediction of vegetation types; occurrence probabilities for all species can be computed as well. In a synoptic vegetation table (see Section 6.6), the relative frequency of occurrence, $s_{k,Vi}$, of species k within a vegetation type, Vi, is an estimate of probability. This species probability has to be multiplied by the probability of occurrence of the vegetation type in which it is present at a given location:

$$q_{k,Vi} = p(Vi|x) * s_{k,Vi} \tag{8.7}$$

It expresses how likely species k is to occur in Vi, given site vector x. As the calculations refer to a site vector, several vegetation types can be involved: the probability of occurrence of a species taking into consideration all vegetation types is:

$$q_k = \sum_{i=1}^{m} p(Vi|x) * s_{k,Vi} \tag{8.8}$$

In order to derive the likely state of the vegetation under the given site conditions, this computation is repeated for all species.

There are various restrictions on the use of Bayes modelling, as discussed by Brzeziecki *et al.* (1993) for example. Among the most obvious is the local relevance of the results. The probabilities only hold within the range of the site factors really measured. Another is the assumption of a Gaussian distribution, when in fact species responses to site factors are monotone within a limited range (Orlóci 1993). Furthermore, it is assumed that the site factors used are independent (which in reality is certainly not true). Many applications, such as those of Brzeziecki *et al.* (1993) prove that lack of independence of site factors hardly ever hampers the results.

8.3 Predicting wetland vegetation (example)

Real-world examples allow testing of methodological and practical questions, two of which I will address in this section:

1 Which proportion of wetland vegetation pattern can be explained by measured site factors (in terms of similarity between real and simulated vegetation)? What is the variation in model performance?

2 Is there a spatial pattern in model performance, for example an edge effect towards the borders of the investigation area, or a spatial gradient?

These questions shall be tested using the ‘Schlaenggli’ data set (Appendix B). The steps involved are:

1 Classify the 63 relevés (scalar transformation $x' = x^{0.2}$, minimum-variance clustering based on correlation coefficient, five groups formed). Plot types according to their location in space (x,y).

2 Build a Bayes probability model using the site factors pH and average water level (a priori probabilities adjusted to 0.2 to compensate for variable group size).

3 Calculate probabilities for all five vegetation types. Plot probabilities according to location in the field (x,y).

4 For all plots, calculate all species probabilities and compare the resulting simulated relevés with the originals.

The two site factors taken (for simplicity) are continuous. For each group, their mean and standard deviation is computed (Table 8.2). For the calculation of the probabilities of the occurrence of groups a Gaussian response function is generated using these parameters.

The result is presented in Figure 8.3. From this, the probability of species occurrence is derived according to Formula 8.8; three examples of this are shown in Figure 8.4. The probabilities express the suitability of the site for the occurrence of the respective species. As would be expected, there

Table 8.2 Mean and standard deviation within groups of pH and water level in the Bayes model of the ‘Schlaenggli’ data set.

Parameter	Group	1	2	3	4	5
pH	mean	6.01	6.14	4.40	5.40	5.00
	stdv.	0.348	0.350	0.426	0.370	0.542
Water level, cm	mean	20.4	12.7	12.8	8.57	12.4
	stdv.	2.13	4.61	3.80	2.64	1.82

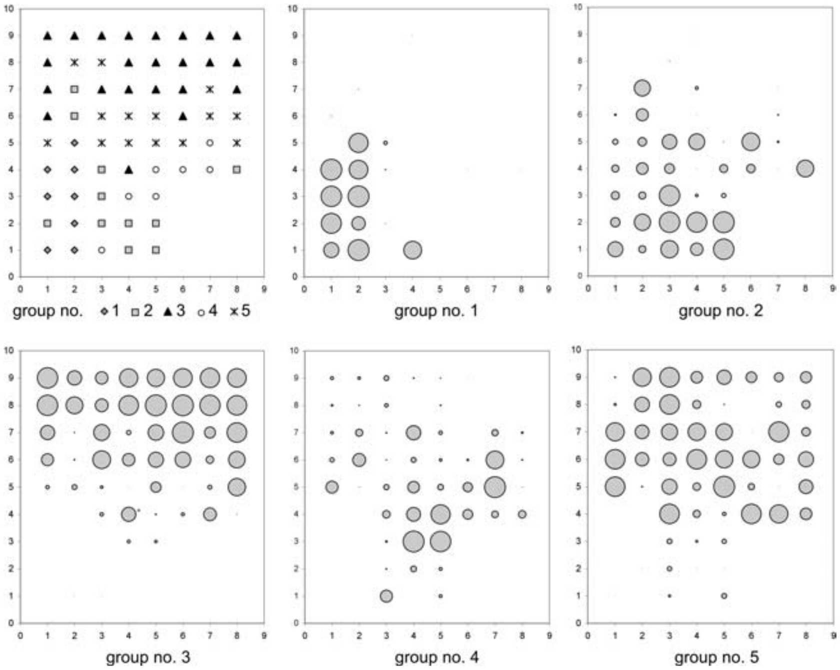


Figure 8.3 Site suitability computed with the Bayes model. The 'Schlaenggli' data set (Appendix B) is used. Upper left: classification of the relevés. Other graphs: probability of occurrence of all five types based on site conditions (pH and average water level). Circle diameter is proportional to probability.

are more plots with suitable site conditions than plots where the species is really observed. In this small meadow, 100 × 100 m, the most important reason for this may be the low population density of many species. Plot size (1m square) may often have been too small to capture an individual from a population that would otherwise be present.

A real model test would require independent data, which are not available in our case. What can be measured is the performance of the model to predict species composition from the measured site factors. This is needed as we would like to know if there is a systematic trend in the variation across the investigation area. For this, the original vegetation data have to be compared to the simulated. There are many different ways to do this, but since the objective is to identify trends (spatial patterns of performance), the method chosen is not too important. Only the relative magnitude of the similarity measures matters.

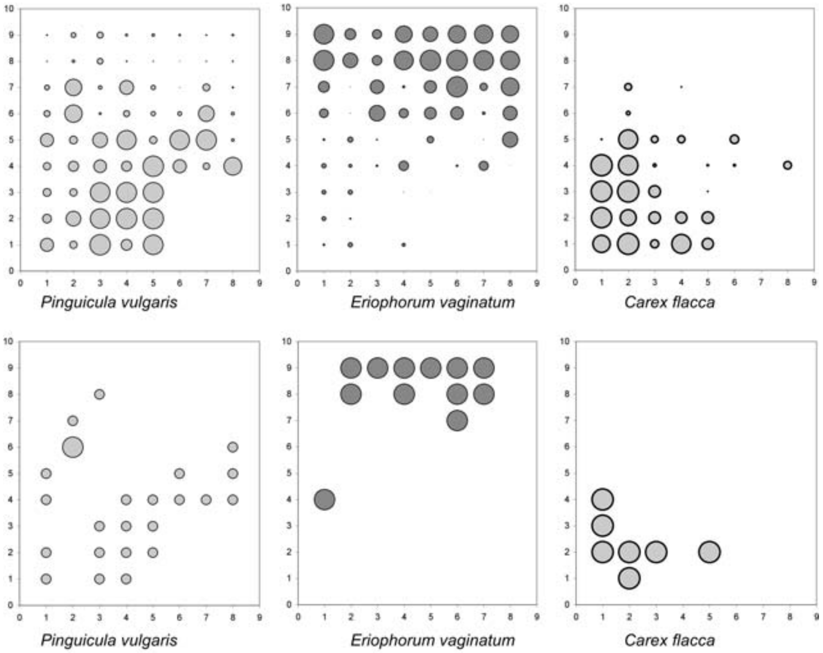


Figure 8.4 Occurrence probability of three selected species (upper graphs) and their occurrence in the relevé data (lower graphs).

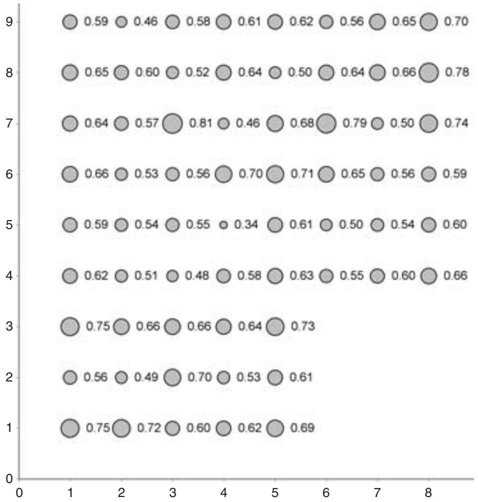


Figure 8.5 Similarity of field relevés and simulated data. The Soerensen coefficient is used. Axes are spatial coordinates x and y.

In a first step the data set with the probabilities is transformed into abundance scores. A presence–absence matrix is derived by setting a threshold; that is, by suppressing low probabilities and setting those remaining to 1.0. In the original data set about 30% of all elements consists of non-zero scores. When cutting the species list at a threshold of $p \leq 0.4$, the simulated data set contains 34% non-zero elements. Hence, whenever the probability for a species to occur in any one plot is $p \leq 0.4$ it is assumed to be absent. Comparison between real and simulated relevés is done with the Soerensen coefficient (Table 4.2), based on presence–absence. The average similarity achieved is 0.611, a fairly high value (cutting the species probability at $p \leq 0.2$ results in 48% non-zero elements and an average similarity of 0.57). A spatial representation of the fit of the model is shown in Figure 8.5. The concern about spatial trends in the simulation is not corroborated, as there are plots with better and others with poorer fit all over the investigation area.

9
Assessing vegetation change in time

9.1 Coping with time

Assessing change of vegetation as a multistate system is a central issue in vegetation ecology (Wildi & Orlóci 2007). One could of course argue that time is but another attribute in a sample, as explained in Section 2.3.2, such that no specific treatment would be indicated. However, time has some unique properties. For one thing it is one-dimensional, unlike space where direction is an issue and a decision may be needed when assessing order. Time always proceeds in the same direction, and even more importantly, it is transient. Once an event has taken place, there is no backtracking as can be done in space. This has consequences for investigating change in environmental systems, as discussed in detail by Green (1979). His hierarchical scheme of impact studies is reproduced in Figure 9.1. The most urgent

Data Analysis in Vegetation Ecology Otto Wildi

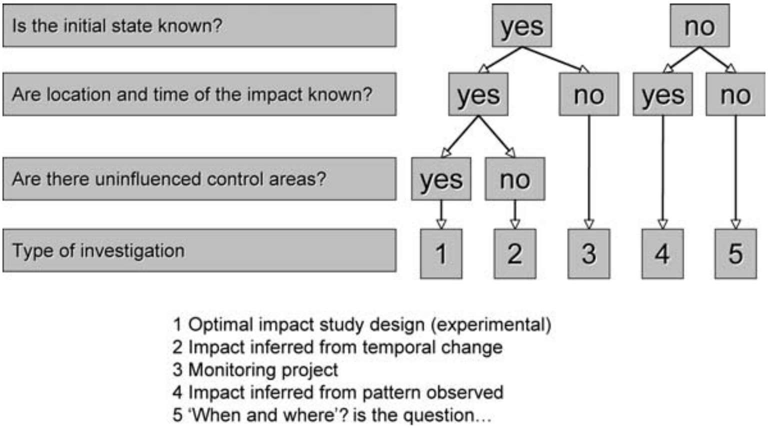

Is the initial state known?
Are location and time of the impact known?
Are there uninfluenced control areas?
Type of investigation
yes
no
yes
no
yes
no
yes
no
1
2
3
4
5
1 Optimal impact study design (experimental)
2 Impact inferred from temporal change
3 Monitoring project
4 Impact inferred from pattern observed
5 'When and where'? is the question...

Figure 9.1 Type of environmental study needed to assess change. Adapted from Green (1979).

question in the investigation of an impact is whether reference plots exist. Undoubtedly, striving for a reference is worthwhile, because once the impact has taken place there is no way to reverse the process. One may succeed in protecting plots from an impact, but plots cannot be sheltered from time passing by. Some more philosophical implications of time dependence are discussed in Legendre & Legendre (1998).

In the first part of this chapter I postpone the question of reference, concentrating on the temporal change of sampling units only. This is classical time-series analysis and in the literature of vegetation ecology it is usually found under the buzzword 'succession' (e.g. Maarel 2005). The ideas presented below focus on the specific problem of change in multivariate data space of high dimension, as is the case in relevé data.

9.2 Rate of change and trend

When sampling is repeated within a plot each new state will differ from the previous as a consequence of limited precision in measuring or change taking place in the system. If change occurs, it can be blurred by noise and a trend may only emerge when sufficiently strong. Consequently, distinguishing randomness from trend is a mandatory prerequisite for any further step when looking at time series, and a method has to be found to investigate the nature and possible causes of change.

In the simplest case a time series consists of subsequent states separated by even time steps (Figure 9.2, left side). The five states documented

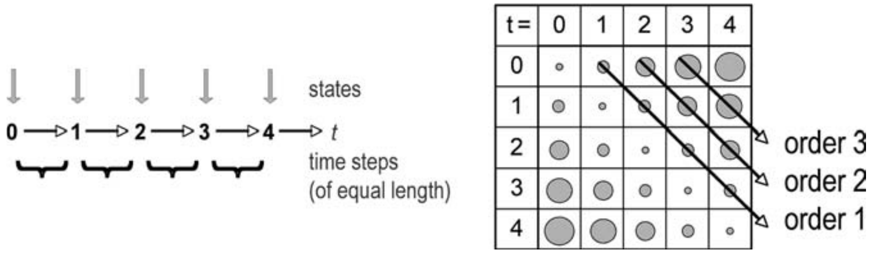


Figure 9.2 Measuring rate of change in time series of multistate systems. The matrix depicts distance (dissimilarity).

by five relevés are then compared by calculating similarities or distances (see Section 4.2), which yields a 5 by 5 distance matrix (Figure 9.2). In the diagonal the self-comparisons are found; these are all zero (identity). All comparisons of states separated by one time step can be found in the first off-diagonal vector. I call these *rate of change of order 1*. The number of comparisons, c, is:

$$c(o) = n - o$$

where n is sample size and o is the order of change. For order 1 there are four possible comparisons, for order 2 only three and so on. The distance matrix can now be interpreted. Simple reasoning leads to the following considerations:

1 A change of any order does not ultimately allow identification of the existence of a trend, as it can arise from any type of error.

2 Small increases in distance from short to long time steps may have been caused by an increase of measurement error. Errors of this kind can usually be avoided by quality control.

3 If distances continuously increase with the number of time steps, the rate of change is likely to be constant in time. A trend emerges.

In most systems, whether disturbed or undisturbed, the rate of change will vary. In succession one can expect phases in which change is fast and others in which it is slow. The same holds for short-time fluctuation as well. Probably the best way to recognize multivariate trends is to display the states in phase space; that is, to ordinate the relevés. If the data originate from one plot only, trends manifest in a single sequence of points in which the order of points accords with time. If this forms a smooth line then there is *temporal dependence* occurring, since each state evolves from the previous, causing minor changes to be likely to happen. To decide whether this is a local phenomenon or a spatially relevant, a spatial sample is needed, as shown in Figure 9.3. This is a subset of the data used by Wildi & Schütz (2000) in which relevés document succession in 8 (out of 59) plots located in the Swiss National Park. The time steps are all adjusted to five-year interval and the investigation period ranges from 1917 until about 1996. The individual series are overlapping, forming a long, horseshoe-shaped temporal gradient. From the symbols one can even see that the whole successional gradient has a range of a great many time steps (approximately 80), corresponding to about 400 years. Different types of succession can be distinguished. Plot Tr6 shows a perfect trend in one main direction: the rate of change is almost constant. Plot Tr5 also fits into the series, but only after the first four time steps; before that there is a different process underway. Pin4, finally, fits perfectly into the temporal gradient but almost no change can be found.

Once we know that plots exhibit a common trend, we can inspect the distance matrices, as explained in Figure 9.2. Taking plot Tr6, there are 16 states available from 1921 until 1994. The resulting distance matrix is shown in Figure 9.4, left side. The distances increase monotonely as the order of rate of change increases – when moving away from the diagonal.

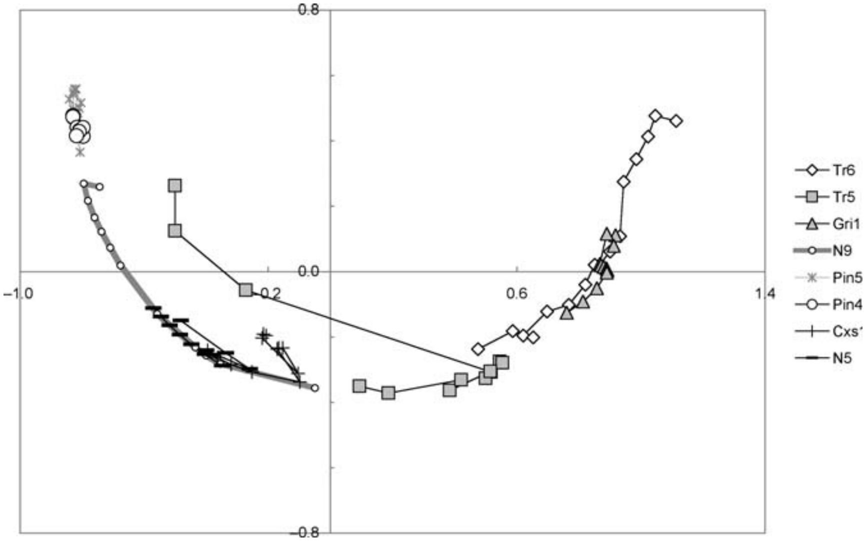


Figure 9.3 Ordination of data from eight plots in the Swiss National Park (Wildi & Schütz 2000). Succession proceeds from right to left.

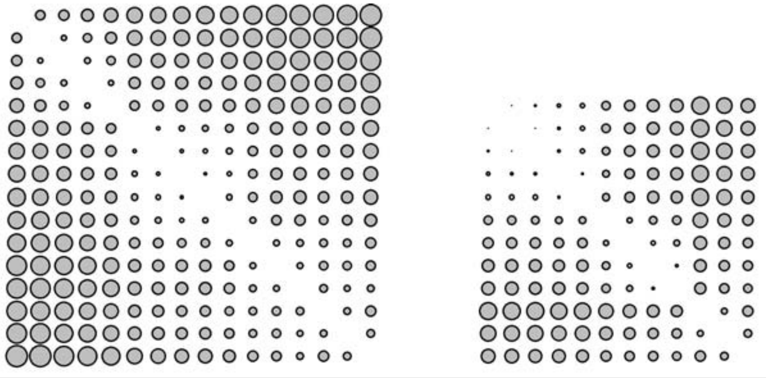


Figure 9.4 Rate of change in plots Tr6 (left, 16 time steps) and Pin4 (right, 12 time steps) in the Swiss National Park (see Figure 9.3). Explanations are given in Figure 9.2.

The pattern confirms that the trend occurs at all temporal scales, at short and also at long time intervals. The right-hand side shows the distance matrix of the plot named Pin4. The series consists of 12 time steps starting in 1940 and ending in 1996. The distance matrix suggests that there is a faint trend in the first nine time steps, but then the average distances start shrinking again. When inspecting the respective plot in Figure 9.3 it can be seen that the trend still perfectly fits the overall successional sequence, but it does not evolve any further and may have reached some equilibrium stage.

9.3 Markov models

In this section a model process capable of reproducing simple cases of multivariate patterns of change is presented. Very much like linear regression, it is elementary by nature in frequently fitting observed patterns locally. It abstracts from noise, complexity and nonlinearity and therefore fails to fit patterns when the rules in systems change, such as the competitional hierarchy of species. However, the process remains fundamental as its application may frequently be the first step in the evaluation of temporal patterns. The functioning and use of Markov models for vegetation surveys is explained below (see also Wildi 2001).

Changes in permanent plots can be interpreted as replacement processes. Several plant populations occupy the same resource. In the context of vegetation ecology, the primary resource is frequently the physical space. Because space can be defined as a fundamental resource occupied by plants, cover percentage is a logical (even though it is two-dimensional only) surrogate for measuring resource consumption. If gains and losses in space are in balance, so that any state of the system can be derived from the previous one, then screening for a Markov process is worthwhile (Usher 1981).

Win and loss of every species is defined in a transition matrix P. This allows derivation of the state of the relevés x at time t from the preceding step:

$$x_{t+1} = x_t P \tag{9.1}$$

In an ordinary permanent plot survey the observation vectors x are vegetation relevés. Unfortunately, this observation does not allow measurement of the elements of the transition matrix: the wins and losses of the species (and this is the main reason why Markov models are not used routinely): if a species wins space then we do not know which other species has lost it, and if a species loses ground, we do not know which one will profit from it. A Markov process, if present, remains undetected. Two plausible assumptions may help to overcome this situation in a method devised by Orlóci *et al.* (1993):

1 If a species loses part of the main resource then any other dominating species will most likely profit (i.e. profit is proportional to the species cover).

2 If a species increases its partition of the resource then the remaining dominant species will lose in proportion to its cover.

Both of these assumptions can be questioned. In succession, a change in abundance of a species may be caused by colonization of space by a new species, but a Markov model in its basic form does not foresee invasion. Similarly, a colonizing species may expand its cover at the expense of rare species. This is one indication why Markov chains cannot be applied in isolation when invasion occurs.

The resource space is not always entirely occupied by the vegetation. For this reason it may be important to add one more variable to the species list: quantifying the *open soil* (see Figure 9.5 for an example). Formally, this functions like any ordinary species. It is important that the sum of all cover values, including open space, exactly amounts to 100%. This is achieved by the appropriate transformation of the initial state:

$$x'_{i,t} = \left(\frac{x_{i,t}}{\sum_{i=1}^{n} x_{i,t}} \right) * 100 \tag{9.2}$$

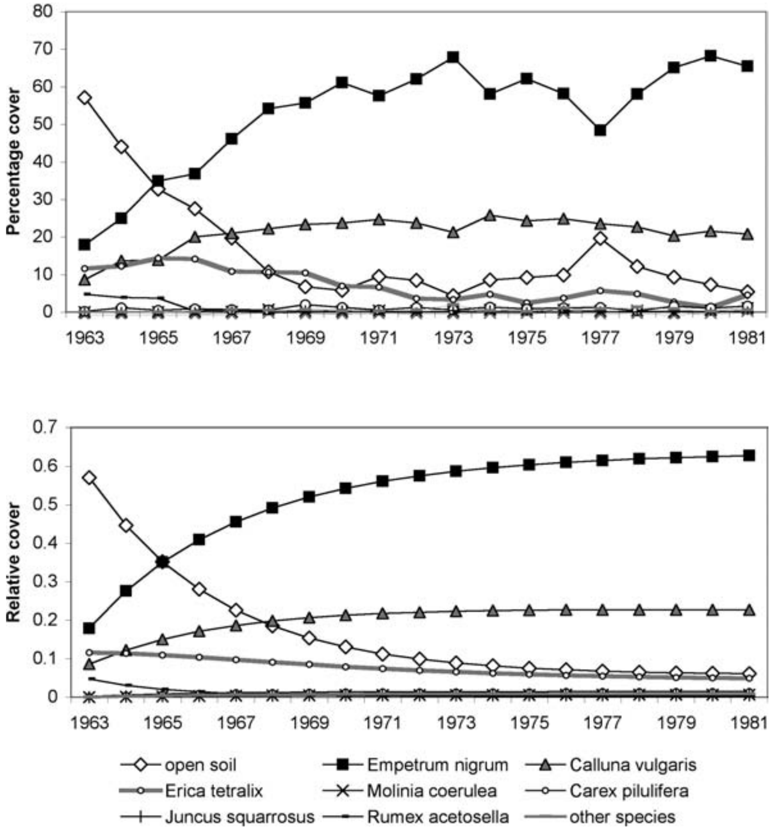


Figure 9.5 A Markov model of the Lippe *et al.* (1985) data set. Upper graph: field data. Lower graph: simulated data.

In this equation, vector x contains the cover values of species i, t is the present time and n is the number of species. In the artificial example used for further explanations (Table 9.1), the sum of cover values is already 100%.

First the transition matrix for time step 1 to time step 2 is calculated (Orlóci *et al.* 1993). For each species i a difference results, expressing change in time:

$$Diff(i) = x_{i,2} - x_{i,1} \tag{9.3}$$

Positive values of $Diff(i)$ signify a gain, negative ones a loss. The transition matrix contains all the losses of species i in row i, and the gains of the same

Table 9.1 Numerical example for demonstrating a Markov process (raw data).

	$x_{t=1}$	$x_{t=2}$	$x_{t=3}$
Species 1	60	40	10
Species 2	25	35	55
Species 3	15	25	35
$\sum$	100	100	100

in column i:

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & \downarrow & \cdots \\ \cdots & \cdots & \downarrow & \cdots \\ \rightarrow & \rightarrow & p_{ii} & \rightarrow \\ \cdots & \cdots & \downarrow & \cdots \end{pmatrix} \tag{9.4}$$

where ↓ is gain and → is loss of species i. The diagonal elements contain the proportions each species covers at the end of the time step: $x_{i,t+1}$. The gains of species i at the expense of species h, as well as the losses of species i from species h, are given by the equation:

$$Dev(h, i) = |Diff(i)| \frac{x_{h,t+1}}{\sum_i x_{i,t+1}} \tag{9.5}$$

This means that gains and losses occur in proportion to the resource (i.e. the space) each species occupies at the end of the actual time step. When processing the first species in our example, it can be seen that it loses 20% of the total ground from $t = 1$ to $t = 2$ ($Diff(1) = -20$). The new diagonal element is 40. The loss of the second element, a portion of 20% of the 35% cover, is 7%. The third element is 20% of the 25% cover -5% – completing the first row:

$$\mathbf{P}(t1; t2; spec.1) = \begin{bmatrix} 40 & 7 & 5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Species 2 exhibits a win of 10% (a factor of 0.1) to be added in column 2. This is again proportional to the covers of the species at time $t + 1$:

$$\mathbf{P}(t1; t2; spec.1 + 2) = \begin{bmatrix} 40 & 7 + 4 & 5 \\ 0 & 35 & 0 \\ 0 & 2.5 & 0 \end{bmatrix}$$

The procedure is completed by applying it to species 3:

$$\mathbf{P}(t1;t2) = \begin{bmatrix} 40 & 7+4 & 5+4 \\ 0 & 35 & 3.5 \\ 0 & 2.5 & 25 \end{bmatrix}$$

After normalizing the rows (the sum adjusted to 1) the transition matrix is:

$$\mathbf{P}'(t1;t2) = \begin{bmatrix} 0.667 & 0.183 & 0.150 \\ 0 & 0.909 & 0.091 \\ 0 & 0.091 & 0.909 \end{bmatrix}$$

For each following time step the procedure is resumed to yield one more transition matrix, as for time steps $t = 2$ to $t = 3$, where it is:

$$\mathbf{P}(t2;t3) = \begin{bmatrix} 10 & 10.5 & 11.5 \\ 0 & 55 & 5.5 \\ 0 & 7.0 & 35 \end{bmatrix}$$

For all time steps, all transition matrices, P, are averaged. The new transition matrix, $\overline{P}$, is assumed to hold for the entire time series. This also means that it is kept constant over time and it will therefore outbalance fluctuations. In our example, after normalizing by rows, we get:

$$\overline{\mathbf{P}} = \begin{bmatrix} 0.500 & 0.2950 & 0.2050 \\ 0 & 0.9091 & 0.0909 \\ 0 & 0.1367 & 0.8633 \end{bmatrix}$$

Through simple matrix multiplication according to Formula 9.1 the simulated relevés are derived (Table 9.2).

Here the first relevé is identical to the field data (Table 9.1). It represents the initial state of the dynamic system; all subsequent states are merely

Table 9.2 Numerical example demonstrating a Markov process (simulated data).

	$x_{t=1}$	$x_{t=2}$	$x_{t=3}$
Species 1	60	30	15
Species 2	25	42.5	51.23
Species 3	15	27.52	33.77

approximations. This becomes obvious from the examples below. Orlóci *et al.* (1993) published a time series documenting recovery of a heathland after fire, using data from an investigation by Lippe *et al.* (1985). The example is presented here as a case where a linear Markov process successfully reproduces the temporal pattern found.

The raw and the simulated data ('Lipperaw' and 'Lippesim' in Appendix B) are shown in Figure 9.5. From the upper graph it can be seen that in the first few years there is a directed change. After about eight years (≈ 1970), an equilibrium state is reached in which merely random oscillation occurs. One objective of the analysis is to determine the equilibrium state for which the Markov model is derived. The transition matrix is calculated as shown above; that is, from the 19 states of the system. It is the mean of 18 matrices calculated for each time step. Then, beginning with the first field observation, 18 Markov relevés are derived. They are shown in the lower graph in Figure 9.5. After 19 years, the model has almost reached an equilibrium state. In the present example, the model fits the field data almost perfectly. However, only the deterministic part of the variation is reflected by the simulated time series; the temporal fluctuation is completely suppressed.

The Markov model perfectly explains multispecies change as long as this change is linear. But why is it called linear? The response curves are not linear, but curved and monotone. However, the similarity pattern is linear and when the data is ordinated it can be seen that the time trajectory is an almost straight line (Figure 9.6, upper graph) and the Markov model

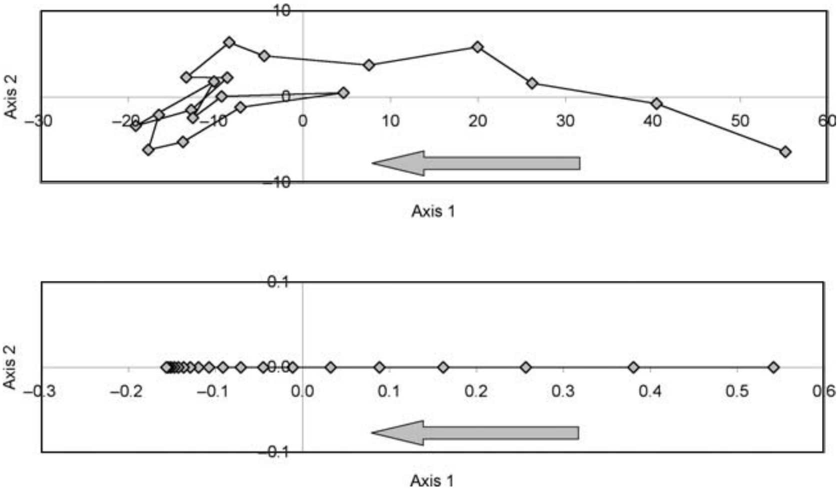


Figure 9.6 PCOA ordination of the Lippe succession data. Upper graph: field data used. Lower graph: Markov data used. Arrow pointing in direction of time.

generates a perfectly linear pattern (Figure 9.6, lower graph). This illustrates that a linear process yields a linear pattern, even if the underlying species response curves are bent!

But when does the linear Markov model fail to reproduce succession? This is simply the case if the ‘rules of the game’ change with time. In terms of the vegetation process, it takes place when the relative competitional power of species changes. Ordinations will reveal such situations in presenting time series as horseshoe-shaped trajectories (see Figure 9.16 for an example). While the data in Figure 9.5 shows a successful application, I also add one demonstrating a failure in Figure 9.7. This is the successional series from the Swiss National Park, presented in more detail in Figure 9.13. There is a fairly good match during the first few of the 81 time steps but the shape of

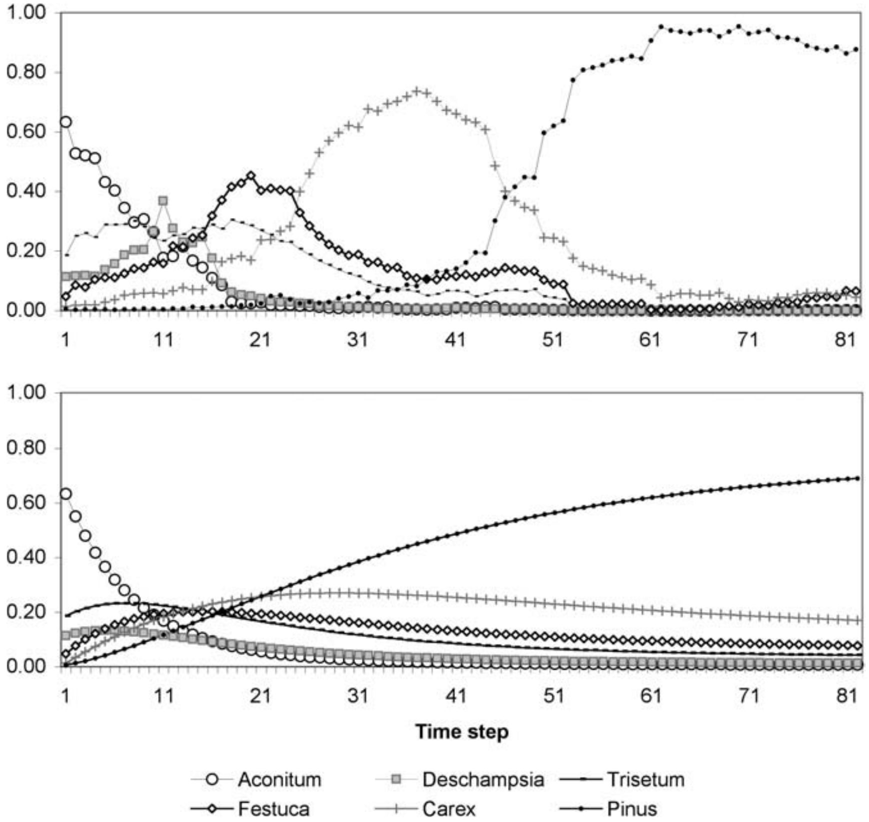


Figure 9.7 A Markov model (lower graph) of the time series of the Swiss National Park (upper graph), taken from Figure 9.13. y-axis shows relative cover.

the real and simulated curves start deviating fundamentally as soon as new species invade and the equilibrium state in Figure 9.7 is far from reality.

A Markov chain of the type shown above simulates a classical succession towards a monoclimax: the vegetation reaches an equilibrium state inherent in the model, from which it does not escape. In systems theory, the final state is called a point attractor. In other systems, however, it can happen that a cycle is reached, as proposed by Watt (1947) (meaning that the system has a cyclic attractor), or there can be apparent random fluctuations like in the data of Lippe *et al.* (1985).

9.4 Space-for-time substitution

9.4.1 Principle and method

In long-term investigations species eventually exhibit a characteristic pattern of change: constancy, increase, decrease, random fluctuation, periodicity and so on (Huismann *et al.* 1993). If the observation time is sufficiently long, many of these patterns turn out to be fragments of a bell-shaped response curve. In space-for-time substitution, one assumes that several different fragments of the same response curve can be found, but occurring in different plots. If these fragments overlap, the entire response curve can be restored. This principle is sketched schematically in Figure 9.8. The heavy lines show hypothetical response curves of the same species over eight time steps. At first glance they seem to be different in nature – sometimes with a tendency

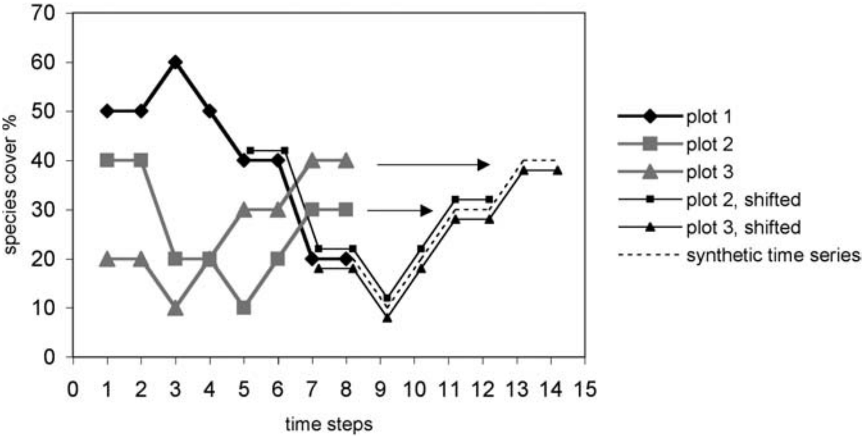


Figure 9.8 The principle of space-for-time substitution in the univariate case (Ghosh & Wildi 2007).

to increase and sometimes decreasing – however, when the curves of plots 2 and 3 are properly shifted (light lines), a single response evolves covering 14 time steps. This is of course just an interpretation, because such curves never fit perfectly. It is therefore essential that they overlap sufficiently, as is the case in Figure 9.8. In real situations, there are many species involved, and an overlap should yield a meaningful result for all of them.

The fact that a successional trend ends within a plot but continues in another plot has been known for long time. It has added much to the monoclimax theory strongly debated in the first half of the twentieth century (Clements 1916, see also Maarel 2005). Although there has never been a doubt that such phenomena exist, the handling and interpretation has led to controversy. Much of this can be read in the review by Pickett (1989): he warns against the uncritical use of such data. Although I give an example of a really successful application below, some of the most frequently mentioned shortcomings and pitfalls are listed first:

- Superimposing time series data is always hampered by statistical noise and disturbance. The resulting synthetic series therefore suffers from some uncertainty.

- Vegetation change within two different plots will never be identical as the plots will most likely differ in site condition as well as in the species pool.

- Succession may not proceed at the same speed in different plots, prohibiting a perfect fit of series.

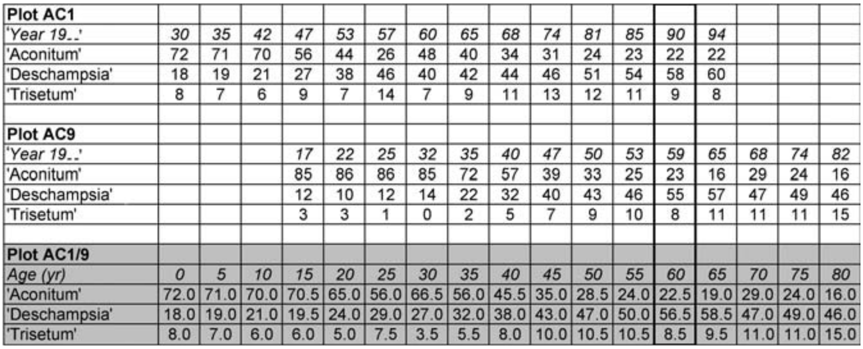
- The overlay of series mimics monoclimax. Alternative paths as in polyclimax can hardly be identified by this approach alone.

There are so far no experiments known to me that would test the potentials and failures of space-for-time substitution, simply because these would last too long. All applications known are from surveys. Often, it is not even envisaged that long-term temporal patterns will be sought, just the data suggested in the course of the analysis. This was the case in the succession data from the Swiss National Park shown below. The method has been developed specifically to screen the more than a hundred time series available for a general temporal trend (Wildi & Schütz 2000). The aim of the method is to find an unequivocal solution to the superposition of several (i.e. more than two!) multivariate time series rather than having to search for an iterative result by trial and error.

The problem with finding the best solution when many time series are available is in identifying the most suitable pairs of response curves for fusion. The best candidates are those where the overlapping observations are the most similar compared to all other time series. In the following the similarity of time series is defined as the similarity of the two most similar observations in any two time series. This is shown in Figure 9.9, an example from two plots in the Swiss National Park. The species set is reduced to three for simplicity. The ‘real’ plots, AC1 and AC9, were surveyed in ‘real’ years: AC1 from 1930 until 1994 and AC9 from 1917 until 1982. When comparing the species composition (using Euclidean distance) the most similar observations are those from 1990 in AC1 and those from 1959 in AC9. The series are now shifted until these two observations are located in the same column. The new series AC1/9 covers 80 years: now just in the sense of age and without any specific dates. It can also be seen that the method requires time steps of the same length. In the present case minor deviations from identical time steps were corrected by interpolations. The steps leading to an unequivocal solution when fusing three or more time series are the following:

- Compute a resemblance matrix of time series. Resemblance (distance) is defined as shown in Figure 9.9.

- Derive the minimum spanning tree of time series (Gower & Ross 1969). This is a graph showing the nearest neighbours of all time series in the form of a tree.



Plot AC1																	
'Year 19..'	*30*	*35*	*42*	*47*	*53*	*57*	*60*	*65*	*68*	*74*	*81*	*85*	*90*	*94*			
'Aconitum'	72	71	70	56	44	26	48	40	34	31	24	23	22	22			
'Deschampsia'	18	19	21	27	38	46	40	42	44	46	51	54	58	60			
'Trisetum'	8	7	6	9	7	14	7	9	11	13	12	11	9	8			
Plot AC9																	
'Year 19..'				*17*	*22*	*25*	*32*	*35*	*40*	*47*	*50*	*53*	*59*	*65*	*68*	*74*	*82*
'Aconitum'				85	86	86	85	72	57	39	33	25	23	16	29	24	16
'Deschampsia'				12	10	12	14	22	32	40	43	46	55	57	47	49	46
'Trisetum'				3	3	1	0	2	5	7	9	10	8	11	11	11	15
Plot AC1/9																	
Age (yr)	*0*	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*40*	*45*	*50*	*55*	*60*	*65*	*70*	*75*	*80*
'Aconitum'	72.0	71.0	70.0	70.5	65.0	56.0	66.5	56.0	45.5	35.0	28.5	24.0	22.5	19.0	29.0	24.0	16.0
'Deschampsia'	18.0	19.0	21.0	19.5	24.0	29.0	27.0	32.0	38.0	43.0	47.0	50.0	56.5	58.5	47.0	49.0	46.0
'Trisetum'	8.0	7.0	6.0	6.0	5.0	7.5	3.5	5.5	8.0	10.0	10.5	10.5	8.5	9.5	11.0	11.0	15.0

Figure 9.9 The similarity of time series. Plots AC1 and AC9 are located in the Swiss National Park.

- Position the observations by overlapping the time series according to the order given in the minimum spanning tree (Figure 9.11). This yields the relative age of each series (Figure 9.12).

- Compute the average composition of the new synthetic time series by averaging all scores pertaining to the same time step.

The minimum spanning tree yields a unique solution to the problem. This is not necessarily the 'true' one, but it is the one delivering the shortest possible series based on the data used.

9.4.2 The Swiss National Park succession (example)

The results shown below are from Wildi & Schütz (2000). The original time series are of unusual length and the first observations date back to the year 1917, when J. Braun-Blanquet established the first permanent plots in the Park with the aim of documenting reforestation of pastures (Figure 9.10).

It took as much as 80 years to detect that the data could be interpreted according to the idea of space-for-time substitution. The plots do not constitute a statistical sampling design but they are dispersed all over the previous pastures in the park. The data set used below includes 59 of them, consisting of 751 relevés (data set 'Snpser59', Appendix B). The species are summarized into six groups, carrying the names of 'dominats'.

Figure 9.10 *Pinus mugo* on a former pasture in the Swiss National Park, just escaping the reach of browsing red deer.

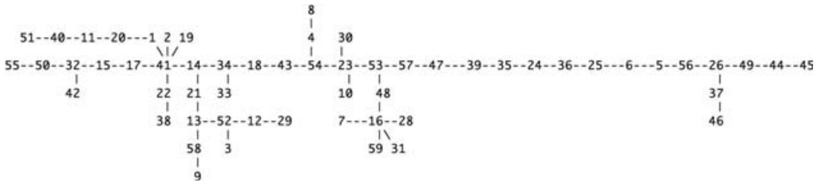


Figure 9.11 Minimum spanning tree of 59 time series from the Swiss National Park.

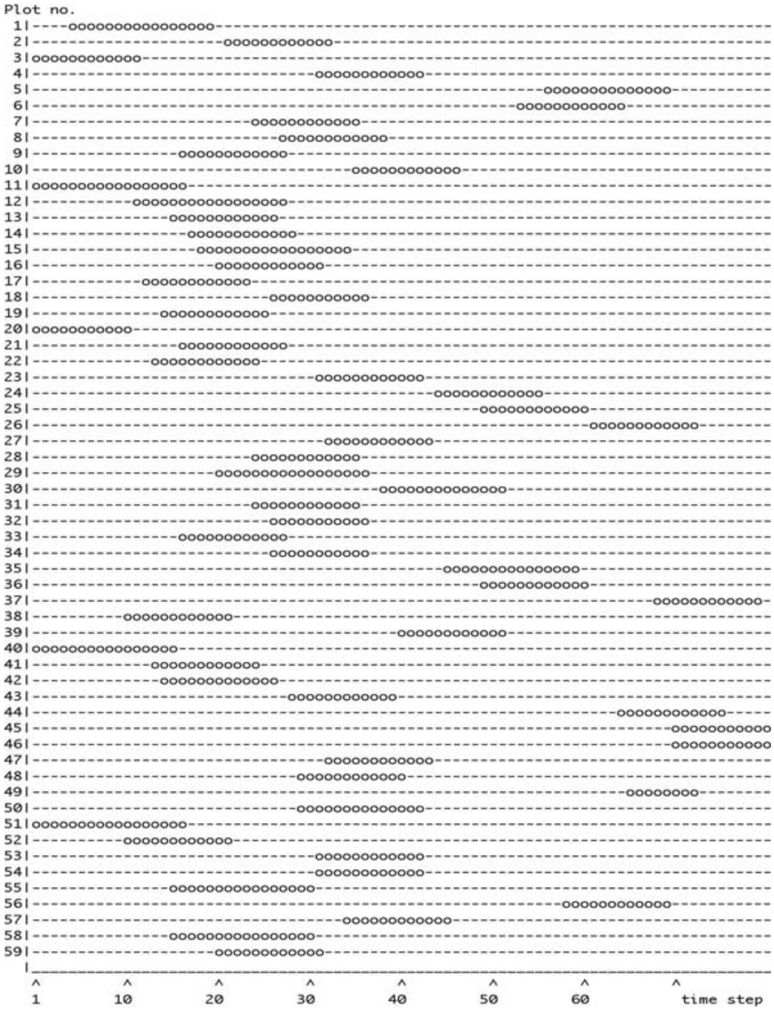


Figure 9.12 Ordering of 59 time series from the Swiss National Park.

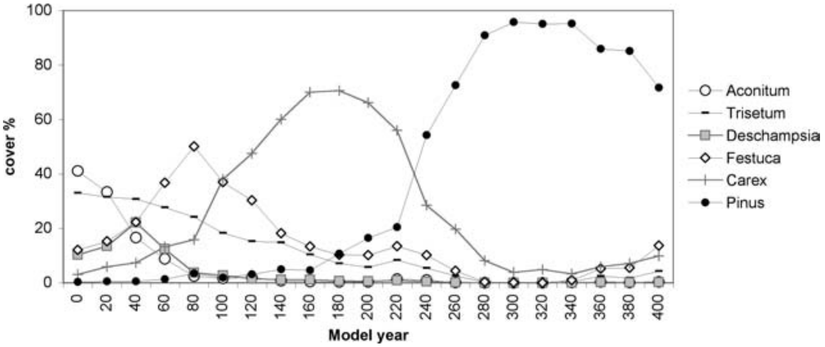


Figure 9.13 Succession in abandoned pastures of the Swiss National Park, derived by space-for-time substitution. Every fourth time step is shown.

Data handling is explained in more detail in Wildi & Schütz (2000). The steps involved in the analysis are the same as those shown in Section 9.4.1. In Figure 9.11 the minimum spanning tree for the 59 time series is shown. This is not just a single line as an ordering principle but a more complex tree. Processing this by fusing time series pairwise yields the arrangement in Figure 9.12. The resulting synthetic time steps (81) have to be multiplied by step length of 5 years, yielding a model time span of 405 years (Figure 9.13).

The overall trend can be interpreted as follows: an initial *Aconitum* phase, resulting from livestock grazing and fertilization, dominates for about 50 years after the cessation of grazing by livestock. A *Deschampsia* phase then emerges and is dominant for about 15 years. A later transition to a grassland dominated by *Festuca rubra* is most likely caused by grazing activity by red deer (Achermann *et al*. 2000). This is followed by a *Carex sempervirens* phase that may last 150 years. Finally, *Pinus montana* seedlings begin to establish, initiating the reforestation phase.

The pattern revealed in this example must be strongly nonlinear, as it has been shown in Section 9.3 that a linear Markov model fails to explain it. It has been shown by Wildi & Schütz (2007) that computed process length, a result of analysis, varies to some extent depending on the transformation chosen for the species scores. Unlike in Section 7.2.3 no reference measurement exists to tell us which of the estimations is best, leaving us with some uncertainty about succession velocity.

9.5 Dynamics in pollen diagrams (example)

Vegetation series encompassing time spans of thousands of years can only be found in fossil records, for example in pollen diagrams (Lischke 2005).

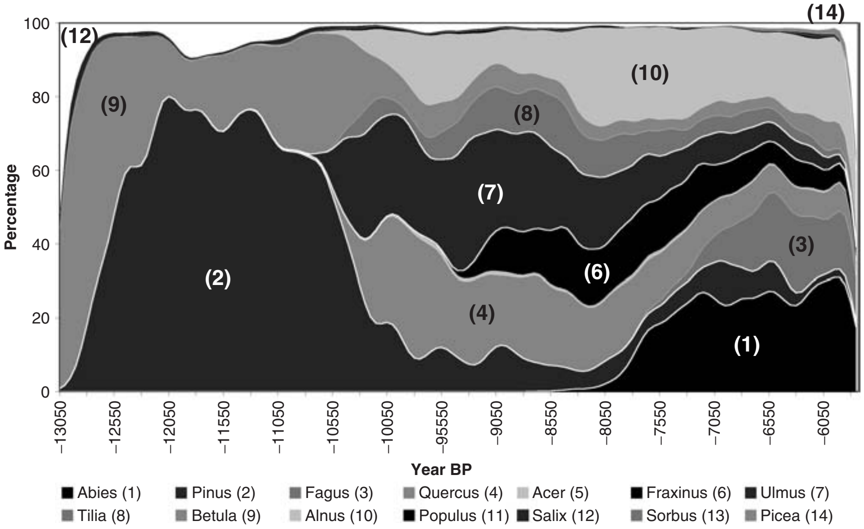


Figure 9.14 Tree species in the pollen diagram from Soppensee (Lotter 1999).

Here extreme nonlinearity can be expected because very long time spans increase the chance that changes in the functional role of species within the vegetation cover will occur, caused mainly by invasions and extinctions. I demonstrate typical patterns using pollen of tree species of the Soppensee profile from Lotter (1999) (see also Lischke *et al.* 2002), documenting the change in tree species on the Swiss Plateau from about 13000 BP until 5700 BP (Figure 9.14, without considering the changes in the technology of 14C dating that have taken place in the meantime).

First we look at the velocity of process, defined as the rate of change in total species composition per time unit:

$$V = \frac{d}{\delta t} \tag{9.6}$$

where d is the Euclidean distance (a measure of dissimilarity) between any two consecutive states in the pollen diagram. This type of calculation allows the derivation of a velocity profile of the change processes over the period of measurement (Figure 9.15). There are different time steps used for this, the shortest of 50 years being given by the temporal resolution of data, whereas a step of 800 years shows the long-term trends. The 50-year time step is also used to interpret the ordination in Figure 9.16. The states in time are the circles and their diameter is proportional to velocity. There are linear phases where velocity is high and others where it is low. Velocity seems to

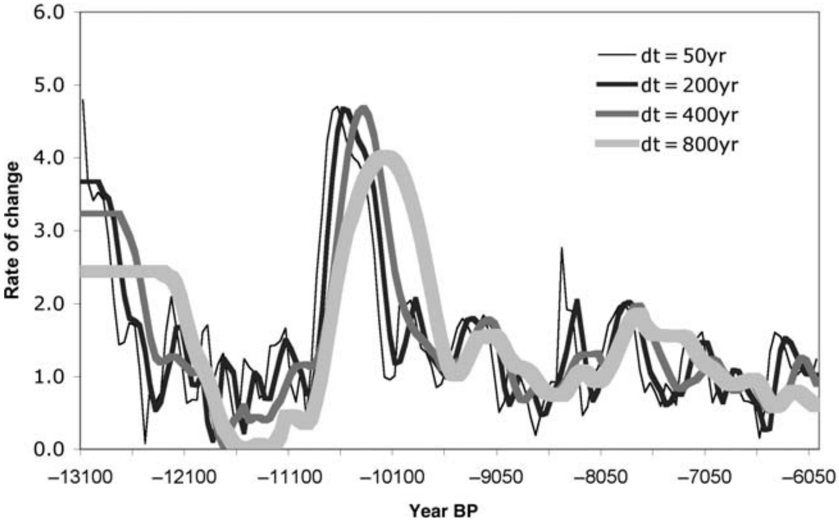


Figure 9.15 Velocity profile of the Soppensee pollen diagram. The time-step length in the data is 50 years.

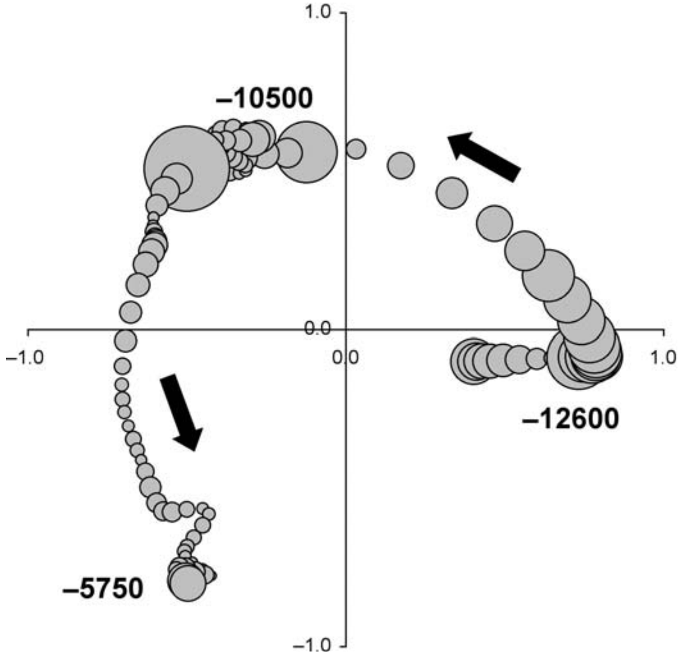


Figure 9.16 Time trajectory of the Soppensee pollen diagram. The diameter of the circles is proportional to the velocity of change.

fluctuate considerably during periods of nonlinearity, when large and small circles alternate.

It is also worth distinguishing the qualitative and the quantitative components in the data. In the analyses shown so far the quantitative view is adopted; that is, the species scores are percentages of total pollen abundance. In these analyses the most abundant types of pollen dominate the result. Rare species, including the invaders in their first years of arrival, do not contribute much to the rate of change. However, species scores can be transformed to presence–absence (or any intermediate scale, see Section 7.2.3) so that the velocity expresses change in species presence. This is shown in Figure 9.17, where the lower lines express change in the quantity of pollen composition while the upper lines are change in quality; that is, the emergence or disappearance of pollen of a specific species. Obviously, these two types of change happen at different points in time. There are phases in which several species emerge (upper lines) but no considerable change in quantity can be observed (lower lines) and vice versa. Numbers 1 through 6 indicate discrete events:

1 Invasion of *Alnus*, followed by two invasions of *Quercus*.

2 Invasion of *Acer*, *Fraxinus*, *Tilia* and *Ulmus*.

3 *Fraxinus* disappearing.

4 *Fraxinus* returning.

5 *Abies* invading.

6 *Fagus* and *Picea* arriving.

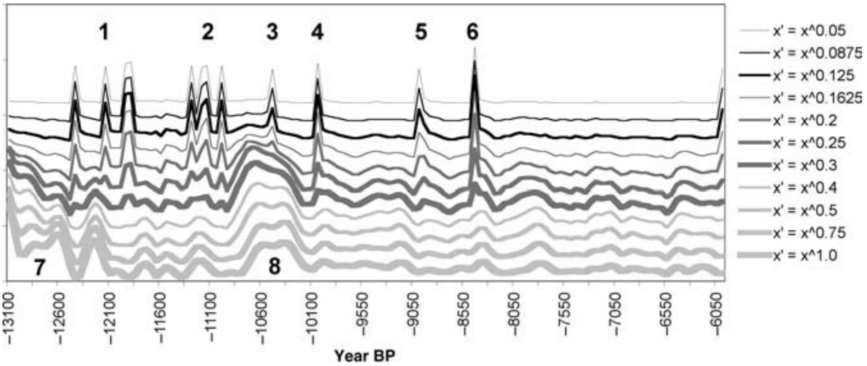


Figure 9.17 Velocity profiles from quantitative (bottom line) towards qualitative (top line). The peaks in the upper lines stem from species invasions.

Numbers 7 and 8 indicate increased velocities in mass changes:

7 Mass expansion of *Pinus*.

8 Mass retreat of *Pinus*.

Even when inspecting the velocity profiles in detail, no evidence is found for velocity being related to nonlinearity, such as that around 12 600 and 10 500 BP. But there is yet another interesting way of looking at similarity, when taking its second derivative, acceleration in the change process, *A*:

$$A = \frac{V_{t+\delta t} - V_t}{\delta t} \tag{9.7}$$

When *A* is positive, the velocity increases; when it is negative, the velocity decreases. In Figure 9.18 acceleration is used to interpret the same time trajectory as in Figure 9.16: circles are proportional to acceleration. The result is striking: linear phases, whether fast or slow, show very low acceleration. Nonlinear, on the other hand, are characterized by huge positive as well as negative accelerations. Hence, strong fluctuations in the dynamics of the systems distinguish nonlinear from smooth, predictable linear phases.

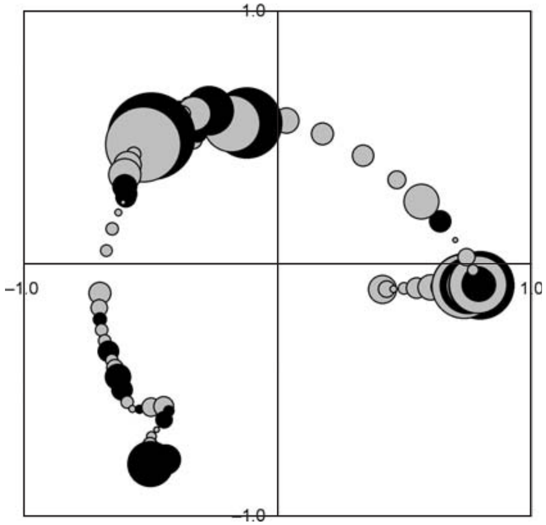


Figure 9.18 Time trajectory of the Soppensee pollen diagram. The diameter of the circles is proportional to the acceleration of change. Grey are positive, black negative values.

Is this finding a generally valid rule? No, it is not. First of all, the observation is restricted to one profile only and in this there are just two phases of strong nonlinearity. Second, fluctuation manifesting in speed and acceleration can be caused by disturbance of the profile. Sedimentation is dependent on many factors and cannot be expected constant over thousands of years. Hence, investigation of many more profiles would be needed to find evidence for validity of the rule.

10
Dynamic modelling

It may come as a surprise to see dynamic modelling in a book on vegetation ecology. Probably the first dynamic models of the type used here served the investigation of systems other than ecological, mainly economic and industrial. Forrester (1968), in his pioneering book *Principles of Systems*, gives a very simple definition of the subject: ‘As used here, a “system” means a grouping of parts that operate together for a common purpose.’ In models of such systems, ‘parts’ are described by state variables such as weight of plant biomass, percentage cover of vegetation, plant nutrients per cubic decimetre of soil, population size of a species in a plot and so on. Hence, dynamic models perfectly serve the analysis of natural systems. But what is the meaning of the buzzword ‘model’? Again, Forrester (1968) gives a simple explanation: ‘A model is a substitute for an object or a system.’ When modelling we are working with this substitute, being a system by itself, just like the real system it describes. When modelling we investigate the substitute by, for

Data Analysis in Vegetation Ecology Otto Wildi

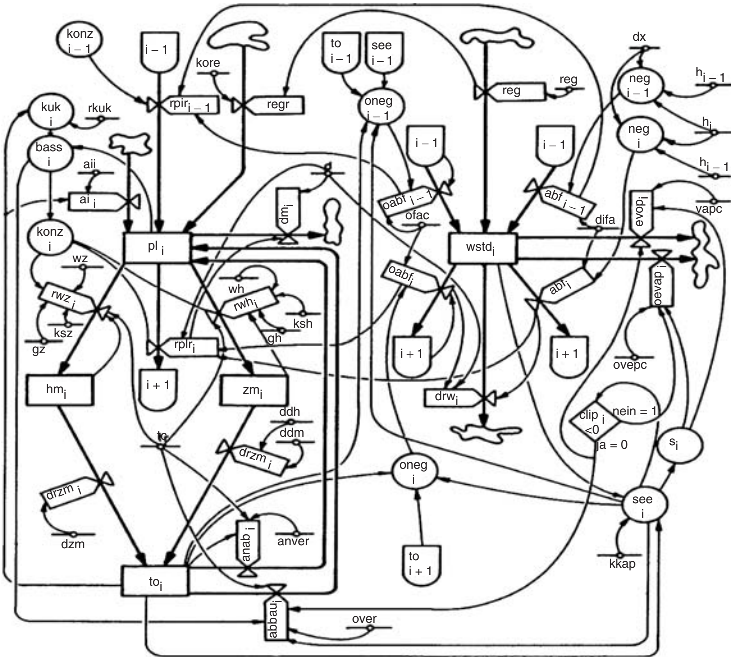


Figure 10.1 Attempt to get a dynamic model under control (Wildi 1976).

example, performing test runs and studying how it succeeds or fails without doing harm to the real system – not even damaging the computer we use. In the early days of electronic computing everyone was fascinated by the apparently unlimited possibilities of simulation, culminating in the world model of Dennis L. Meadows, by which he justified his *Limits to Growth* (Meadows *et al.* 1972). Limits to modelling were experienced later because the computer models proved difficult to handle when complexity increased, as illustrated in an early attempt shown in Figure 10.1. Even simple models may be difficult to handle and to understand; small is often beautiful.

Dynamic modelling became popular because it is easy to understand and easy to do – even for the mathematically less well trained. The rules by which state variables change are described by one or several differential equations. Based on initial conditions given by the modeller the resulting change of the entire system in time is derived through numerical integration, all carried out by the computer. When introducing the method I start with the simplest temporal systems, comprising one state variable only, and subsequently add interacting variables, then finally extend the principle to spatial systems.

10.1 Simulating time processes

Time has only one dimension and the simplest type of model does not consider space. The temporal change of a system is described in the form of differential equations. Numerical integration of these equations yields a state vector describing the state of the system at any one point in time. An example is the exponential growth equation:

$$\frac{\delta X}{\delta t} = rX \tag{10.1}$$

where X (the state variable) is, for example, the number of individuals at time t, and r the is rate of population increase (Maynard Smith 1974). This equation has the well-known deterministic solution:

$$X = X_0 e^{at} \tag{10.2}$$

This assumes that within a short time span, δt, each individual will give rise to a fraction, $r\delta t$, of new individuals. It is of course more adequate to reason in stochastic terms and so $r\delta t$ is the probability of an individual having an offspring within δt, or no offspring with the probability of $1 - r\delta t$. As shown by Maynard Smith (1974), the formula for the average population size is:

$$\hat{X} = X_0 e^{at} \tag{10.3}$$

and the variance of X is:

$$var(X) = X_0 e^{2at}(1 - e^{-at}) \tag{10.4}$$

Integrating more complex differential equations is difficult, and frequently impossible. This is where numerical integration comes into play: the principle shown in Figure 10.2. In numerical integration, one specific outcome is calculated based on one assumed initial state of the system. The example starts with a population size of $X_{t=0} = 10$. The growth rate is $r = 0.1$, meaning that any given individual has a probability of having an offspring within a single time step Δt of 0.1. For each time step, the new state of the population size is determined based on the previous state:

$$X_{t=1} = X_{t=0} + \frac{\Delta X}{\Delta t} \tag{10.5}$$

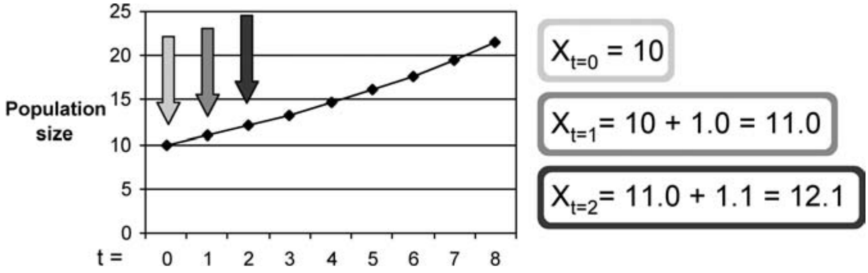


Figure 10.2 Numerical integration of the exponential growth equation (Formula 10.1) using Euler's rule.

Table 10.1 Approximating integration (column 1) by numerical integration (columns 2 and 3).

Time	$\Delta t \to 0.$	$\Delta t = 0.5$	$\Delta t = 1.0$
$X_{t=1}$	11.0517	11.025	11.000
$X_{t=2}$	12.214	12.155	12.100
$X_{t=3}$	13.498	13.401	13.310
$X_{t=4}$	14.918	14.774	14.641
$X_{t=5}$	16.487	16.289	16.105

This kind of calculation is also known as Euler's rule. After one time step, $X = 11.0$, and after two steps we get $X = 12.1$. In numerical integration the symbol Δ is used instead of δ because the time step has a finite length, chosen by the user. This is of course just an approximation to the real process, which is continuous and not discrete. An ideal population of large size will grow from the very beginning of the process, leading to a somewhat faster growth than in a discrete case. In numerical simulation the approximation is improved when recalculating X in smaller time steps. The effect is shown in Table 10.1. Obviously, precision increases with reduced time-step length for the calculation. Setting the time step too short, however, may lead to computational errors due to the limited precision current computers provide.

In the simulation of ecological systems there are usually several state variables involved. To illustrate this I develop a simple artificial model describing, for example, the overgrowth of an open water pond by two competing, floating plant species, $X1$ and $X2$. Hence, $X1$ and $X2$ are the state variables. All parameters composing the system are:

$X1$	size of population $X1$
$X2$	size of population $X2$
$r1$	growth rate of $X1$
$r2$	growth rate of $X2$
C	carrying capacity of the system

Because the pond has limited surface I also introduce C, the carrying capacity. This is the total number of individuals of any kind the open water surface can hold. The carrying capacity acts as a limiting factor in the logistic growth equation shown below:

$$\frac{\delta X1}{\delta t} = rX1\frac{C - X1}{C}. \tag{10.6}$$

As the state variable $X1$ approaches the carrying capacity, C, the growth becomes zero and the population stops growing. In the following example, I extend this model to two equations, describing the growth of two interacting populations, $X1$ and $X2$. Both follow the rules of the logistic growth, but they rely on the same carrying capacity, C:

$$\frac{\delta X1}{\delta t} = rX1\frac{C - X1 - X2}{C} \tag{10.7}$$

$$\frac{\delta X2}{\delta t} = rX2\frac{C - X1 - X2}{C} \tag{10.8}$$

The two populations shall differ in size and growth rate. For a simulation, the following initial conditions are assumed:

$$\begin{aligned} X1_{t=0} &= 10 \\ X2_{t=0} &= 50 \\ C &= 200 \\ r1 &= 0.25 \\ r2 &= 0.30 \end{aligned}$$

Using Euler's rule, the result of a simulation run over 50 years is shown in Figure 10.3 (model 1). The two populations co-exist and reach an equilibrium when the carrying capacity is exhausted. This takes about

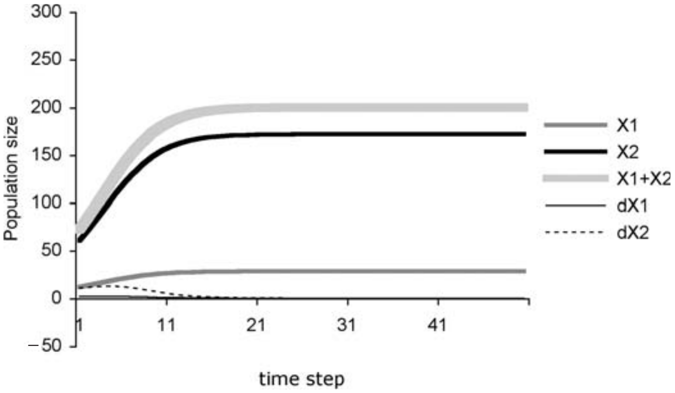


Figure 10.3 Logistic growth of two populations; model 1.

20 time steps. The final population size for $X1$ is 28.25 and for $X2$ is 171.74. The sum, $X1 + X2$, reaches the carrying capacity as stated in the model – within minor rounding errors.

So far, no serious complications have been observed in this model. To demonstrate critical issues I now change the model slightly. In model 2 the populations grow differently. $X1$ now is no longer affected by $X2$, and it will therefore outcompete the latter. $X2$, on the other hand, is still limited by itself and $X1$. Hence, we get:

$$\frac{\delta X1}{\delta t} = rX1\frac{C - X1}{C} \tag{10.9}$$

$$\frac{\delta X2}{\delta t} = rX2\frac{C - X1 - X2}{C} \tag{10.10}$$

The result shown in Figure 10.4 (model 2) reveals that the very simple model defined by the equations 10.9 has unexpected properties. The assumption that $X1$ is not limited by $X2$ results in the carrying capacity being exceeded in an early stage of development. The model slowly corrects for this and for some period of time growth of $X2$ becomes negative. Finally, the model approaches an equilibrium where $X2$ becomes extinct and $X1$ uses the full capacity, C, of the system. The carrying capacity is unlikely to be exceeded in the real system, leading to the conclusion that the assumptions are not realistic. This demonstrates how easy it is to implement assumptions into a model, while carefully evaluating the results alone may prevent logical errors.

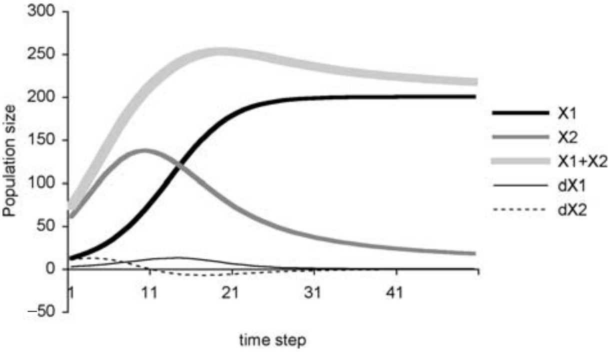


Figure 10.4 Logistic growth of two populations; model 2.

The result of model 2 gives rise to the discussion of three common pitfalls in numeric simulation:

Is the balance of matter, energy and information correct? Unlike natural systems, models can easily violate any law of physics or other laws. In a numerical model, matter leaving a source pool need not necessarily reach the target pool; it may simply disappear due to wrong assumptions. Model 2 is not critical in this regard as it omits for example consideration of the nutrients needed for the growth of the plants. If these were part of the model – say state variable $X3$ – growth would require a flow of matter from $X3$ to $X1$ and $X2$. It is good practice to check whether the total is the same at the beginning and the end of a simulation.

Are the numerical calculations correct? Often, the number of steps used in modelling is very large. Even minor rounding errors may accumulate due to the limited precision of today's computers. Very small rates of change may corroborate the calculations completely. Very large rates of change, on the other hand, may lead to numerical instability, a property the numerical model may not share with the real system. In model 3 I attempted to reproduce results of model 2 using five times as many time steps ($\Delta t = 0.2$). The differences models 2 and 3 generate are given in Table 10.2 for a few states in time. The shape of the curves are pretty much the same and overcrowding of the carrying capacity persists. Comparing models 2 and 3 suggests that the response pattern observed really reflects the property of the equations and not, as must always be suspected, the limited precision in computation.

Table 10.2 Comparison of model 2, $\Delta t = 1$, and model 3, $\Delta t = 0.2$.

	Model 2			Model 3		
t	x	y	$x + y$	x	y	$x + y$
1	15.3	72.0	87.3	13.2	63.3	76.5
2	18.8	84.2	107.2	16.5	75.2	91.8
5	34.2	118.4	152.6	31.7	109.8	141.5
10	80.5	136.0	216.5	78.6	130.7	209.3
20	180.5	71.6	252.2	177.9	72.7	250.6

Are the intrinsic assumptions of the model realistic? Again, a reference to model 2 illustrates the point. Is it possible that overcrowding can occur? Is the assumption made in this model that population $X1$ is not affected by $X2$ possible at all? In any case, the model will exhibit the outcome of these assumptions and will eventually no longer show the properties of the real system.

Sometimes the real system differs considerably from what we think would be logical, such as when societal systems frequently exhibit slow feedback mechanisms. It is easy to implement this into logistic growth models: taking Equation 10.7, I implemented a delay function. Using $X1_{t-5}$ instead of $X1$ and replacing $X2$ by $X2_{t-5}$ in limiting growth causes a backlash in the memory of the system, as it now considers the state of the system five time steps back rather than in the present. What happens is shown in Figure 10.5,

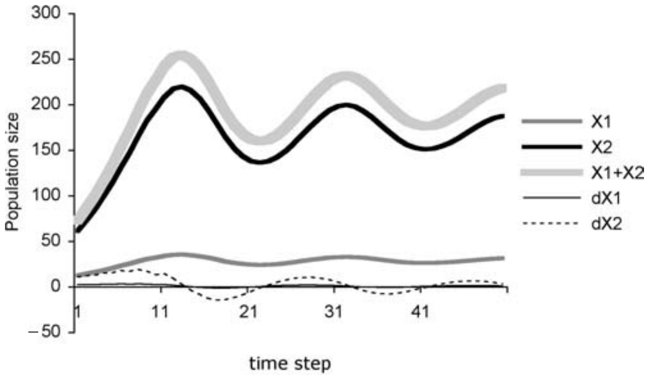


Figure 10.5 Logistic growth of two populations; model 4. Model 4 is initially identical to model 1, but a delay of five time steps is built in, causing oscillations.

where poor control causes the system to oscillate, just as the economy does when people continue spending money while forgetting about the desperate condition of their bank account.

10.2 Including space processes

In space–time models, space is assumed to be discrete, just like time in numerical integration (Section 10.1). For simplicity, two-dimensional spatial systems are frequently designed as rectangular, systematic grids. Spacing grid cells regularly and assuming finite extension facilitates computations. The state variables, which in time models imply no spatial extent, account for the content of these cells. Spatial interactions proceed through exchange between cells; this may involve matter, energy or information. Exchange is of course also a function of time and the model has to express how much is moved from one cell to the next per time unit. This transport is either directed or undirected (diffuse).

In ecosystems a diffuse process will hardly happen in isolation. Simultaneously, a temporal process is taking place inside all grid cells and these therefore contain their own temporal models. The entire model claims to describe a space–time process.

Assumptions have to be made about the exchange process between cells. In Figure 10.6, left-hand side, exchange takes place in the horizontal and vertical direction. An alternative would be to also allow fluxes in the direction of the diagonals. In Figure 10.6 two time steps are needed to reach the next diagonal cell.

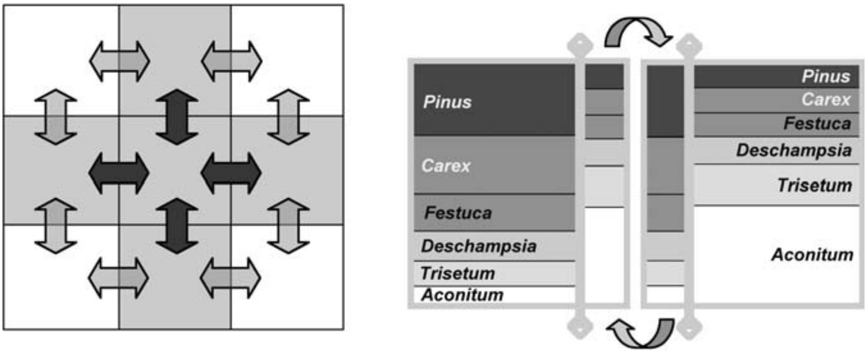


Figure 10.6 The mechanism of spatial exchange. Left: a model design in which exchange proceeds horizontally and vertically only. Right: partial exchange of total content of neighbouring cells to simulate diffusion.

When designing the exchange process, the balance of matter, energy and information has to be maintained. What leaves one cell has to arrive either in the next or in a controlled sink. An example in a directed process is water flow induced by gravity, for which orientation of the slope determines direction.

A different process is diffusion, where no specific direction is defined. An assumption of this kind is used in the succession model of the Swiss National Park (Wildi 2002, see Section 9.4.2), where diffusion applies to plant species and these potentially spread in any direction (Figure 10.6, right). I assume that a small fraction of the content of any two cells is exchanged at each time step. Some of the species newly arrived in a cell will successfully establish and spread, while others will disappear; this depends on the local conditions, set by the temporal model within each cell.

10.3 Processes in the Swiss National Park (SNP)

This is an example of the application of modelling techniques presented in the previous sections. The aim of model building was to reproduce the temporal pattern of succession revealed by space-for-time substitution, explained in Section 9.4.2. The description of the model follows Wildi (2002).

10.3.1 The temporal model

To keep complexity under control, species are grouped into six guilds (assemblages of species) (Wildi & Schütz 2000): these are the state variables. The basic process considered is thus colonization of plots and subsequent species interactions. The plots (the cells in the model) accord to the research grid established in the SNP for the purpose of investigation. Plot size is 20 m by 20 m and the number of plots within the unforested investigation area, Alp Stabelchod, is 286 (Achermann *et al*. 2000, see Figure 10.8). For simplicity, it is assumed that the total surface the species guilds occupy never exceeds 100% of the plot. The model plot is eventually overgrown by one or several species guilds, so that in the end no open soil is left.

Next, the objective is to quantify overgrowth and also replacement. In the original time series from the permanent plots it can be observed that overgrowth always starts slowly (Wildi & Schütz 2000). With increased cover of the guilds, spread accelerates. When approaching 100%, overgrowth

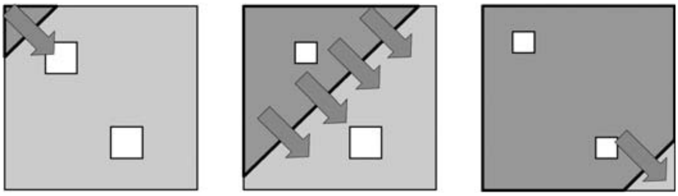


Figure 10.7 Overgrowth of a plot by a new guild. The white squares indicate patches inappropriate for growth, causing it to slow towards the end of the invasion process.

slows down. This finding is illustrated graphically in Figure 10.7. A function that mimics this behaviour is the logistic growth equation; in case of only one guild, it has the general form:

$$\frac{dX}{dt} = Xr\frac{100\% - X}{K} \tag{10.11}$$

(Wissel 1989, see also Equation 10.6). Here, r is the growth rate of guild X and K is the carrying capacity; that is, 100% of the plot surface. As X is also measured in terms of percentage, the space not yet occupied is $100\% - X$. Colonization stops when X reaches 100%. The growth is regulated by X itself, as a result of intra-specific competition. It must be noted that logistic growth requires all guilds X_i to be present in a minimum quantity at the beginning of any simulation run.

Competition comes into play because of two assumptions. First, the gain in cover of guild X_i is at the expense of any other guild's lower competition power (or open ground). In order to keep the cover percentages balanced, the growth equation will have two components: a gain by population growth and a loss to better competing species. Second, $100\%{-}X_i$ is the available space only for the best competing guild, i. If there is another, more successfully competing guild, X_j, then the space reduces to $100\% - X_i - X_j$. As will be seen in the description of the model, the mechanism has to make provision for many more competing guilds; six in the present case. Based on previous findings (Wildi & Schütz 2000) the following order of competition power was assessed:

$$Pinus(1) \rightarrow Carex(2) \rightarrow Festuca(3) \rightarrow Trisetum(4) \rightarrow$$
$$Deschampsia(5) \rightarrow Aconitum(6)$$

The logistic growth equation for $Carex(X_2)$, which is out-competed by $Pinus(X_1)$, is given by:

$$\frac{dX_2}{dt} = X_2 r_2 \frac{100\% - X_1 - X_2}{K} - \left(\frac{dX_1}{dt}\right) \frac{X_2}{\sum_{i=2}^{6} X_i} \tag{10.12}$$

Carex is growing according to the logistic growth equation. But in addition a portion of the surface that *Pinus* (X_1) is winning is subtracted. For *Festuca* (X_3) there is additional proportional loss to *Pinus* and *Carex*:

$$\frac{dX_3}{dt} = X_3 r_3 \frac{100\% - X_1 - X_2 - X_3}{K} - \left(\frac{dX_1}{dt} + \frac{dX_2}{dt}\right) \frac{X_3}{\sum_{i=3}^{6} X_i} \tag{10.13}$$

The growth equations for all subsequent guilds are built accordingly.

The third important factor in this pasture is recurrent disturbance; that is, trampling by grazing deer (Krüsi *et al.* 1998). I assume that it affects all the plants within a plot similarly. The intensity will of course vary depending on animal density. Trampling is a very fast process, instantly generating open space. This causes a loss t_i for guild i, which is simply proportional to its state, X_i. Re-colonization c_i is also fairly fast. I assume that it happens instantly; that is, within the relative short time span of one year, the standard time-step length of the model. It is proportional to the exponential growth of each guild. Direct competition, as happens in species replacement, is not assumed. Trampling and re-colonization are balanced within the year:

$$\sum_{i=1}^{6} t_i = \sum_{i=1}^{6} c_i \tag{10.14}$$

This assumes that growth is sufficiently fast to colonize any gap that has occurred within one year. Furthermore, trampling leads to a yearly shift of the guilds, favouring the fast growing, provided the growth rates r_i differ.

10.3.2 The spatial model

The following notation is used:

$$\vec{x}_{i,x,y,t}|i = 1, \ldots, 6;\ \ x = 1, \ldots, 25;\ \ y = 1, \ldots, 30;\ \ t = 1, \ldots, 400| \tag{10.15}$$

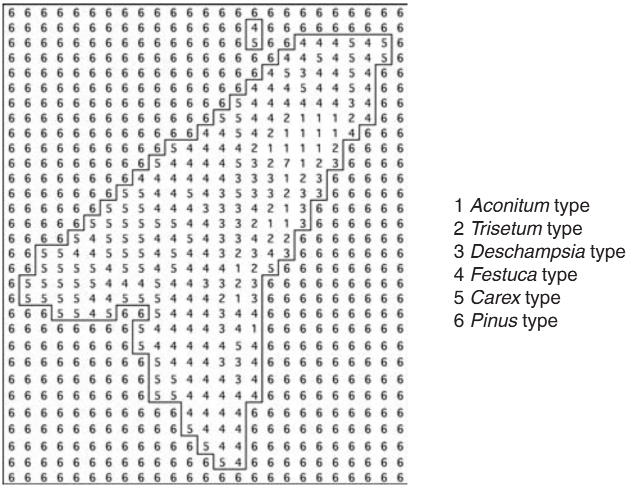
where i stands for guild, x and y are the spatial coordinates and t is time, in years. The model space is a grid of 25 by 30 plots (Figure 10.8). Not only the pasture but also the adjacent forest stands fit into this rectangle. The spread of any one guild happens by spatial exchange. A portion of the content of any plot is transferred yearly to the neighbouring plots, as shown in Figure 10.6. The gains, g, and losses, l, are balanced:

$$\vec{x}_{i,x,y,t+1} = \vec{x}_{i,x,y,t} + \vec{g}_{i,x,y,t} - \vec{l}_{i,x,y,t} \tag{10.16}$$

$$\vec{g}_{i,x,y,t} = d(\vec{x}_{i,x-1,y,t} + \vec{x}_{i,x+1,y,t} + \vec{x}_{i,x,y-1,t} + \vec{x}_{i,x,y+1,t}) \tag{10.17}$$

$$\vec{l}_{i,x,y,t} = 4d(\vec{l}_{i,x,y,t}) \tag{10.18}$$

From Equation 10.17 we see that the gain always comes from all four directions. The losses in all four directions (Equation 10.18) are the same as they are proportional to the composition of the central plot. The velocity of exchange is given by factor d. This is assumed constant, even though spatial



6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	4	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	5	6	6	4	4	4	5	4	5	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	4	4	5	4	5	4	5	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	4	5	3	4	4	5	4	6	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	4	4	4	5	4	4	5	4	6	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	5	4	4	4	4	4	4	3	4	6	6
6	6	6	6	6	6	6	6	6	6	6	6	6	5	5	4	4	2	1	1	1	2	4	6	6
6	6	6	6	6	6	6	6	6	6	6	6	4	4	5	4	2	1	1	1	1	4	6	6	6
6	6	6	6	6	6	6	6	6	6	5	4	4	4	4	2	1	1	1	1	2	6	6	6	6
6	6	6	6	6	6	6	6	6	5	4	4	4	4	5	3	2	7	1	2	3	6	6	6	6
6	6	6	6	6	6	6	6	4	4	4	4	4	4	3	3	3	1	2	3	6	6	6	6	6
6	6	6	6	6	6	6	5	5	4	4	5	4	3	5	3	3	2	3	3	6	6	6	6	6
6	6	6	6	6	6	5	5	5	4	4	4	3	3	3	4	2	1	3	6	6	6	6	6	6
6	6	6	6	6	5	5	5	5	5	5	4	4	3	3	2	1	1	3	6	6	6	6	6	6
6	6	6	6	5	4	5	5	5	4	4	5	4	3	3	4	2	2	6	6	6	6	6	6	6
6	6	5	5	4	4	5	5	5	4	5	4	4	3	2	3	4	3	6	6	6	6	6	6	6
6	6	5	5	5	5	4	5	5	4	5	4	4	4	1	2	5	6	6	6	6	6	6	6	6
6	5	5	5	5	5	4	4	4	5	4	4	3	3	2	3	6	6	6	6	6	6	6	6	6
6	5	5	5	5	4	4	5	5	5	4	4	4	2	1	3	6	6	6	6	6	6	6	6	6
6	6	6	5	5	4	5	6	6	5	4	4	4	3	4	4	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	5	4	4	4	4	3	4	1	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	5	4	4	4	4	4	5	4	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	5	4	4	4	3	3	4	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	5	5	4	4	4	3	4	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	5	5	4	4	4	4	4	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	4	4	4	4	6	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	5	4	4	4	6	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	6	5	4	4	6	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	6	6	5	4	6	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6

1 *Aconitum* type
2 *Trisetum* type
3 *Deschampsia* type
4 *Festuca* type
5 *Carex* type
6 *Pinus* type

Figure 10.8 Spatial design of the SNP model and the initial state (see Table 10.4 for contents of the cells).

processes may be faster where more animals prevail. Having no measurements of exchange at hand, I keep it at the very low level of $d = 0.001$.

Along the edges of the system, outside the meadow, the exchange is mirrored. All these plots are covered by *Pinus mugo* forest, the final state of succession considered in the model.

10.3.3 Simulation results

A comparison of the different graphs in Figure 10.9 shows that, with a suitable choice of parameter values, the temporal model can reproduce the basic pattern of the time series. In the model output (graphs on the right-hand side) fluctuations are absent. The results for the 415-year and the 585-year simulations were obtained only after carefully adjusting the initial conditions (the state variables) and approximating the growth rates of the guilds by trial and error. Initial conditions for both model runs are shown in Table 10.3. The initial states of the model are within the range of the values observed in the field (Wildi & Schütz 2000). It must be noted that the observed initial cover values are themselves affected by random fluctuation, whereas in the deterministic model a fixed value is needed. Simulation runs show that the time of emergence of late successional guilds depends on the initial states.

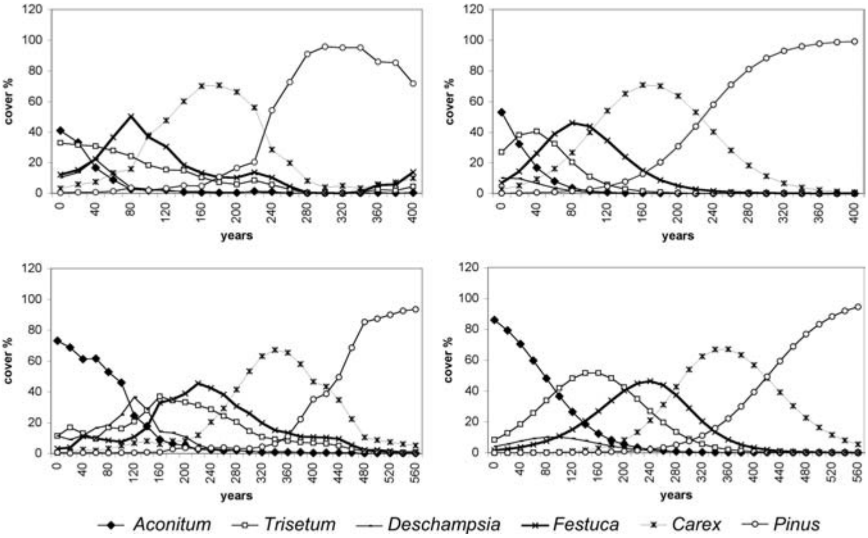


Figure 10.9 Original (left) and simulated (right) temporal succession. Upper row: 415-year version. Lower row: 585-year version.

Table 10.3 Initial values in the times-series data and initial state variables (percentage cover of six guilds) and growth rates used in the models.

series	415 year			585 year		
variable	state, % field data	state, % model	growth rate model	state, % field data	state, % model	growth rate model
Aconitum	41.1	53.0	0.050	73.1	86.0	0.018
Trisetum	33.0	27.0	0.045	11.4	8.4	0.022
Deschampsia	10.3	10.0	0.045	11.3	4.0	0.022
Festuca	12.1	7.0	0.045	3.0	1.6	0.020
Carex	3.1	2.8	0.035	0.6	0.03	0.026
Pinus	0.3	0.2	0.030	0.6	0.01	0.022

In other words, the initial state of the guilds determines the speed of succession. This is not a realistic feature as it does not consider invasion. Even worse, in order to allow growth to occur in a solely temporal model, all species guilds have to be present in the model (i.e. within all 268 plots considered) from the very beginning of the simulation.

The approximated growth rates yielding a realistic model behaviour do not differ much between guilds. That is, growth rates do not much affect the temporal pattern the model generates (Table 10.3). However, the two time series differ in their growth rates: in the 415-year model, growth has to proceed twice as fast as in the 585-year model. If all rates are set to 0.045 (not shown here), succession will last about 400 years; when taking 0.022 this will be close to 600 years.

Including spatial extent in the model creates problems with the initial condition of the meadow; that is, the state of all 268 plots in the year 1917 (outset of succession), which is not known to us explicitly. From the present state of the meadow and the direction and rate of change observed in many similar plots we are able to suggest a simplified state using the same composition for all cells belonging to the same type as initial conditions in the year 1917 (Table 10.4, Figure 10.8). Hence, at the beginning of simulation all plots of the same succession stage have identical species scores and the system consists of a limited number of discrete states, whereas in reality the vegetation forms a continuum. As soon as simulation begins, diffusion causes differentiation of cells and all maps become continuous.

Table 10.4 Six discrete vegetation states used as initial conditions of the six state variables (guilds) in spatial modelling, cover percentage. Maps in the first column ($t = 0$) in Figure 10.10 are composed of these states.

Guild	no.	state 1	state 2	state 3	state 4	state 5	state 6	$\sum$
Aconitum	1	50.00	17.50	17.50	10.00	5.00	0.00	100.00
Trisetum	2	10.00	35.00	35.00	15.00	5.00	0.00	100.00
Deschampsia	3	7.00	15.00	35.00	35.00	6.00	2.00	100.00
Festuca	4	2.00	3.00	30.00	42.00	20.00	3.00	100.00
Carex	5	1.00	1.00	10.00	15.00	65.00	8.00	100.00
Pinus	6	0.00	0.00	1.00	1.00	8.00	90.00	100.00

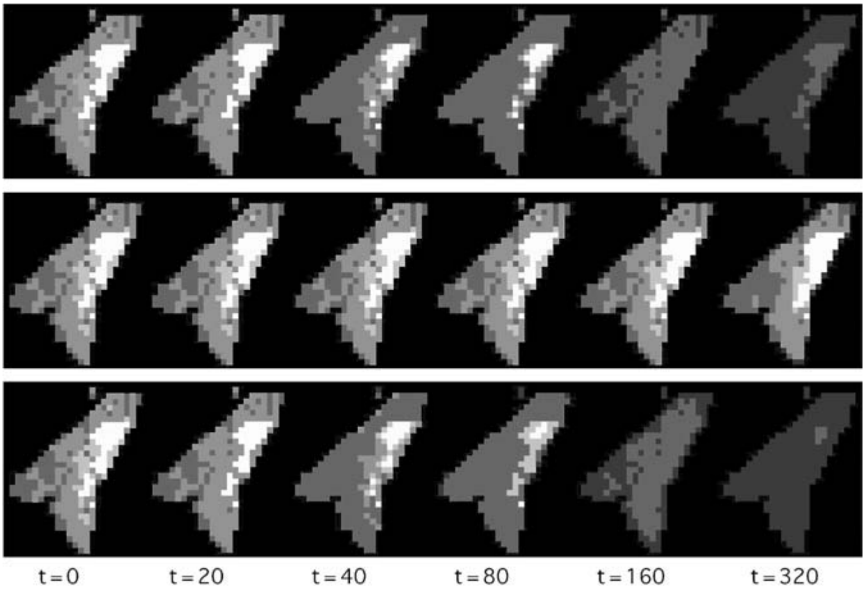


Figure 10.10 Results of the spatial simulation of succession, Alp Stabelchod. First row: temporal model only. Second row: spatial diffusion only. Third row: spatial and temporal processes combined. The sequence of white towards dark grey accords with succession states: *Aconitum* (white pixels) – *Trisetum* – *Deschampsia* – *Festuca* – *Carex* – *Pinus* (dark grey shading).

The persistence of the meadow through subsequent successional guilds and finally *Pinus* forests lasts about 500 years in simulations using the temporal model. Because of the lack of spatial interactions, vegetation boundaries do not move and only vegetation composition changes. As a result, the pattern formed by the edges remains unchanged over the entire simulation time. This can be seen in all states of the simulation run shown in the first row of Figure 10.10. Vegetation boundaries finally only vanish because all plots reach the final state of succession.

The effect of spatial exchange among pixels can be simulated in isolation. Assuming an extremely low rate of exchange of $d = 0.001$, the state of the system after 320 years is shown in the middle row in Figure 10.10. Overall composition is almost the same as at the beginning, but the compositional pattern of the maps has become more homogeneous compared to the initial state. Along the forest edges, *Pinus mugo* has invaded the first row of cells. Other types have spread as well. The *Aconitum* stage (white cells) has increased in surface. The spatial process changes the state of the meadow very slowly and does not explain the results from the permanent plot survey.

Finally, the spatial and temporal processes are run simultaneously. This accelerates the simulation of succession considerably and the meadow is almost covered by *Pinus mugo* after 320 years (Figure 10.10, bottom series). It can now be seen that vegetation boundaries have moved and differ from their initial state. The diffusion process causes *Pinus* to invade the meadow from the edges towards the centre.

The lesson to be learned from this simulation exercise is that neither the temporal model alone nor the spatial is suited to explain succession. In order to understand the major processes invasion has to be allowed, suggesting a spatial component be added to the temporal model. While simple models may be easy to handle and understand, extra complexity is often indispensable.

11 Large data sets: Wetland patterns

11.1 Large data sets differ

As collaboration among vegetation scientists evolves and large spatial and temporal scales are analysed in order to explore global change effects (Kienast *et al*. 2007) the data sets available for access are growing. While at first glance the problems in the analysis of large data sets appear to be the same as for those of normal size, in practice they differ in many ways. I consider two extreme data types:

- Type A. The survey data encompass a relatively small number of vegetation types. Consequently, any vegetation type will be represented through a large number of relevés. In such data sets redundancy tends to be high, the similarity pattern is frequently blurred by much random noise, but

Data Analysis in Vegetation Ecology Otto Wildi

correlations among sampling units are almost linear (because the relevés tend to be rather similar). This promises classifications (and hence synoptic tables) and ordinations will be efficient in displaying the underlying patterns.

- Type B. The survey data encompass a wide range of vegetation types in which, when comparing the extremes, few or even no common species exist. In this case nonlinearity is expected to be strong and the similarity pattern very high in dimension. Even measuring resemblance becomes unreliable. Hence, linear methods will suffer from low efficiency and the emerging similarity patterns will be difficult to present graphically.

In addition, some statistical and technical hurdles occur in both cases:

- Multipurpose software used for data organization may no longer be suitable: many spreadsheets even nowadays limit the number of columns to 255; these columns may need to hold the relevés. Using databases such as JUICE (Tichý 2002), TURBOVEG (Hennekens & Schaminée 2001) or VEGEDAZ (Küchler 2009) becomes a must.

- When large research teams are involved in the survey process taxonomical problems arise. The software mentioned above supports unifying different taxonomies.

- Not all large surveys rely on statistical sampling. Preferential sampling, as often encountered in databases, may come in line with over- and under-representation of vegetation types compared to the real situation. Ordination and classification will further accentuate this pattern. The result fails to yield an unbiased picture of reality.

- The chance of including aberrant observations – that is outliers – is increasing. Whether reflecting the real situation or just resulting from measurement errors or mistakes in data input, outliers are hampering most multivariate analyses.

- Graphical presentation may also become ineffective. Synoptic tables of just hundreds of relevés are difficult to print and even more so to inspect; in ordinations with thousands of data points groups and gradients vanish. When printing dendrograms of even a few hundred sampling units overview is lost.

In the sections below I concentrate on the analytical issue. Most approaches attempt to reduce the data in the hope that patterns will emerge despite the large samples and the high number of attributes. In most examples discussed below the sample size is drastically reduced, either by sampling units (reducing the number of relevés) or by the attributes (e.g. by eliminating redundancy in species lists or substituting species with different variables). This may appear to be an unjustified manipulation of reality, but the reduction holds only temporarily in order to detect a pattern; then a step back is taken and the full information is reconsidered.

The data used in this chapter are taken from the first survey of the Swiss mire-monitoring project (Grünig *et al*. 2005), referred to as 'Swiss wetland vegetation data' below. This data set captures a very wide ecological gradient of wetlands, ranging from ombrotrophic peat bogs through fens to open-water reed vegetation at altitudes from about 300 m a.s.l. to about 2400 m a.s.l. The full sample consists of 17608 relevés. Depending on the method used, subsets of this sample belong to either data type A or B as described above.

11.2 Phytosociology revisited

Whenever a data set encompasses a large range of vegetation types the description of content becomes cumbersome. Allocating such sets to established classification systems is one way of achieving a quick overview. The result of allocation will of course reflect the properties of the classification used. In the following example I focus on phytosociological classification, as introduced by Braun-Blanquet (1932). It is generally assumed that this system – at least for Europe – is comprehensive, robust, inert against spatial and temporal variability and so on (Dengler *et al*. 2008). In recent years a project surveying North America (Jennings *et al*. 2003) has promised increased reliability due to revised standards. I abstain from discussing the strengths and weaknesses of phytosociology here. In the example below I concentrate on the technique of allocation and the presentation and interpretation of the results. Flaws in the classification system will inevitably emerge if they exist. The example is taken from Graf *et al.* (2010). The starting point is the phytosociological system (or any other established system), offering one unique classification. Despite prescribing some rules for classification (Mueller-Dombois and Ellenberg 1974) the outcome in phytosociology always represents the opinion of individual experts (Ewald 2003); nonetheless I intend to suggest a mathematical solution for the allocation of relevés to this system, one that can be reproduced anytime and anywhere. As far as possible this should do what experts would do.

Using a variety of data transformations and resemblance measures we compared several published data sets classified by experts to the phytosociological database of Switzerland released by Pantke (2003). Pantke gives a list of all associations published to date (about 650). The names given to the associations and all higher hierarchical levels (alliances, orders, classes) conform to the standards used in phytosociology (Mucina 1997). The definitions of the units consist of species lists. These lists are incomplete as they include so-called character species, differential species and others that occur with high frequency only. This prevents us from using recently proposed methods such as Bruelheide's (1995, 1997) 'cocktail classification', which requires a complete and classified database of relevés (see also Kŏci *et al.* 2003). While the incomplete set of species will not deter us from using resemblance measures to compare relevés, it adds some uncertainty to the results. For our assignments of the 'Swiss wetland vegetation data' we applied the combination of transformations and resemblance functions where we found the consensus between the experts and the mathematical solution reached a maximum. For the transformation of the ranks of the Braun-Blanquet scale, this was $x' = log(x + 10)^{2.5}$ (see Table 3.3), a solution close to presence–absence. The product moment correlation coefficient (Table 4.3) performed slightly better than Ochiai's coefficient and Maarel's similarity ratio (definitions in Wildi & Orlóci 1996). It can easily be imagined that many correlations between the Swiss wetland relevés and the lists of Pantke (2003) are rather poor; to avoid misclassifications we set a threshold for assignment of $r \geq 0.2$ and also excluded ambiguous cases (i.e. relevés between two associations). Of the 17608 relevés only 2265 were assigned to any of the associations and remained in the data set; the associations were then assigned to the alliances simply by applying Pantke's hierarchy. This rigorous selection was then compared to the situation in the full data set: a random subsample of $n = 2265$ relevés was drawn from the full 'Swiss wetland vegetation data'. All relevés were also assigned to the associations and alliance of Pantke, but of course without using any threshold.

For ordinations, both data sets were subjected to principal coordinates analysis; in Figure 11.1 the random selection is shown. The proportion of total variance the first two Eigenvalues account for is $\lambda_1 = 7.7\%$ and $\lambda_2 = 5.1\%$. Highlighted distributions of the six most frequent alliances help in the interpretation of the point cloud. Obviously, the wetland data represent a gradient system forming a roundish triangle. The corners are formed

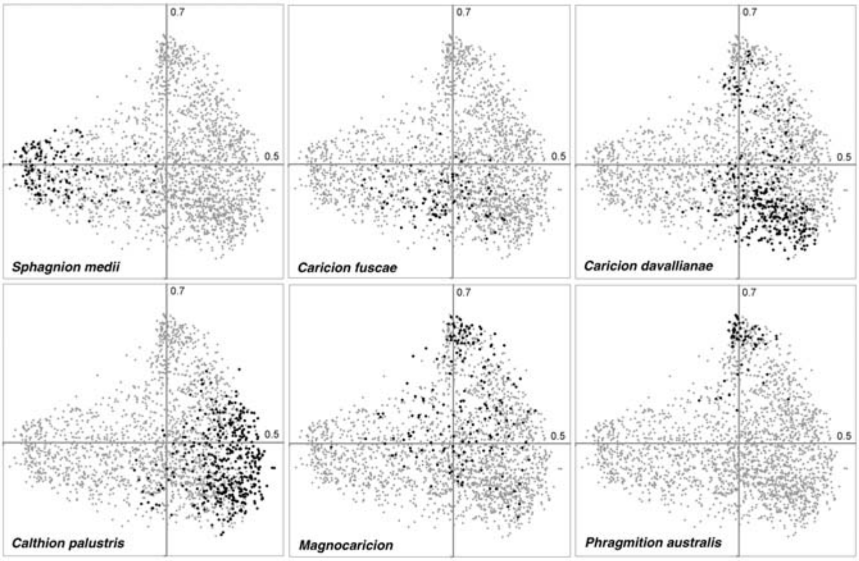


Figure 11.1 Six alliances represented in a random sample ($n = 2265$) of the Swiss wetland vegetation data (Graf *et al.* 2010).

by the alliances *Sphagnion medii*, *Phragmition australis* and *Caricion davallianae*.

In Figure 11.2 the same is done for the relevés best fitting the phytosociological system. The alliances are the same as in Figure 11.1. These ordinations clearly differ from the previous ones; this can be seen for example in the Eigenvalues accounting for $\lambda_1 = 10.4\%$ and $\lambda_2 = 7.7\%$ of total variation, and the triangular shape of these ordinations is far more pronounced than the same in Figure 11.1. What are the lessons to learn from this? First of all, phytosociologists managed to capture the extreme types of the wetland gradient, but they failed in documenting much of the variation in between. In many of the alliances discrete groups emerge: in the *Caricion davallianae*, the *Magnocaricion* and the *Phragmition australis* (Figure 11.2) for example. These subgroups are associations. The phytosociological classification on the level of alliances is hence inconsistent and may deserve revision. On the other hand, the pattern emerging in Figure 11.2 is really striking. In this regard phytosociology has considerable potential as a tool for visualizing pattern. Clearly, however, the Swiss database desperately needs to be supplemented by the full set of published species and its inherent classification deserves revision.

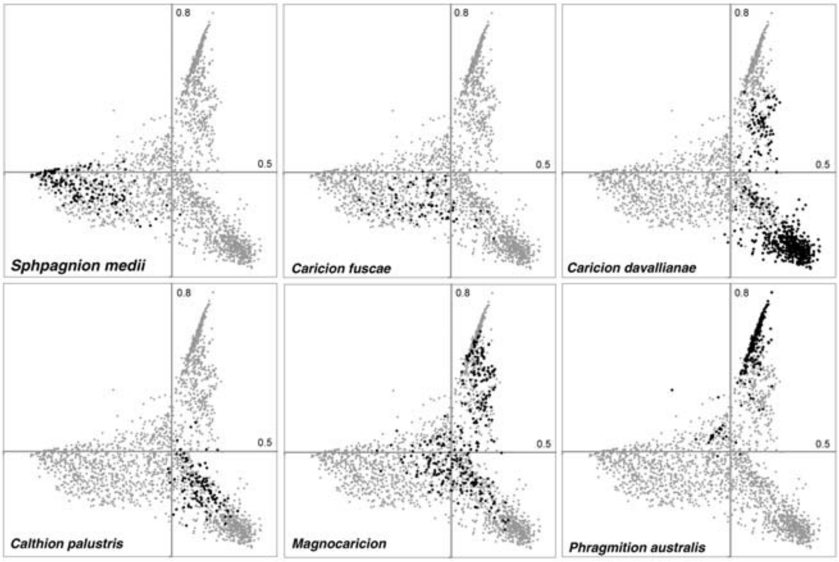


Figure 11.2 Six alliances represented in a relevé sample ($n = 2265$) best fitting the phytosociological classification of the Swiss wetland vegetation data (Graf *et al.* 2010).

11.3 Suppressing outliers

From a practical point of view an outlier relevé is a single representative of a 'vegetation type' different from all others present in a sample. How different it really is can be measured: it is the similarity (or distance) to the most similar relevé in the sample. This most similar relevé is the 'nearest neighbour'. The nearest neighbour of each relevé can easily be found in the similarity matrix by searching for the highest similarity value in the row or column pertaining to this relevé (except the diagonal element). To identify and erase an outlier a threshold is needed. This can be chosen preferentially: one can decide that any neighbour correlation of, say, $r \leq 0.5$ is deemed an outlier situation. However, before doing so it is good practice to find out what a 'normal' nearest-neighbour situation is in any one data set. Looking at the distribution function of nearest-neighbour resemblances does this.

An example is the bar chart in Figure 11.3. The data used is the random sample of 2265 relevés from the 'Swiss wetland vegetation data', with $n = 17608$ relevés, or 12.86% of the total. While this much smaller data set allows faster and more flexible processing it must be noted that taking a random subsample changes some of the data's properties considerably: as there

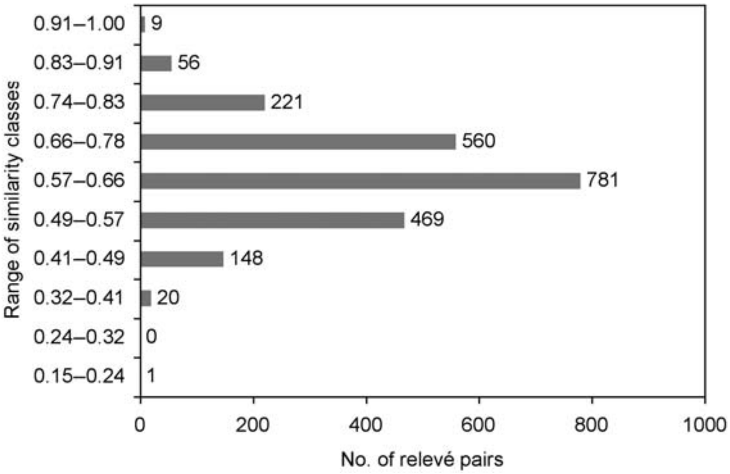


Figure 11.3 Frequency distribution of nearest-neighbour pairs of relevés in a random sample ($n = 2265$) of the Swiss wetland vegetation data. Product moment correlation coefficient r used.

are now much fewer sampling units involved, the mean nearest-neighbour distance will increase (but probably not the shape of the distribution function). Taking a random subsample will reduce the mean correlation among sampling units and therefore further accentuate the nonlinearity.

In Figure 11.3 the nearest-neighbour correlation coefficients in the data set are almost normally distributed. A real outlier would have to be far below these values. Clearly, no real outlier in the statistical sense exists in this data set. Despite this I'm now removing some of the most isolated relevés. This may be seen as a filtering process. In the present case I decided to remove all relevés having a nearest-neighbour similarity of $r < 0.6$. This reduces the sample from $n = 2265$ to $n = 1440$, or by about one third. The eliminated 'set of outliers' has size $n = 825$. The effect of filtering is shown in Figure 11.4: ordinations using principal coordinates analysis.

To help the interpretation, specific symbols highlight data points representing three extreme vegetation types. On the left-hand side the entire sample is included and the outliers are marked by dark symbols. The point cloud they form is the same as in Figure 11.1. On the right-hand side the ordination is shown using the remaining 1440 data points only. The group and gradient pattern is now much more striking: it confirms the triangular shape and the gradient pattern extending between the corners. One has to keep in mind, however, that the true pattern remains the same as that on

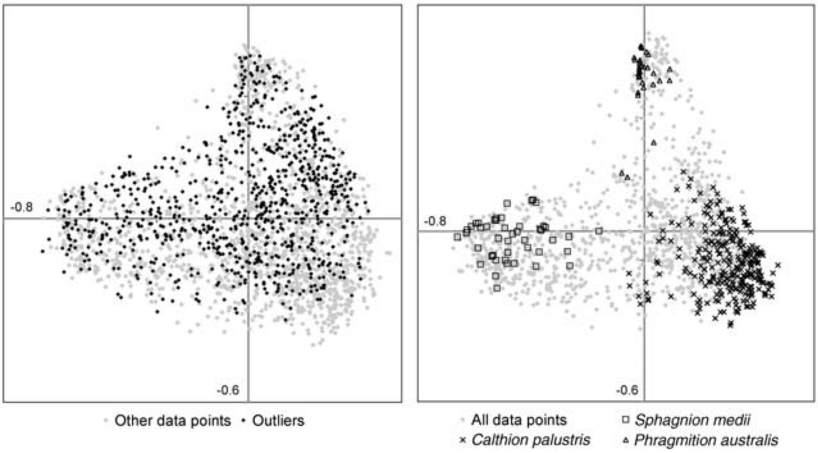


Figure 11.4 Ordination of a random sample of the Swiss mire vegetation data (left). Black data points are 825 'outliers'. The remaining 1440 relevés are ordinated on the right-hand side with the result that the triangular pattern appears pronounced.

the left-hand side, whereas the ordination on the right-hand side exaggerates it. It is easy to find the reason for this: the outliers are most frequent in low-density areas of the ordination. The high-density areas will therefore constitute the group pattern of the reduced sample. This is a welcome property which helps in the interpretation in many cases.

Should outliers be eliminated in order to successfully order synoptic tables as well? The answer is yes. Unfortunately, doing this with the present example would lead to disappointing results. I summarize my experiences and some rules of thumb when using synoptic tables for large data sets in Section 11.5.

11.4 Replacing species with new attributes

Analysis and interpretation is tremendously simplified when the number of attributes (e.g. the species lists) can be reduced. Because the number of species is usually very large the original relevés are over-determined: the number of attributes exceeds the number of sampling units. The goal of species reduction is therefore to erase redundancy; that is, where information is carried by two or more species simultaneously. This is most efficiently done by the RANK algorithm (Orlóci 1978, see Section 5.6). As explained there, the reduced list of species usually accounts for a surprisingly large amount of explained variance, but it still operates in the same similarity space (the species space), although reduced. Alternatively, we may project our data into a different resemblance space. Several of these spaces are well established in vegetation science:

- The species indicator values, as proposed by Ellenberg (1974), or those of Landolt (1977).
- Growth forms of dominating species, as introduced by Raunkiaer (1937).
- Plant functional types (Box 1996), in which the functioning of the species in its environment is considered.
- Character set types (Orlóci & Orlóci 1985, Orlóci 1991), where anatomical and physiological features of the individual plants represent the new attributes.
- In recent years the term 'trait' (i.e. species described by traits) has increasingly been used (Pillar *et al.* 2009) to investigate trait patterns of vegetation types.

The formal process projecting relevés via species lists into a new variable space is the same as that used in principal component analysis (Section 5.2). It is achieved through matrix multiplication and the new variables are linear combinations of the original; this is shown schematically in Figure 11.5,

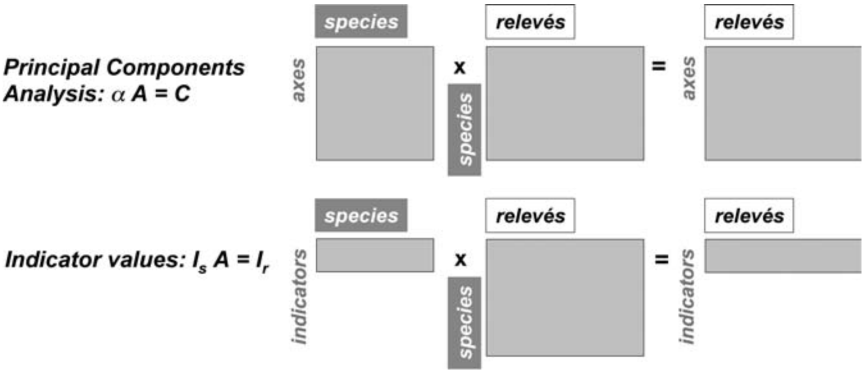


Figure 11.5 Projecting a given sample into a new resemblance space. Top: the example of PCA, where species are substituted by axes. Bottom: the species list is replaced by a list of indicator values. See also Pillar *et al.* (2009).

where the new attributes replacing the species are indicator values (see e.g. Pillar *et al.* 2009 for a similar illustration). What is needed in addition to the survey data is a matrix of species by indicator values. Here I use the list published by Landolt (1977) calibrated for wetlands, as explained by Feldmeyer-Christe *et al.* (2007). In the example below the 2265 randomly selected relevés from the Swiss wetland data are processed accordingly. In this way the 1413 species are replaced by as few as 8 indicator values. To reveal the new similarity pattern of the relevés and its relation to the indicator values I choose correspondence analysis (Section 5.4); the result is shown in Figure 11.6. The resulting point cloud once again has a triangular shape. As

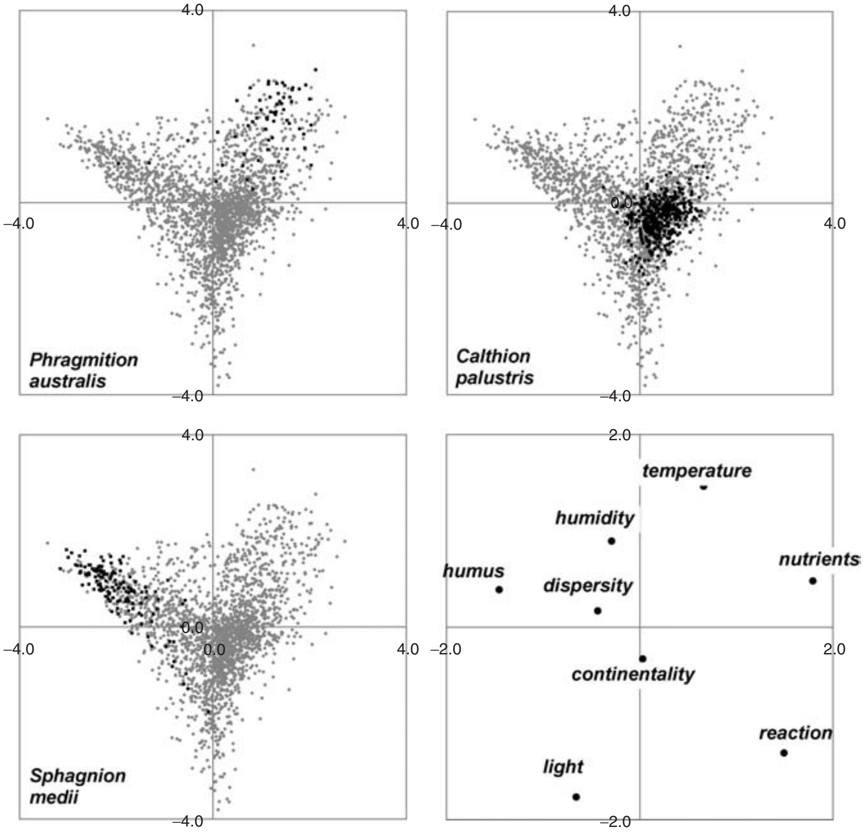


Figure 11.6 Ordination of the wetland sample in the indicator space using correspondence analysis. The location of three vegetation types is shown. The usually superimposed plot of indicator values is shown separately in the lower-right ordination.

in Figure 11.1, two of the corners represent the alliances *Phragmition australis* and *Sphagion medii*, but this time the alliance *Calthion palustris* forms the centre of the ordination. The explanation resides in the characteristics of the method: this alliance is very well represented in the sample (398 relevés). Because correspondence analysis operates with deviations from the overall expected state of the sample, frequent types are considered 'normal' and therefore projected close to the centre of the ordination.

The space of indicator values separates the established vegetation types just as well as the species space does. In Figure 11.6 the ordination of the indicator values (lower-right graph) is not superimposed on the ordination of relevés, as is commonly practised in correspondance analysis, but is printed separately. The direction of the data points as seen from the centre accords with locations of high values. The length of that same vector expresses the explanatory power of the respective indicator value.

The species space and the indicator space can also be superimposed, as shown in Figure 11.7. The ordination coordinates are the same as in Figure 11.1. The diameter of the bubbles representing the relevés is proportional to one indicator value at a time. One could therefore generate eight plots of the kind: one for each indicator value. Figure 11.7 yields an ecological interpretation of the ordination. The corner on the left-hand side, for example, representing the *Sphagnion medii*, carries high humus values (the peat) and low nutrient values, as is found in peat bogs.

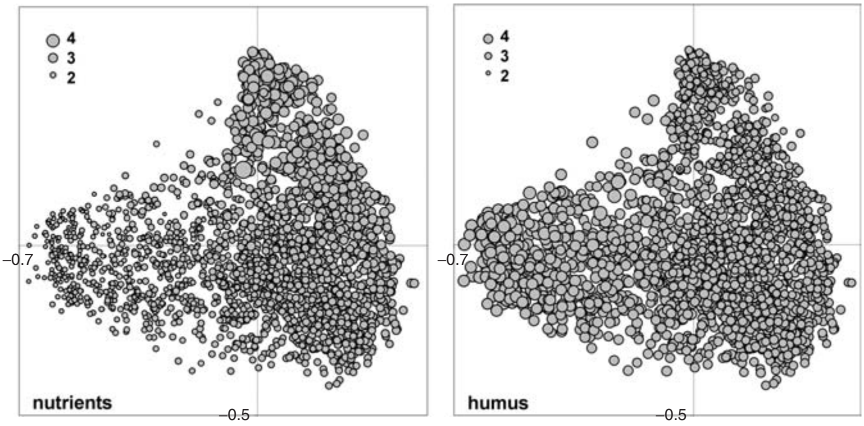


Figure 11.7 Indicator values superimposed on ordinary ordination. The diameter of the bubbles representing the relevés is proportional to one indicator value at a time.

11.5 Large synoptic tables?

Synoptic tables visualize vegetation patterns by using the field data directly. Some limitations have been mentioned in the introduction to this chapter, such as the excessive size and the difficulties in printing and inspecting. When analysing large data sets of thousands of relevés, printing the full sample is out of the question and a selection is needed. The data sets generated in Section 11.2 are an example of this, where two samples of similar content promise to have different properties despite belonging to type B (addressed in Section 11.1), the more critical case. It can be expected that the alliances to which the relevés were assigned exhibit low internal variation (i.e. high similarity) in the case of the selective sample but high internal variation in the case of the random sample. This results in two questions: (i) Does rigorous selection yield well-defined alliances? (ii) Does rigorous selection generate well-separated alliances? These questions shall be answered by inspecting a variety of parameters yielded by the analyses. In order to be able to produce printable tables in an ordinary book like this an additional reduction of the sample size was indicated, from 2265 relevés to about 100 or so. As for the species, probably no more than 80 would fit a double page.

Determining the within-alliances variation is only meaningful if a sufficient number of relevés per alliance are involved. The first step in both data sets therefore was to eliminate all relevés except those belonging to the 10 most frequent alliances. Just by chance these are the same in both cases. In most other aspects the data sets differ:

- In the *random sample* 70 alliances were found, most represented by just a few relevés. After reducing these to 10 alliances, 1548 relevés remained out of 2265, with 1185 species involved. For final ordering I took a systematic sample of every 16th, resulting in 97 relevés and 612 species.

- In the *selective sample* only 50 alliances were found. After reducing these to 10 alliances 1937 relevés remained out of 2265, with 1125 species involved. For final ordering I took a systematic sample of every 20th, resulting in 97 relevés and 500 species.

Before selecting the 10 alliances there was clearly more variability in the random sample. Even afterwards the still higher number of species shows higher dimensionality compared to the selective sample. For a comprehensive presentation of within- and between- variation, similarity matrices are

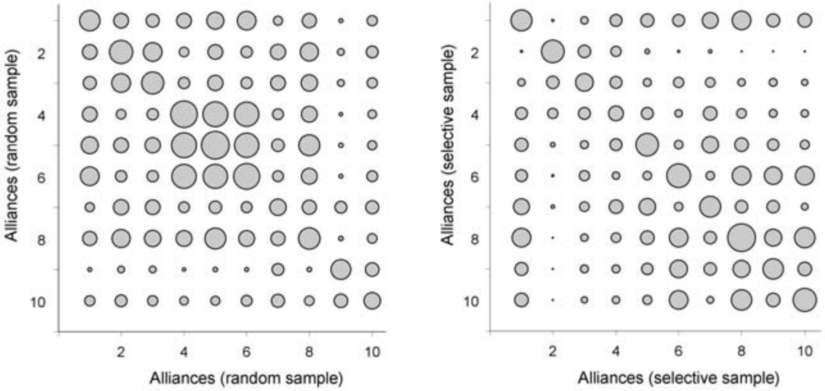


Figure 11.8 Similarity matrices of 10 alliances from the Swiss wetland vegetation data. Left: random sample. Right: selective sample.

derived, as introduced in Section 4.6. The result from the random sample is shown on the left side of Figure 11.8; that from the selective sample is on the right. The two subsamples show totaly different dispersion patterns. Assigning all relevés to the alliances found in Pantke's (2003) database results in a poor group structure (left graph in Figure 11.8); for example, alliances 4, 5 and 6 have almost the same composition, the within-group similarity being the same as the between-group similarity. Off-diagonal elements (the between-group similarity) are generally large, suggesting a more continuous pattern is to be found than in the selective sample to the right. There, the groups are more distinct. Within-group similarity, on the other hand, is hardly higher than in the random case and heterogeneity within the alliances has not really been reduced by the selective procedure.

Both data sets are then analysed to form a structured synoptic table, as explained in Section 6.6. The number of relevé groups chosen is 10 (assuming that the underlying alliances were governing the pattern), the number of species groups 100, with 80 species involved in printing. While visual inspection of Tables 11.1 and 11.2 reveals obvious differences, analysis of concentration (Section 7.4) involved in the ordering process offers measurements:

- The *random sample* achieved a fairly high mean square contingency coefficient of $C = 0.279$ (Formula 7.16). The first three Eigenvalues accounted for the following explanatory power: $\lambda_1 = 22.48\%$, $\lambda_2 = 19.57\%$ and $\lambda_3 = 13.38\%$.

Table 11.1 Synoptic table of a random sample of the Swiss mire vegetation data. 97 relevés in 10 groups, based on 100 species groups, 80 species displayed.

```
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|     RELEVE NO.                                       |2  1 1|  22 21|11  2 111   1 111 11 |11212211|1  1 11      |1 11  11 |1 111  2|1 1 111 1| 1 11 |1 1   211 |
|                                                      |125932|1 12327|09111548332 019983463|00080025|386971626 767|2956  675|28345681|145401497|214239|6587490785|
|                                                      |655773|3731945|907903347019119203665|75183556|3295961804734|9689 2432|75502699|334531941|381099|7062727728|
|                                                      |549652|8615595|640058052216283511712|45566730|0982704059569|284646310|45180991|952436973|175300|9165369710|
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|     RELEVE GROUP NO.                                 |      |       |111111111111111111111|        |             |         |        |         |      |          |
|                                                      |444444|5555555|000000000000000000000|88888888|9999999999999|666666666|77777777|222222222|111111|3333333333|
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 652 Phalaris arundinacea L.                        91|34   1|      3|                     |        |             |         |        |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 239 Phragmites australis (Cav.) Steud.             45|444241|4444 4 |      2    2         |        |             |         |        |         |      |          |
|  61 Carex elata All.                               45|2 34 4|       |     1     3 2       |        |             |         |        |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 372 Typha latifolia L.                             65|      |22   1 |                     |        |             |         |        |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 175 Holcus lanatus L.                              74|      |14     | 2  1231  22 3 31   1|        |             | 1       |        |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 148 Galium mollugo aggr.                           35|      | 422 22|            1        | 1      |1        2   |2        |        |         |      |          |
| 256 Poa pratensis aggr.                            35|      |1 24  2|            3       2|   4    |             |         |        |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 142 Filipendula ulmaria (L.) Maxim.                70|      |4423 14|    2 21 2 332  222 2|  2 22  |         1   |1132  12 |        |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 159 Geum rivale L.                                 56|      |2   4 2|    1         2      |        |1       1    | 2211  22|   2    |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 140 Festuca pratensis Huds.                         7|      | 2  1  |   3      1          |2 2 33 3|             | 12    1 |   2    |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  79 Cerastium fontanum Baumg. subsp. vulgare (Har  33|      |       |          1     1    |  2 2112|             |         |   12  2|    1    |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 111 Drepanocladus cossonii (Schimp.) Loeske        18|      |       |             1     1 |        | 2 3  1   432|         |        |         |      |          |
|  57 Carex davalliana Sm.                           18|      |       |  3   4  2    3  22  |        |2342343323 23|     3   |2  223  |         |      |          |
|  49 Campylium stellatum (Hedw.) J. Lange & C. Jen 18|  1   |       |  3  31      1   3   |        | 2 33333 2 22|     2 1 |   1 3  |     1   |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  68 Carex nigra (L.) Reichard                      26|      |       |3 321  4231 4 3 13 24|13233 33|2 2 3 123234 |     1 1 |   1223 |343432  2|      |2 2221   2|
| 268 Potentilla erecta (L.) Raeusch.                26|      |    21 |32332 33223223 2 4323|        |2323232 22 21|11 112122|  21 2 4|2134121 4|      |2 23    22|
| 101 Dactylorhiza maculata (L.) Soó                 26|      |       |   1   1111   2 1 4  |        |1 1    2 1   |1 12 1 21|111121  |     1   |      |        1 |
|  19 Anthoxanthum odoratum aggr.                    26|      |1    12|2 331 321222223232314|32432233|2 22  21     | 1 2 2 3 |212232 3|2133222 3|      |        2 |
| 350 Succisa pratensis Moench                       26|      |       | 1321 233213    13322|   2    |     3 3 2   |    111  |   1    |         |      |  1     2 |
|  46 Caltha palustris L.                            26| 1    |    32 |   32    2322 3    1 |2 242 24|1 222 222 2  | 23224243|4  32   |2   22   |      |          |
| 141 Festuca rubra aggr.                            26|      |2  2 2 | 2222 22 32132  4 212|32 3 33 |  2   2 3    | 2  13 2 |1 323234| 23      |      |   2      |
| 367 Trifolium pratense L. subsp. pratense          26|      |       |   2121 222       312|434 4233|2 22 13 2    |   2     |2 4222  |  3 11  2|      |          |
|  94 Crepis paludosa (L.) Moench                    26|      |     12|  314   2222  3   2  |21 32   | 132   2 1   |2222 3123|2  321  |        1|      |          |
| 370 Trollius europaeus L.                          26|      |       |  3 2   41      122  |3223322 |2 33  2331   |2  2 3124|2 32    |         |      |          |
|  36 Briza media L.                                 26|      |       |   11 3 1312  311232 |22 2223 |2 32  222   2|     2 1 |  22 2  |     1   |      |          |
| 283 Ranunculus acris L.                            26|      |       | 1221 11 13   3 412  |32343233|     12 2    |       2 |   221  |         |      |          |
|  86 Climacium dendroides (Hedw.) Web. & Mohr       26|      |   2   |   22 132223334 43334|33 33122|2 3    34    |         |    21  |  3      |      |          |
| 128 Equisetum palustre L.                          26|1     |   2 3 | 22 3 323423 23 13211|2 1  32 |1423 42423   | 121 3132|2  2 1  | 23     2|      |          |
|  70 Carex panicea L.                               26|   4  |       |2  31223 212 24231321|2 1    2|1 22332332 22|       2 |   2 2  |  2      |      |          |
| 226 Myosotis scorpioides L.                        26| 1    |    1  |   11  1211   2 1 21 |21222122|      1      |         |1  1    |  2      |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 227 Nardus stricta L.                              15|      |       |3  2   2  2    12 223|13      |      3 4    |         | 34 3244|4  3122 2|      |         1|
| 368 Trifolium repens L. subsp. repens              15|      |       |   1            2   2|23   433|             |         |  31    |  2      |      |          |
| 262 Polygonum bistorta L.                          15|      |3   3 3|  33              2 3|222 3123|  3     3    |   211 3 |   1    |  3      |      |          |
| 324 Silene flos-cuculi (L.) Clairv.                15|      |       |   1      122   1 2 2|212 2 1 |             |         |   1    |         |      |          |
|  96 Cynosurus cristatus L.                         15|      |2      |   2      1     1 3  |33 33333|      1      |         |  31    |  3      |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  58 Carex flava aggr.                              16|      |       |3 21    22   132  3  |11    2 |12 2 421322  |       12|1  223  |         |      |          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 385 Veronica officinalis L.                       76|      |       |                   1 |       2|             |         |1 1  1 2|         |      |          |
| 357 Thymus serpyllum aggr.                        76|      |       |        2 1          |       1|             |         |1 1221 2|         |      |          |
| 152 Galium pumilum Murray                         76|      |       |              2   1  |       2|   1         |         |1 22 1 3|         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 155 Gentiana asclepiadea L.                       28|      |       |        2       12   |        |1  2 3 2     |2  111111|   2 2  |         |      |          |
| 581 Primula farinosa L.                           28|      |       |        2       1    |        |22 2 4122   1|         |   1 2  |         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  81 Chaerophyllum hirsutum aggr.                   6|      |       |                     |   2    |             |414424143|1  22  1|      2  |      |          |
|   4 Aconitum napellus aggr.                        6|      |       |                     |        |1       1    |2 3   122|   22   |         |      |          |
| 158 Geranium sylvaticum L.                         6|      |      3|         2        1  |        |1  1         |  2212 32|   1   1|         |      |          |
|  71 Carex paniculata L.                            6|      |       |                     |1 2 2   | 2 2     2   |  33 2233|   3    |         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 870 Centaurea montana L.                          51|      |       |                     |        |             |   1  1 1|        |         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 236 Philonotis calcarea (B. & S.) Schimp.         39|      |       |                     |        |             |         |   223  |         |      |          |
| 443 Cratoneuron falcatum (Brid.) G. Roth          39|      |       |                     |        |  4342 2     |         |   334  |         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 399 Alchemilla conjuncta aggr.                     9|      |       |                     |        |             |         |1  21  2|     1   |      |          |
| 511 Poa alpina L.                                  9|      |       |                     |        |             |         |2  22132|         |      |          |
| 825 Soldanella alpina L.                           9|      |       |                     |        |      1      |         |1  1 11 |         |      |          |
| 424 Campanula cochleariifolia Lam.                 9|      |       |                     |        |             |         |1 22 1  |         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 261 Polygonatum verticillatum (L.) All.           21|      |       |                     |        |             |11  2    |        |         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|   3 Achillea millefolium aggr.                     5|      |       |                     |        |             |         |  1 1  1|         |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 404 Athyrium distentifolium Opiz                  31|      |       |                     |        |             |         |        |2   23   |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 105 Dicranodontium denudatum (Brid.) Britt.       59|      |       |                     |        |             |    2    |        |    332  | 2 312|          |
| 264 Polytrichum formosum Hedw.                    59|      |       |                     |        |             |    3    | 1 1    |    233  |  4244|          |
| 110 Dicranum scoparium Hedw.                      59|      |       |        2            |        |             |    2 2 2|22      |    323  |3 3321|          |
| 375 Vaccinium myrtillus L.                        59|      |       |        2          2 |        |           1 | 2  3 1 2|13 1 2 2|2 1 434 2|443443|1    3  2 |
| 327 Sorbus aucuparia L.                           59|      |       |                     |        |             |12  2  11|31      |11   13  | 31212|          |
| 241 Picea abies (L.) H. Karst.                    59|      |   3   |                   3 |        |             |14 23 434|43 324 2|22  443 2|344444|23   31   |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 263 Polytrichum commune Hedw.                     12|      |       |                     |        |             |         |        |44422 4 4|32 112|   22     |
|  55 Carex canescens L.                            12|      |       |                    2|        |             |         |        |34 22 22 | 1   1|          |
|  42 Calliergon stramineum (Brid.) Kindb.          12|      |       |3      1   1     3   |        |             |         |        | 11   121|      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 335 Sphagnum magellanicum Brid.                   24|      |       |  2         3        |        |             |    2    |        |43  22  4|2  4 3|44   131 4|
| 133 Eriophorum vaginatum L.                       24|      |       |  2         2        |        |   2         |        1|        |33 3211 3|     1|3344243233|
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 376 Vaccinium oxycoccos L.                        30|      |       |                     |        |             |         |        |   1     |      |132  2 12 |
| 332 Sphagnum recurvum P. Beauv.                   30|      |       |                     |        |             |    1    |        | 4 4   34|   3 4|443 44 332|
| 374 Vaccinium uliginosum aggr.                    30|      |       |  2                  |        |           3 |         | 3   1  |         |      |  434 3422|
| 329 Sphagnum capillifolium (Ehrh.) Hedw.          30|      |       |        3   4        |        |           4 |         |        | 2 242   |333  2|  44 2 423|
| 255 Pleurozium schreberi (Hedw.) Mitt.            30|      |       |  3    3    2        |        |           2 |         | 2      |  33324 1|433  3|  4 4  33 |
| 377 Vaccinium vitis-idaea L.                      30|      |       |                   1 |        |           2 |         |14   2 2|  1233   |3232 1|  12   1 2|
|  45 Calluna vulgaris (L.) Hull                    30|      |       |                 2 2 |        |           2 |         | 2 122 2| 2 322   | 2    |  422 2441|
|  27 Aulacomnium palustre (Hedw.) Schwaegr.        30|  1   |       |  4 3   3  33    3214|        |        2  2 |    1    |       2|23342 2 4|3     | 23122 333|
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 219 Melampyrum pratense L.                        60|      |       |                     |        |             |         |        |    222  |      |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 211 Lycopodium annotinum aggr.                     2|      |       |                     |        |             |         |        |         | 3  1 |          |
| 528 Rhytidiadelphus loreus (Hedw.) Warnst.         2|      |       |                     |        |             |    2    |        |     2   | 23312|          |
|   1 Abies alba Mill.                               2|      |       |                     |        |             | 4  1    |        |         | 31321|          |
| 299 Rubus fruticosus aggr.                         2|      |       |                     |        |             |2   4    |        |         | 1 12 |          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 114 Drosera rotundifolia L.                       47|      |       |                     |        |             |         |        |         |      |22    1  1|
|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
```

Table 11.2 Synoptic table of a selective sample of the Swiss mire vegetation data. 97 relevés in 10 groups, based on 100 species groups, 80 species displayed.

```
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    RELEVE NO.                                      |1 21 |1  1211 2|122  1 2|1  2121221  122  1222 22|2211 |1 11|211 121222111|  1 |1 2   2  2  |11222 21 1 21|
|                                                    |35635|376274266|16716525|133271330534442413354244|5360 |6641|6252724121566|7154|246 2355 345|00010101 0 20|
|                                                    |41028|811511282|89324910|517223648532319060978646|5749 |9663|5708478008272|3979|047945247276|0645642144252|
|                                                    |58438|813164698|13800975|829633668367757263483202|32913|6616|1766942820742|6053|573917904405|3818721275705|
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    RELEVE GROUP NO.                                |     |         |        |                        |11111|    |             |    |            |             |
|                                                    |55555|888888888|66666666|444444444444444444444444|00000|9999|7777777777777|2222|333333333333|1111111111111|
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 101 Dicranum bergeri Hoppe                       60| 24 2|         |        |                        |     |    |             |    |            |             |
| 353 Vaccinium oxycoccos L.                       60|13233|         |        |                        |     |    |             |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 310 Sphagnum recurvum P. Beauv.                  65|33134|  2      |     24 |                        |   3 |    |         32  |    |            |             |
| 307 Sphagnum capillifolium (Ehrh.) Hedw.         65|4 233|2    4   | 42  2  |  3                     |   2 |    |     2       |    |            |             |
| 230 Pleurozium schreberi (Hedw.) Mitt.           65|44334|2 2  2   | 42  42 |                 2      |     |    |   3         |    |            |             |
| 351 Vaccinium uliginosum aggr.                   65|133 4|3    22 3|22  2   |    2       1           |     |    |             |    |            |             |
| 354 Vaccinium vitis-idaea L.                     65|24 44|2    2  2|13  113 |  21        1 1     1   |     |    |     1       |    |            |             |
| 127 Eriophorum vaginatum L.                      65| 1233|2  2 2233| 3    2 |       1   2      3 11  |     |    |             |    |            |             |
| 228 Pinus mugo Turra subsp. uncinata (DC.) Domin 65|44 23|     11  |    4   | 1                      |     |    |             |    |            |             |
| 313 Sphagnum magellanicum Brid.                  65|42 22|4 4  4334|        |                        |     |    |             |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  29 Betula pubescens Ehrh.                       21|   21|         |24 443 3|                      1 |     |    | 2      2    |    |            |             |
| 417 Dicranum scoparium aggr.                     21|     |1        | 2122 2 |                        |     |    |     2       |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 206 Melampyrum pratense L.                       42|     |2    11 2|        |                        |     |    |             |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 352 Vaccinium myrtillus L.                       34|44 43|3 2  4  2|233 4342|   2 1           1   2  |     |    |     2    2  |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 248 Polytrichum formosum Hedw.                    2|     |  2  2   |  343422|   1        1           |     |    |             |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  25 Aulacomnium palustre (Hedw.) Schwaegr.       40|  4 1|2 2 21  3|      2 |  2             1       |   32| 221|             |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  64 Carex nigra (L.) Reichard                    24|     |3 3 12 22| 2    3 |333 2122223421  1 2  1 3|44443|  32|  44322   222|    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 110 Dryopteris carthusiana (Vill.) H. P. Fuchs   62|     |         |  3 4213|                        |2    |    |        2    |    |      1     |             |
| 171 Hypnum cupressiforme Hedw.                   62|     |         | 22 22 2|                        |     |    | 2        2  |  2 |            |             |
| 280 Rubus idaeus L.                              62|   1 |         |  2 1 22|                        |     |    |          1  |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 148 Gentiana asclepiadea L.                      26|     |         |        |  22  1  1 1 11   2 11 2|     |    |             |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  82 Climacium dendroides (Hedw.) Web. & Mohr     13|   1 |  1      |     2  |2 3 4 1     2    2      |   21|    |323332  3233 |    |     2      |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 505 Blysmus compressus (L.) Link                  6|     |         |        |2322 2  3  221  3 2 1   |     |    |           21|    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 679 Ligusticum mutellina (L.) Crantz             93|     |         |        |                        | 32 2|    |             |    |            |             |
| 699 Saxifraga stellaris L.                       93|     |         |        |                        | 22  |    |             |    |            |             |
| 648 Arabis subcoriacea Gren.                     93|     |         |        |                        | 22  |    |             |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 227 Pinguicula vulgaris L.                       67|     |         |        |  21 1    2 11  1213 12 |     |    |3    1    1  |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 211 Molinia caerulea (L.) Moench                 55|  3  |24 332222|    4  4|1 4 3 32323  32 44224434|     |    | 2 3  3   4  |3433|            |             |
| 340 Trichophorum cespitosum (L.) Hartm.          55|  4  | 4 34 44 |        |3 3 4323 41   2  4  44 4|     |    |          4  |    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 230 Plagiomnium affine aggr.                     50|     |         |     1  |      1  2         2 2 2|     |    |4 44 233 2232|    |          11|             |
| 265 Ranunculus aconitifolius L.                  50|     |         |        |  222 2    11        1  |     |    |312 3 13 32 3|    |            |             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 117 Epilobium palustre L.                          53|      |         |     1  |             1          |1    | 222|  11  11  1 21|  1 |         1   |              |
|  72 Carex rostrata Stokes                          53|      |  2   12 |        |             2          |     |3444|  44      4   |    |       4     |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  42 Calliergonella cuspidata (Hedw.) Loeske        25|      |  2      |        |4 222 2 2 1223  2 22  23|     |    |4 4 2232  2432|23  |      2 4 2 2|              |
| 122 Equisetum palustre L.                          25|      |  222    |     2  |3 223 4  222232  123 214|     | 1 3|414 224    432|    |       12 2  |              |
|  44 Caltha palustris L.                            25|      |  2  1   |     2  |2 223131   311  1 22 2 3|     |  23|3 433323  4333|    |      2   2  |              |
| 355 Valeriana dioica L.                            25|      |         |     2  |2  1222   32 2  2 12 2 2|     |    |3 3  21    231|    |          1  |              |
| 252 Potentilla erecta (L.) Raeusch.                25|      |313221223|     12 | 22232312 3222222 232 23|     |   2|3222 22  23   |2   |      1      |              |
|  89 Crepis paludosa (L.) Moench                    25|      |     1   |        |3 21 1222 3222    22 213|     |    | 14 122    3 2|    |             |              |
| 344 Trifolium pratense L. subsp. pratense          25|      |  2      |        |3222 23 1 2232  2 2  122|     |    |3 4 32      32|    |             |              |
|  96 Dactylorhiza maculata (L.) SoÛ                 25|      |  1  1  1|        |2 1222212 2 1 1 1  1   2|     |   1|     1    12  |    |             |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 147 Galium uliginosum L.                           15|      |         |        |                        |     |    | 1 21 11   121|    |        2    |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  47 Campylium stellatum (Hedw.) J. Lange & C. Jen 31|      |         |        |4322242 232  2 3 123 213|     |    |4    2     2  |2223|      1 3  1 |              |
|  66 Carex panicea L.                               31|      |   2     |        |23222332232321324 2234122|     |   2|4122133     2 | 223|            2|              |
|  60 Carex hostiana DC.                             31|      |         |        |23 2    323 122  314 213|     |    |              |313 |             |              |
| 175 Juncus alpinoarticulatus Chaix                 31|      |  1      |        |4222 212 4 232123 12113 |     |    |4   123       |    |            1|              |
|  58 Carex flacca Schreb.                           31|      |         |        |13 3 2 1213213131 3311 3|     |    |2          1  |1 1 |             |              |
|  54 Carex flava aggr.                              31|      | 12      |        |3322242132322313 22412 3|     |    |3 1  22    1  |2   |      1      |              |
| 258 Prunella vulgaris L.                           31|      |         |        |222223223 222 21221 1122|     |    |2   222      1|    |      2      |              |
| 187 Linum catharticum L.                           31|      |         |        |12221111   1111 211 1 1 |     |    |              |  1 |             |              |
| 347 Trollius europaeus L.                          31|      |         |        |2 324222  22321  12 1213|     |    |4122 21    3  |    |             |              |
|  34 Briza media L.                                 31|      |         |        |122222122 32212 1 121122|     |    |2 1 211     2 |    |             |              |
| 194 Lotus corniculatus aggr.                       31|      |         |        |  3212221 21 11 211  223|     |    |3 2  2      2 |    |             |              |
| 217 Parnassia palustris L.                         31|      |    1  1 |        |22311  1 2  1312221 1 12| 2   |    |2          1  |1   |             |              |
| 126 Eriophorum latifolium Hoppe                    31|      | 1       |        |2322 1 2 24 141 31234  2|     |    |  3  2        |    |             |              |
|  93 Dactylorhiza fistulosa (Moench) H. Baumann &  31|      |         |        |2212 11111 11 1 1 1   1 |     |   1|4 1 12     1  |    |             |              |
| 157 Gymnadenia conopsea (L.) R. Br.                31|      |         |        |1  1 1 22  1121   1 1122|     |    |2             |    |             |              |
| 183 Leontodon hispidus L.                          31|      |         |        |223134223 2 2 12 32 2112|     |   1|3 3 12        |    |             |              |
|  53 Carex davalliana Sm.                           31|      |         |        |333333323233242222334224| 23  |    |4 3 221    3  |    |             |              |
| 652 Carex ferruginea Scop.                         31|      |         |        |2 3442413   3242 4  3313|     |    |2    3     3  |    |             |              |
|  90 Ctenidium molluscum (Hedw.) Mitt.              31|      |         |        |12 434  2 33332  23312 2|     |    |              |    |             |              |
| 522 Primula farinosa L.                            31|      |         |        |232 24 2   1 222232 2  2|     |    |              |    |             |              |
| 336 Tofieldia calyculata (L.) Wahlenb.             31|      |         |        |2222111222312122 2131111|     |    |     1     1  |    |             |              |
|  22 Aster bellidiastrum (L.) Scop.                 31|      |     1   |        |3223231231311 212 241223| 32  |    |  1  3     4  |    |             |              |
| 382 Bartsia alpina L.                              31|      |         |        |2 22322121 22 22 12 22  | 11  |    |3    1     1  |    |             |              |
| 490 Selaginella selaginoides (L.) Schrank & Mart.  31|      |         |        |1 1212   1  1111 21 1 1 | 1   |    |              |    |             |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 136 Filipendula ulmaria (L.) Maxim.                16|      |         |     1  |           1            |     |    |332322214   33|    |     334     |              |
| 246 Polygonum bistorta L.                          16|      |         |     1 1|           2            |     |    |3 4333 1    33|    |      22     |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 139 Frangula alnus Mill.                           18|  4   |         |  1  1  |                        |     |    | 3            | 133|             |1             |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 554 Hydrocotyle vulgaris L.                        17|      |         |        |                        |     |    |              |  21|1            |              |
|  28 Betula pendula Roth                            17|      |         |        |                        |     |    |              |  13|             |              |
| 229 Pinus sylvestris L.                            17|      |         |        |                        |     |    |              |  11|             |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 541 Cladium mariscus (L.) Pohl                     80|      |         |        |                        |     |    |              |1342|4       4    |              |
| 270 Rhamnus cathartica L.                          80|      |         |        |                        |     |    |              | 11 |             |              |
| 579 Schoenus nigricans L.                          80|      |         |        |                        |     |    |              |44 2|             |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 536 Carex acutiformis Ehrh.                        35|      |         |        |                        |     |    |              |    | 224 344  3  |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 203 Lysimachia vulgaris L.                         72|      |   2     |        |                        |     |    |              |   1|1 1 12232112 |       1      |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 533 Calystegia sepium (L.) R. Br.                  47|      |         |        |                        |     |    |              |    |12   21   2  |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 224 Phragmites australis (Cav.) Steud.             46|      |         |4  3    |              2         |     |    |         1    |21 2|344243  13442|31421444444444|
| 204 Lythrum salicaria L.                           46|      |         |        |                        |     |    |         1    | 111|121 1   11  1|       1      |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
```

- The *selective sample* allowed for a mean square contingency coefficient of $C = 0.406$, indicating very high concentration of species scores. The first three Eigenvalues accounted for the following explanatory power: $\lambda_1 = 19.15\%$, $\lambda_2 = 15.03\%$ and $\lambda_3 = 13.47\%$.

The resemblance pattern of the random subsample is closer to linearity, as can be seen in the Eigenvalues. As a result of this, more non-zero scores are involved in the structured part of Table 11.1 than in Table 11.2. In the latter a source of nonlinearity is the distinct group pattern, as seen in Figure 11.8. However, neither table is really convincing in documenting the vegetation pattern. Clearly, there are not enough species involved to display the very high-dimensional similarity space. For example, in Table 11.2 there is only one species (*Pragmites australis*) identified as typical for the last group (with label no. 1). As a straightforward conclusion it can be suggested that synoptic tables might be most useful for displaying consistent parts of large data sets rather than providing an overview when the range of the vegetation gradient is large and dimensionality of the similarity pattern is high.

12
Swiss forests: A case study

12.1 Aim of the study

This case study illustrates some applications of the methods explained in the preceding chapters. It is a ‘real world’ example assessing ecological and also methodological questions. The vegetation data originate from a survey across Switzerland (Wohlgemuth *et al.* 2008) which aimed to reveal relationships between vegetation composition and the growth rate of tree species. Just as in an experimental approach, questions were posed prior to the investigation and a sampling design is developed below to obtain the answers. But as is often the case in large surveys the order of steps is partly reversed, as in some examples the data appear at the outset of the exercise and the questions that they could potentially answer are identified later. This imposes restrictions on the analytical methods, but it also allows exploration of the variable set far beyond its primary scope.

Data Analysis in Vegetation Ecology Otto Wildi

The analyses shown below concentrate on species composition and species spatial distribution. An alternative would be to focus on species richness, as done by Wohlgemuth *et al.* (2008), who found that there is a high correlation between diversity (the number of species per plot) and the canopy cover of trees (a surrogate for light availability inside the forest stands). However, they also detected that the spatial resolution of the survey is probably not the best considering biodiversity, as correlations become higher after clumping the sampling units into landscape patches of about 100 km^2 in size.

Hence, the scale as well as the variables chosen in sampling design determine the potentials and restrictions of application. As will be shown below the sampling area is the territory of Switzerland (∼41000 km^2), restricted to its forested area (∼30% of the surface). The strong elevation gradient is the main cause of spatial variation in climate, and climatic relationships are therefore an issue. Concentrating the study on forests means that human influence is weak and controlled, a consequence of the strong regulations imposed on forest management by public law. The question of the strength of human disturbance will be raised below. Due to the extent of the study area, traces of the post-glacial history of the vegetation can be expected, as well as patterns related to the diversity in geology; that is, the parent material for soil formation. There is no explicit temporal information included in this data set. Yet, the fact that young trees (seedlings and shrubs) are distinguished from grown-up trees opens a window on change in time. Variables describing the soil conditions and the geological pattern are as yet scarce. The statistical analysis should reveal whether these are needed to reveal the relationship between vegetation and site factors.

12.2 Structure of the data set

The vegetation data are organized in the tradition of forest ecology: each tree species is recorded three times (in the herb, the shrub and the tree layer). Shrubs are recorded twice (herb and shrub layer) and herbs and mosses just once. Furthermore, all locations are sampled three times within concentric circlic plots of 30 m^2, 200 m^2 and 500 m^2 respectively. Most of the examples shown below use the 200 m^2 data set.

In the environmental data sets described below three categories of variables are to be found. First, variables recorded or verified in the field: elevation, slope and pH of the upper soil layer. Second, climate data interpolated from meteorological stations, as described by Zimmermann & Kienast

(1999). Third, ‘azonal variables’ taken from the Swiss Soil Quality Map, interpolating properties of soil types to yield a rank scale with range 1–6. Quality is limited due to the low resolution of the original map (Scale 1:200000) and the arbitrarily selected soil types. These variables are labeled with ‘soil map’. Environmental factors included in this data set are the following (order according to field records, climatic data and data from soil map):

Sampling plan (x-, y-coordinates). The sample is a subset of the grid used in the Swiss National Forest Inventory (NFI, Brassel & Brändli 1999). Sampling units are located at each intersecting point of the 4 km × 4 km coordinate grid of Switzerland. Only forest stands are taken (definition of NFI). The resulting sample size is $n = 726$ and it represents an unbiased state of the Swiss forests. True locations of plots deviate by 5–30 m from the grid: the precision of the navigation tools used at the time was rather limited.

Elevation (m a.s.l.). Two variables are given for technical considerations. One stems from the field survey, the other is derived from a digital terrain model (DTM). The range of the elevation scores does not reflect the Swiss topography as plots above the timber line are not part of the sample.

Slope (degrees). This is derived from the digital terrain model and therefore it is affected by errors inherent in this.

pH upper soil layer. Samples of the upper soil layer were taken in the first survey of the Swiss National Forest Inventory (Brassel & Brändli 1999). I replaced 43 missing values with the mean of the sample (pH 5.1095).

Degree days ($^{\circ}C$) x days. Daily temperature is interpolated from a set of climatic stations (Zimmermann & Kienast 1999). Degree days are the integral of the daily mean temperature curve above the zero line.

Yearly precipitation (mm). Precipitation exhibits higher yearly spatial variation than, for example, temperature. Accordingly, interpolated variables are also less reliable.

Frost days during growing season. On frost days the night temperature drops below the freezing point. The growing season lasts from March until the end of September.

Coldest mean monthly temperature ($^{\circ}C$). This is a proxy for the risk of frost drought. Trees are exposed to this, whereas plants below the snow cover are protected. Coldest month usually is January.

Mean yearly global radiation. This is the sum of direct and diffuse radiation over the entire year.

Yearly water balance (mm x $yr^{-}1$). The monthly water balance – water gain by precipitation minus water loss by evapotranspiration – is integrated over the whole year.

Moisture index, i.e. water balance in July (mm x $yr^{-}1$). This is the water balance as explained above, but for July only. July is the warmest month in Switzerland.

Soil depth (soil map). This is estimated on a 1–6 scale addressing suitability for agriculture.

Nutrients (soil map). A 1–6 scale expresses availability of main plant nutrients.

Water capacity (soil map). This is the maximum possible water holding capacity estimated on a 1–6 scale.

Water permeability (soil map). Permeability estimated on a 1–6 scale.

Soil wetness (soil map). Excess average water content to limit agricultural use, estimated on a 1–6 scale.

12.3 Methods

All calculations are based on the same data transformation: Braun-Blanquet code transformed to a rank scale and further according to $x' = x^{0.25}$ (see Table 3.3). This is the basis for ordinations and classifications. The standard ordination method is principal coordinates analysis (PCOA, Section 5.3) and for classification minimum-variance clustering is used (Section 6.3). The number of groups (vegetation types) chosen to illustrate the examples is eight, a low number considering the sample size of $n = 726$. This, however, simplifies presentation of results in tables (Table 12.1), ordinations (Figure 12.1) and maps (Figure 12.2). VEGEDAZ (Küchler 2009) was used

Table 12.1 Composition of eight vegetation types in terms of tree layer and some site factors. Rows and columns are rearranged according to the first axis of a correspondence analysis.

	Group no.	1	3	7	8	6	2	5	4
	Group size	135	119	45	83	46	38	171	89
	Elevation, mean	1690	1260	1380	1100	674	690	729	649
	stdv.	263	368	114	187	209	242	230	162
	Degree days, mean	1660	2180	1910	2320	2910	3450	2810	2920
	stdv.	377	527	137	300	298	477	363	250
	Precipitation, mean	1420	1320	1690	1610	1260	1730	1320	1230
	stdv.	331	357	261	241	238	216	253	212
	pH, mean	4.23	5.61	5.42	5.00	3.68	4.35	5.70	5.65
	stdv.	1.15	1.20	1.13	1.25	0.64	0.89	1.27	1.20
738	*Pinus cembra*	17	1						
732	*Pinus mugo ssp. arborea*	3	3						
735	*Pinus mugo ssp. prostrata*	1							
210	*Sorbus aucuparia*	7	11	12	9			5	2
75	*Alnus incana*	2	4	18	3		6	5	
32	*Picea abies*	72	75	78	76	87		58	53
216	*Sorbus aria*	1	12	3	2		19	5	20
69	*Betula pendula*	3	9		8	3	14	3	
51	*Salix caprea*	3	6		3		3	2	2
280	*Acer pseudoplatanus*	3	11	29	25	9	6	44	34
29	*Abies alba*	2	21	20	52	61	3	41	31
35	*Larix decidua*	52	20		3	9	3	7	3
84	*Fagus sylvatica*	3	16	9	75	70	35	74	95
371	*Fraxinus excelsior*	1	8	3	14	7	45	46	38
1043	*Quercus pubescens*		2				3		

(continued overleaf)

Table 12.1 (*continued*)

38	*Pinus sylvestris*	6	21			9		9	22
93	*Ulmus glabra*		1		4			9	7
755	*Castanea sativa*	1				3	74	1	2
303	*Tilia cordata*		1			3	14		2
72	*Alnus glutinosa*						8	1	
87	*Quercus petraea*		2			5	22	4	8
232	*Prunus avium*		3			5	16	9	12
43	*Taxus baccata*							2	2
274	*Acer platanoides*							4	11
300	*Tilia platyphyllos*							9	8
78	*Carpinus betulus*					3	3	7	11
90	*Quercus robur*					11	6	9	18

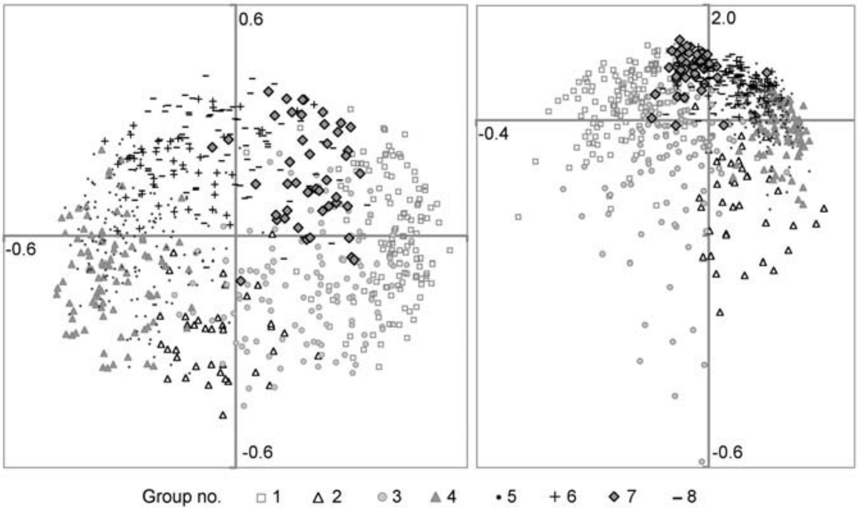


Figure 12.1 Principal coordinates analysis (left) and correspondence analysis (right) with eight vegetation types of the Swiss forest data set overlayed.

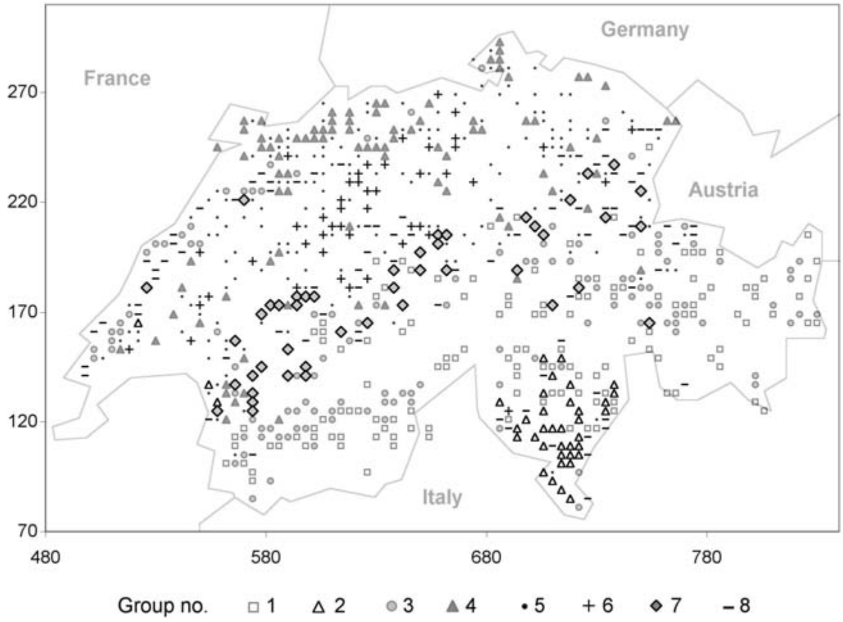


Figure 12.2 Vegetation map (eight groups from cluster analysis).

to update species lists as well as to merge data sets. Separate data sets were derived for plot sizes of 30 m^2, 200 m^2 and 500 m^2 ($n = 726$) and for a merged one with $n = 2178$. A data set with reduced attribute list of 29 tree species is the source for Table 12.1, derived from the 200 m^2 plot-size data set, omitting tree species which are very rare or have been introduced from gardens and parks.

12.4 Selected questions

12.4.1 Is the similarity pattern discrete or continuous?

The worst case scenario in terms of pattern recognition is patterns caused by the sampling plan itself, as may happen in preferential sampling. This, however, is unlikely to occur in the systematic sampling plan used here. One could of course argue that discrete patterns are unlikely to occur because the territory of Switzerland encompasses a huge altitudinal gradient, which is continuous by nature. But there are other factors with the potential to generate discontinuity, such as bedrock type and forest management. In order to explore the resemblance pattern I use classification and ordination applied to the entire data set, as searching for local patterns would probably require the analysis of subsets.

In Figure 12.1 classification (eight types) is superimposed on ordinations by PCOA (left) and CA (right), both confirming continuity of pattern. Unlike in the application to small data sets (Figures 5.6 and 5.7), differences between the two methods here are striking. In PCOA strong nonlinearity is responsible for a round point cloud, whereas CA generates a horseshoe (left to right) with outliers towards the bottom of the graph. In conclusion, strong resemblance gradients prevail but even faint traces of discontinuities are lacking.

12.4.2 Is there a scale effect from plot size?

All locations have been surveyed threefold using concentric plots of size 30 m^2, 200 m^2 and 500 m^2 respectively. When plot size is increased, generally the number of species increases. Do additional species found in large plots add to the distinction of relevé groups (vegetation types)? Could different plot sizes yield an alternative classification of forest types? A joint analysis of plots of different size is needed to explore these questions, for which data sets have to be merged. To reveal patterns caused by difference in species richness, adjustments of relevé vectors have to be avoided, by using Euclidean distance as a resemblance measure and PCOA for ordination for example, as shown in the upper row of Figure 12.3. Clearly, plots of size 30 m^2 generate the smallest point cloud, those of 500 m^2 the largest. When using correlation instead of Euclidean distance, the effect of plot size vanishes doe to the intrinsic standardization of the relevé vectors (Figure 12.3, middle row). Because intrinsic transformation in correspondence analysis is distinct from these two examples, I also provide CA ordinations of the combined data (Figure 12.3, bottom row). Careful inspection of the three ordinations reveals a shift along the first axis, with plots from larger size being located further to the right. In summary, differences in plot size emerge when using Euclidean distance for comparison and disappear when using correlation. Interpretation of CA ordination remains difficult because the solution neither suppresses the effect completely nor reveals it to its full extent.

Ordinations shown in Figure 12.3 also illustrate differences in methods when applied to large data sets. This concerns not only the shape of ordinations but also the performance of axes in accounting for variance. For the upper row (PCOA applied to resemblance matrix of Euclidean distance), explained variance is $\lambda_1 = 7.2\%$ and $\lambda_2 = 5.4\%$. When using correlation

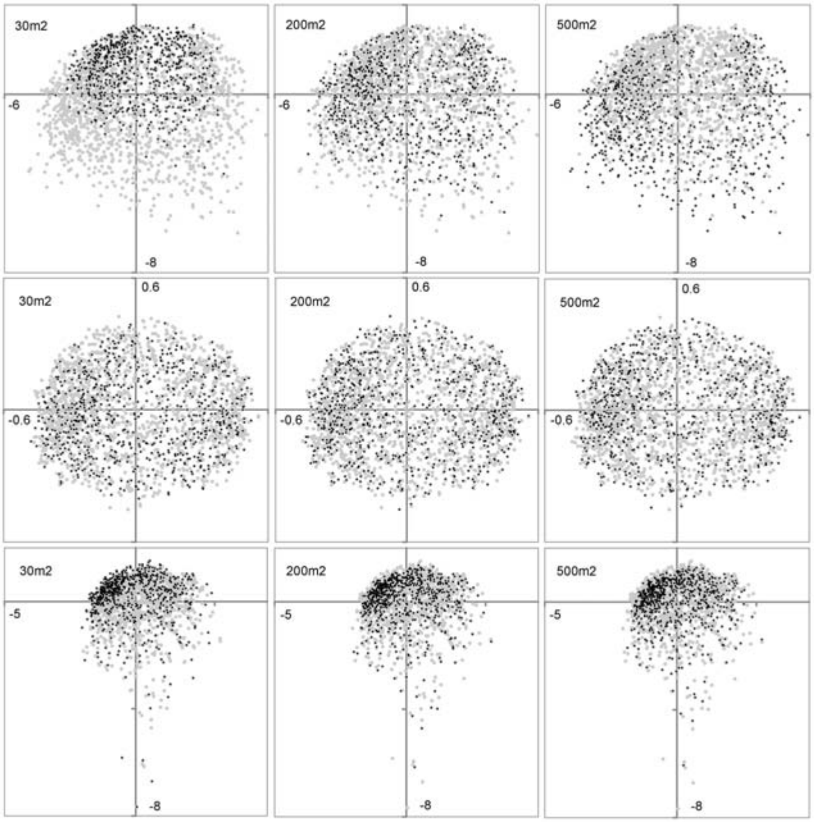


Figure 12.3 The effect of different plot sizes on similarity pattern. Upper row: Euclidean distance used in PCOA. Middle row: correlation coefficient used in PCOA. Bottom row: CA used.

instead of Euclidean distance we get $\lambda_1 = 6.3\%$ and $\lambda_2 = 5.4\%$, whereas for CA it is an astonishingly low $\lambda_1 = 2.2\%$ and $\lambda_2 = 1.5\%$.

12.4.3 Does the vegetation pattern reflect the environmental conditions?

This conforms with the main question asked in Chapter 7. Because the present survey concerns a large, well-known area, an effort to interpret an ordinary vegetation map like the one in Figure 12.2 seems promising. Even though the map considers discrete types only, some of the symbols accord with altitudinal zones: along the upper-left part in Figure 12.2 the Jura mountains, ranging from south-west to north-east, can easily be identified. Dominating groups are numbers 8, 3 and 4. Parallel to this follows the Plateau (Mittelland), where types 5 and 6 prevail. In the Pre-Alps groups 7 and 8 are most common, while the high alpine zone is mainly above the timber line. Embedded in the Alps are Central alpine valleys, the driest locations in the investigation area with the highest abundance of group 1. The sequence ends in the south where relevé group 2 corresponds with the Insubrian climate.

Maps revealing far more details are continuous, although they require separate layers for every vegetation type or species considered. I choose a Bayes probability model deriving probability maps of vegetation types (see Section 8.2 for details). This is not used as a simulation tool but as a method to spatially display part of the joint variance of vegetation and environmental factors. For simplicity – and to keep the number of degrees of freedom as low as possible – I used four of these factors (see Table 12.2 for the full list): (1) elevation taken from DTM, (2) degree days, (3) yearly precipitation and (4) pH of the soil. As a measure of performance of the model I counted the number of plots where the model predicts the same vegetation type as the field survey. This is 381 of the total of 726, or 52.4%. I found this a surprisingly high match considering human impact on forest stands and the simplicity of the model (see also Brzeziecki *et al.* 1993). The resulting maps are shown in Figure 12.4, where circle diameter is proportional to the probability of occurrence. Symbols occur on forested plots only; that is, about 30% of the area of the country. In general the spatial distribution pattern accords with the oreographic pattern dominated by the altitudinal gradients. Patterns related to geology are rare and their recognition would require more detailed analysis.

Whether there is a significant relationship between site factors and vegetation groups can be tested using the available data. The eight groups used for mapping form the basis for variance testing of the site factors as explained in Section 7.2. The results are shown in Table 12.2. Obviously, the forest types greatly reflect the abiotic environmental conditions; that is, forest management seems to be less important – or foresters have planted tree species within their ecological range.

12.4.4 Is tree species distribution man-made?

This question is related to the well-known ‘Potential Natural Vegetation’ (PNV) issue introduced in a systematic manner by Tüxen (1956). His fairly complicated definition is further extended in Lindacher (1996). Lindacher

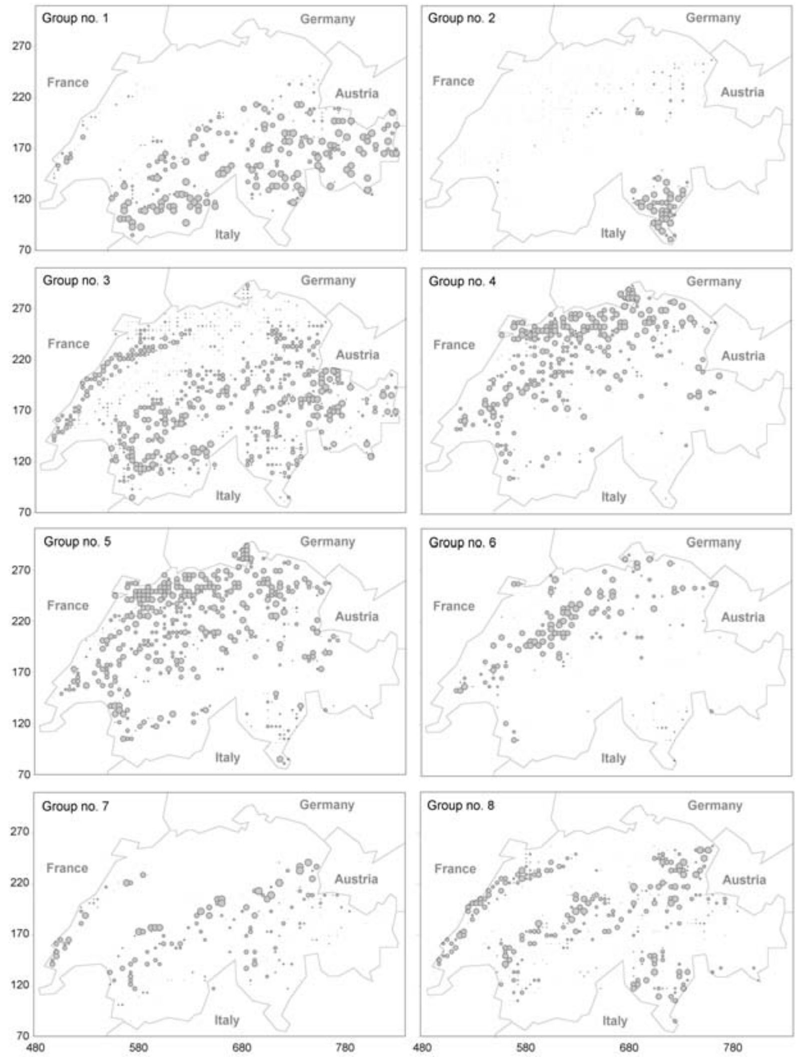


Figure 12.4 Vegetation probability map (eight groups from cluster analysis distinguished).

also mentions extensions of the concept to adapt for effects like climate change and environmental pollution, addressed in the definitions of Kowarik (1987) for example. In many forests human influence dominates a given site and the vegetation may be *anthropogenic* rather than natural (Küchler 1988). One may want to reconstruct the vegetation as it would be without *Homo sapiens* living on earth, as Neuhäusl (1984) explained.

Table 12.2 F-values of site factors based on eight forest vegetation types. $Df1 = 7$, $df2 = 721$. All F-values are significant at the 1% error probability level.

Variable	F-value
Elevation, DTM	253.51
Elevation, field	246.60
Degree days	207.87
Lowest monthly temperature	182.18
y-coordinate (N–S gradient)	78.660
Soil depth (soil map)	69.464
Nutrients (soil map)	66.942
Water capacity (soil map)	60.254
Soil skeleton (soil map)	45.251
Moisture index July	43.839
Slope, deg.	39.100
pH (upper soil layer)	35.563
Yearly precipitation	30.341
Water permeability (soil map)	30.774
Frost days in growing season	27.338
Yearly water balance	19.300
x-coordinate (E–W gradient)	13.796
Yearly radiation	13.307
Soil wetness (soil map)	8.653

In all references given above there is a general agreement that PNV could only be derived from a land-use history of the past few thousand years. As this is not known, all findings will finally remain hypotheses. Furthermore, tree species planted within their natural range do not in principle change the entire vegetation composition. Hence, I restrict the evaluation to the discussion of four selected tree species: *Fagus sylvatica*, *Fraxinus excelsior*, *Larix decidua* and *Castanea sativa*. If the occurrence of any of these is man-made, this should become evident in either the *geographical*, the *compositional* or the *ecological* distribution pattern.

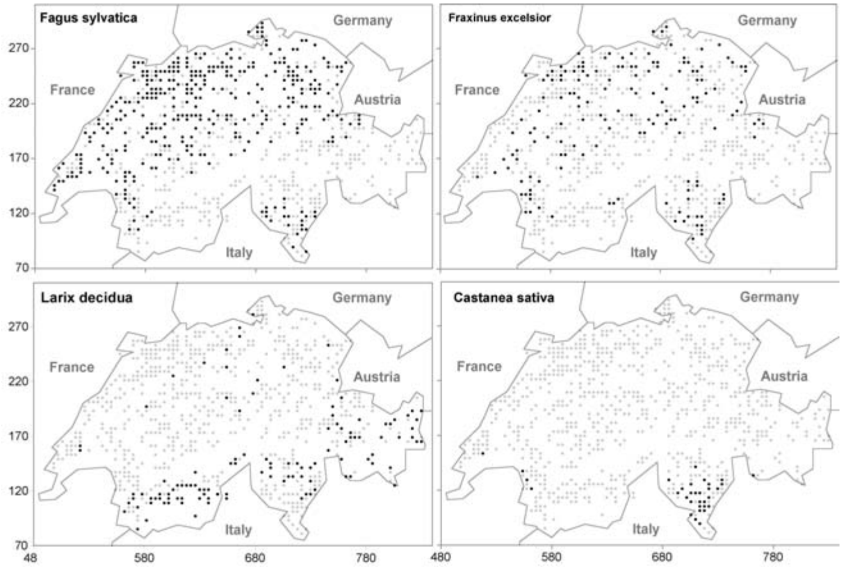


Figure 12.5 Distribution of four selected tree species occurring in the tree layer.

Can tree plantations be seen in the *geographical* distribution pattern? It is known that some tree species have been planted outside their ecological range. Can this be observed in the present data set? Yes, it can. In Figure 12.5 *Larix decidua* is such a species. Under natural conditions it mainly grows in the central alpine belt from the west to the east of the country. Isolated, scattered plots north of this belt mark the locations where plantations may have been made in the past. However, this is an interpretation, not a proof.

Even more striking is the pattern of *Castanea sativa*. This is restricted to southern Switzerland, but a few clumped plantations in the western part of Switzerland emerge in the survey grid. Are there cases where this method fails? No, not really. *Fraxinus excelsior* also grows in *Quercus* forests, which are too dry for *Fagus*, for example. However, (like *Fagus*) it does not grow at higher altitudes, which are outside its physiological range (Figure 12.5). It cannot be planted there, so no wonder that plantations are not found.

Can tree plantations be inferred from the *compositional* pattern? Under natural conditions it is rather unlikely that any plant species would exhibit a disjunct compositional distribution pattern. Where such a phenomenon occurs it is likely that plantations have taken place. In Figure 12.6 the

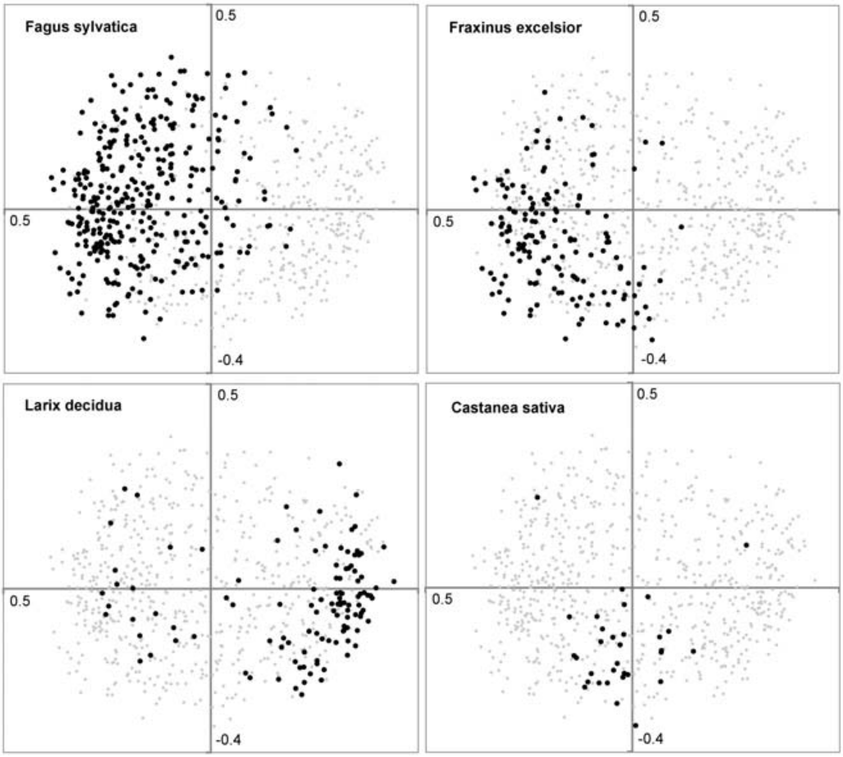


Figure 12.6 Ordination of forest stands. Four selected tree species marked.

occurrence of the same species within an ordination is shown. In *Larix decidua* and *Castanea sativa* a compositional centre of occurrence exists, again with some scattered remote points: in these the two trees grow in common with a totally different set of plant species, supporting the evidence derived from the geographical pattern that plantations exist. Unlike in the geographical map, for *Fraxinus excelsior* a centre of occurrence is now visible. It is the lower-left quadrant of the ordination. In all remaining sites of the ordination it is absent. This supports the hypothesis that *Fraxinus excelsior* stands are either natural or are planted where they would occur under natural conditions. The same interpretation applies to *Fagus sylvatica*.

Can tree plantations be inferred from the *ecological* pattern? Ecograms offer a similar kind of interpretation to maps and compositional ordinations, as shown in Figure 12.7. These point patterns show the climatic conditions below the timber line. There is a striking *main gradient* from the lower-right corner (warm, dry) towards the upper-left (cold, wet). Along this gradient,

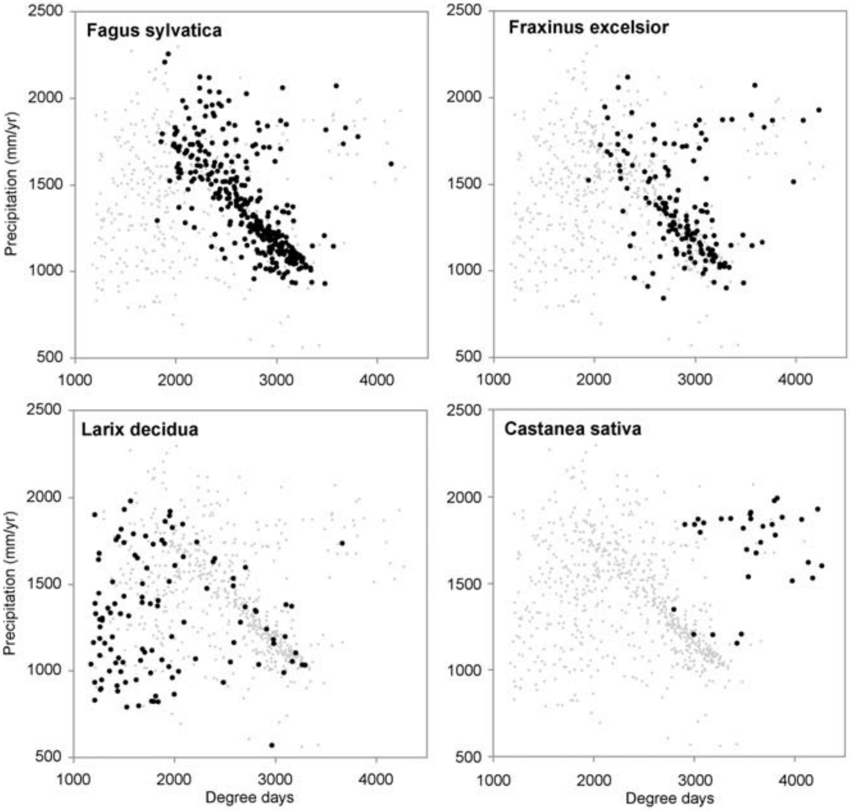


Figure 12.7 Ecograms of forest stands. Four selected tree species marked.

warmth and water supply are strongly correlated. But there are two areas following a different pattern. The first is the left-hand lower edge of the ecogram. These are the central alpine regions, with low temperature and dry conditions. The second is the upper-right corner, where the opposite holds and it is warm and wet. This is the climate of southern Switzerland, the Insubric area.

Looking at the same tree species as before, *Fagus sylvatica* spreads over the main gradient and also in the direction of the Insubrian conditions, avoiding the central alpine growth conditions. That is where *Larix decidua* has its centre of distribution. If the previously stated hypothesis is correct then the few locations along the main altitudinal gradient are plantations (including one isolated stand in the Insubrian climate). Even more striking is the pattern of *Castanea sativa*, with its centre of distribution in the Insubrian part of the ecogram. Five stands in the main altitudinal gradient are really disjunct and definitely artificial from the ecological point of view.

12.4.5 Is the tree species pattern expected to change?

This question is based on the idea that the next generation of tree species is already present in the form of seedlings and saplings. As can be seen in Table 12.3 the tree species behave totally differently in this regard. Some species with large seeds, such as *Abies alba*, *Pinus cembra* and *Fagus sylvatica*, show almost the same frequency for the tree and the herb layers. But most interesting are the species with a much higher frequency in the herb layer: these have the potential to extend their area in the face of changing climate much faster than species which must expand through propagation of seeds. Typical cases are *Abies alba*, *Quercus robur*, *Acer pseudoplatanus*, *Fraxinus excelsior* and *Sorbus aucuparia*.

In Figure 12.8 three species are compared in this regard. *Fraxinus excelsior*, with close to twice as many plots with seedlings than plots with trees, has almost identical distribution patterns for both. The seeds of this species are very mobile and are wind-dispersed, though the seedlings have not so far established in areas of different ecological conditions. Mobility of seeds, however, suggests that extension of area may still be rapid in the case of climate change. *Larix decidua* is the opposite case: rejuvenation is almost exclusively restricted to sites where it occurs naturally; where it is planted, for example along the main altitudinal gradient, seedlings and saplings are very rare. Finally, *Sorbus aucuparia* is rather abundant in the herb and shrub layers. The range of the seedlings exceeds that of the trees considerably. Under changed conditions *Sorbus aucuparia* could rapidly grow up to the tree layer, thereby expanding its present area. Along the main altitudinal gradient the species is mainly lacking in the tree layer, probably due to competition of taller growing trees. For it to expand successfully, other species would have to reduce vitality.

12.5 Conclusions

This chapter demonstrates the use of different methods in the context of a fairly large real-world data set resulting from a standard vegetation survey. Method and plot size conform to established standards; even the Braun-Blanquet cover-abundance scale is used. Less common, however, is the sampling plan, which consists of a regular net of 4 km grid width. Hence, the

Table 12.3 Number of plots where selected tree species occur in the tree, shrub and herb layers.

Species	Tree layer	Shrub layer	Herb layer
Abies alba	205	183	305
Acer platanoides	15	3	51
Acer pseudoplatanus	161	183	397
Alnus glutinosa	4	8	3
Alnus incana	26	30	35
Betula pendula	55	30	33
Carpinus betulus	22	25	35
Castanea sativa	32	18	35
Fagus sylvatica	345	293	348
Fraxinus excelsior	153	142	335
Larix decidua	113	52	52
Picea abies	471	362	409
Pinus cembra	23	24	26
Pinus mugo arborea	7	6	5
Pinus mugo prostrata	1	4	3
Pinus sylvestris	69	14	15
Populus nigra	1	2	3
Prunus avium	35	41	116
Quercus petraea	25	9	40
Quercus pubescens	3	4	5
Quercus robur	38	18	80
Salix caprea	16	57	75
Salix eleagnos	1	2	2
Sorbus aria	49	101	141
Sorbus aucuparia	43	127	320
Taxus baccata	4	10	8
Tilia cordata	8	7	15
Tilia platyphyllos	22	22	32
Ulmus glabra	24	47	58

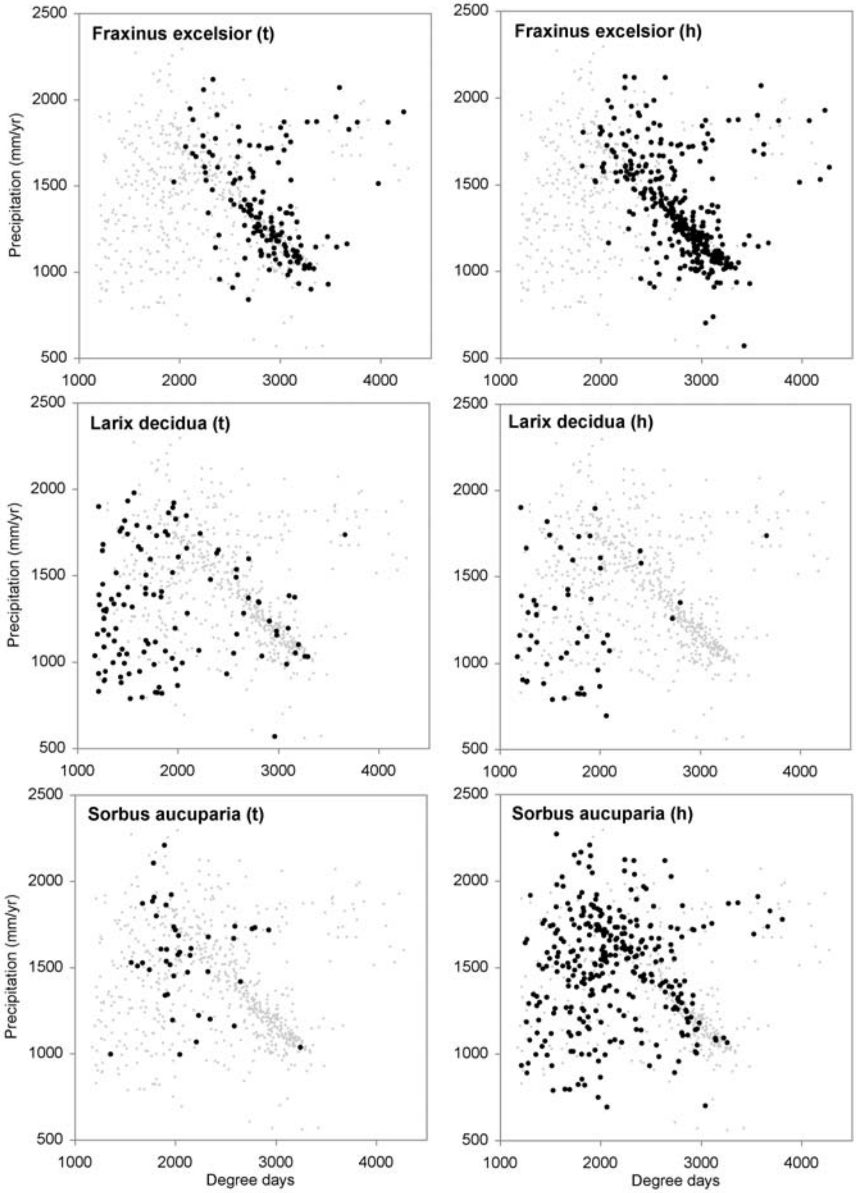


Figure 12.8 Distribution of three species in ecological space. Left: tree layer. Right: herb layer.

location of plots is given by the sampling plan and homogeneity – a standard requirement in classical phytosociology – is not an issue. But systematics is vital, in that many of the explorations rely on this sampling design. Geographical patterns could not be interpreted if preferential sampling had been used and the same is true for patterns in ecological and resemblance space. Patterns revealed are not 'representative' for the population just because of the sampling plan, but also due to the sample size. Given the topographical variability of the sampling area, $n = 726$ is probably still small. To reveal patterns of the kind shown in this chapter definitively requires large samples. Hence, the most striking conclusion does not concern methods but the quality of data sets required to allow pattern recognition and interpretation, which is out of the question when using preferentially sampled data only. There are no methods to correct for biased data sets.

Ordinations are much easier to interpret when classification is superimposed, even if this is not the scope of analysis. Although not shown here (but explained in Section 7.4), the opposite holds too: that is, the interpretation of classifications by ordinations, in which the data points are group centroids of relevés, revealing continuity among groups. Just like the axes in ordinations, groups may become reference points in a continuous system, shown here in the form of a continuous vegetation map expressing probability of occurrence of all types in all plots. Continuous maps can also be derived where only vegetation data are available, applying fuzzy classification, an as yet underexploited approach.

But which ordination and classification method should be used? Even though I prefer some (where I know exactly what the method does to my data) and try to avoid others (e.g. where iterations alter the results in order to eventually find a 'stable' solution), all those previously and recently used work very well in general, as long as they are handled with care. Accidents may happen if data transformation gets out of control: analysis of the effect of plot size is one example, where one resemblance measure reveals the difference, which is hidden by a different one. Hence, the selection of the ordination is not the crucial point, but rather the selection of the proper options. Computer programs always use default parameters, which by chance may be the ones needed – or may be the wrong choice. Escaping this pitfall is easy, through careful study of the options the software provides, getting familiar with the method in gathering experience and eventually doing the same with other software, if available.

Appendix A
On using software

In Section 3.1 I pointed out that we tend to measure depending on what measurement tools we have at our disposal, causing some bias in our sampling design. The same happens when choosing methods for data analysis, which are often dictated by the options the software packages offer. Clearly, single all-purpose programs for data input, testing for input errors, data manipulation, data analysis and graphical presentation are most practical, a nice example being CANOCO, as explained in Lepš & Šmilauer (2003). While this is the user-friendly method, my book cuts its own way by emphasizing specific issues (e.g. data transformation, measuring resemblance) and applications beyond standard methods (e.g. space-for-time substitution in Section 9.4.1, measuring rate of change of different orders in Figure 9.2, evaluating the second derivative of change in Figure 9.18, etc.). Multiple methods cannot be found in a single software package, such as programs for dynamic modelling (Chapter 10) and the use of Markov models (Section 9.3). Rather than giving specific suggestions, I disclose my path through the examples in this book, accepting the downside that this partly represents a historic travel through software development in recent times.

A.1 Spreadsheets

Almost all small data sets used to illustrate the functioning of methods eventually passed though Excel, for either data input, simple analyses or data presentation. In spreadsheets, data are usually organized as shown in Figure 2.8; that is, in a single data matrix with relevés organized in columns and species and environmental variables in rows. For import and export into

Data Analysis in Vegetation Ecology Otto Wildi

other programs I used commas, semicolons or tabulators separating data fields, which are accepted by almost all databases and statistical packages. The tasks I have conducted using spreadsheets are:

- Plotting scatter diagrams (ordinations). Excel offers sufficient formatting options for graphs and all are set automatically. Subsequent manual formatting is needed as ordination axes x and y must be identically scaled and this is achieved only when manually setting the range of the axes and the width and height of the graphs.

- Plotting bubble graphs to display within- and between-group similarities (Figures 4.5 and 11.8), illustrating change of similarity in time series (Figures 9.2) and ordinations of time series exhibiting velocity (Figures 9.16) or acceleration (Figures 9.18). Three vectors are needed for input: x-axis, y-axis and the diameter of the bubbles.

- Numerically integrating differential equations for dynamic modelling. All models presented in Section 10.1 are implemented in Excel. A column is chosen for each state variable; that is, $X1$, $X2$, as well as its derivative $\delta X1/\delta t$, $\delta X2/\delta t$. For time step $t = 0$ the initial values are written in a row; in the following row the formula for deriving the state at $t = 1$ is entered. Dragging these cells down by, for example, 100 rows reveals the state of the system at $t = 100$. Increasing complexity of dynamic models, however, limits the application of spreadsheets.

A.2 Databases

Databases are becoming indispensable when data sets of increasing size are analysed. JUICE (Tichý 2002) and TURBOVEG (Hennekens & Schaminée 2001) are probably the best known. I used VEGEDAZ (Küchler 2009), a program with features typical for the purpose. The functions mentioned below refer to this, but all databases designed for handling vegetation data do so similarly:

- There are functions to test input data, such as format of date, double entries in relevé and species labels, range of variables, double entries of entire relevés and species vectors, identification of outliers (Section 11.3) and so on.

- There are tools for data exploration, mainly options for display of scatter plots and frequency diagrams.

- Further, one of the functions concerns the data selection used for example when double entries have to be erased. Subsets may be generated based on stratification criteria (Section 2.3.2). Reduced samples are derived using systematic or random subsamples, as shown in Section 11.2.

- The program also supports handling of taxonomic problems existing in all fields of organismic biology (plants, animals, fungi, algae, bacteria, etc.). For all species names, alternative names are displayed for selection; names form different taxonomies – outdated as well as valid ones. This function is essential for combining data sets, as in Chapter 12.

- Further support is given by providing auxiliary variables, mainly valid for the geographical area considered, such as political borders, rivers, digital terrain models and, in the present case, indicator values of Central Europe (Landolt 1977), as used in Section 11.4.

- The program package is also an analytical tool. Many basic functions used throughout this book are implemented in the program menu. Furthermore, there is an interface to the R statistical package, where, for example, all functions of the VEGAN package (Dixon 2003) and others are available. Output of R programs, like ordination coordinates and classifications, can be embedded into the data set being analysed.

A.3 Software for multivariate analysis

There are several specialized program packages available for multivariate analyses, of which I mention PC-Ord, CANOCO and SYN-TAX2000 (Gilliam & Saunders 2003). I consider my own program package MULVA-5 (Wildi & Orlóci 1996) partly outdated as many of the ordination and classification methods used these days are missing. However, I implemented operations presented in this book into MULVA using FORTRAN code:

- Computing within- and between- similarity of relevé groups in Section 4.6.

- Ordering synoptic tables (Section 6.6).

- Computing directed spatial dependence (Section 7.3.3).
- Implementing a simple Bayes probability model (Section 8.2).
- A means to extract rate of change of different order (Figure 9.2).
- Smoothing Markov series (Section 9.3).
- Deriving synthetic time series via space-for-time substitution (Section 9.4.1).

The space–time model of succession in the Swiss National Park (Section 10.3) is a standalone FORTRAN program as complexity exceeds flexibility of multipurpose programs. Packages for dynamic modelling exist, such as STELLA and SIMILÉ, generating flow charts as shown in Figure 10.1 upon input of differential equations and providing a flexible graphical output.

In many computer programs the default options used are hidden in program descriptions or parameter lists; this is one of the flaws of R procedures. It is good practice to compare output from different programs to avoid misunderstanding; I did so when preparing examples, using PC-Ord in parallel with other programs.

The number of methods available in the R computing environment is growing quickly and the VEGAN package (Dixon 2003) offers many methods previously available only in specialized program packages. I used R procedures for constrained ordination in Section 7.5 and also to check other ordinations processed in MULVA. In view of the rapidly growing specific functions the computing environment R offers, this may soon become the dominating platform for data analysis in vegetation ecology. Rumours confirm detailed instructions to be underway.

Appendix B

Data sets used

Most of the data sets used in this book are available on the Internet (http://www.wsl.ch). All are text files with UNIX line endings. These may have to be adjusted depending on the operating system and software used. The data sets are provided in two different formats:

- Extension .txt refers to the matrix arrangement as shown in Figure 2.8
- Extension .m5 indicates MULVA-5 organization (Wildi & Orlóci 1996).

Name	Dimensions	Species scores	Reference	Comments
nzz	11 relevés 21 species 7 site factors	Braun-Blanquet *blank*, +, 1, 2, 3, 4, 5	Wildi & Orlóci 1996)	Artificial data x-, y-coordinates, time (yr)
nzzt	11 relevés 21 species 7 site factors	Braun-Blanquet Rank scale Square-root transformed	Wildi & Orlóci (1996)	The same as nzz, but numerical
Schlaenggli	63 relevés 119 species 22 site factors	Braun-Blanquet *blank*, +, 1, 2, 3, 4, 5	Wildi (1977)	Wetland data x-, y-coordinates
Schlaengglit	63 relevés 119 species 22 site factors	Braun-Blanquet *blank*, +, 1, 2, 3, 4, 5	Wildi (1977)	The same as Schlaenggli, but numerical data
Lipperaw	19 relevés 9 species	Cover %	Lippe *et al.* (1985)	Adjusted to 100% cover, raw data
Lippersim	19 relevés 9 species	Cover %	Lippe *et al.* (1985)	The same as Lipperaw, but simulated
Snpser59	751 relevés 6 species guilds	Cover %	Wildi & Schütz (2000)	59 plots, SNP five-year time steps

References

Achermann, G., Schütz, M., Krüsi, B.O. 2000. Tall-herb communities in the Swiss National Park: long-term development of the vegetation. *Nat Park-Forsch Schweiz* **89**: 67–88.

Allen, T.F.H. and Starr, T.B. 1982. Hierarchy: perspectives for ecological complexity. The University of Chicago Press.

Anand, M. 1997. The fundamental nature of vegetation dynamics: a chaotic synthesis. *COENOSES* **12**(2–3): 55–62.

Anderberg, M.R. 1973. Cluster analysis for applications. Academic Press, New York, San Francisco, London.

Austin, M.P. 2005. Vegetation and environment: discontinuities and continuities. In: Maarel, van der, E. (ed.) Vegetation Ecology. Blackwell Publishing, Malden, Oxford, Victoria. 52–84.

Batschelet, E. 1975. Introduction to Mathematics for Life Sciences. 2nd ed. Springer-Verlag, Berlin, Heidelberg, New York.

Belden, A.C. and Pallardy, S.G. 2009. Successional trends and apparent *Acer saccharum* regeneration failure in an oak-hickory forest in central Missouri, USA. *Plant Ecology* **204**: 305–322.

Benzécri, J.P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: S. Watanabe (ed.) Methodologies of Pattern Recognition. Academic Press, New York. 35–60.

Boyce, R.L. and Ellison, P. 2001. Choosing the best similarity index when performing fuzzy set ordination on binary data. *Journal of Vegetation Science* **5**: 439–440.

Box, E.O. 1996. Plant functional types and climate at the global scale. *Journal of Vegetation Science* **7**: 309–320.

Bradfield, G.E. and Kenkel, N.C. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. *Ecology* **68**: 750–753.

Data Analysis in Vegetation Ecology Otto Wildi

Brassel, P. and Brändli U.B. (eds.) 1999. Schweizerisches Landesforstinventar. Ergebnisse der Zweitaufnahme 1993–1995. Birmensdorf, Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft. Bern, Bundesamt für Umwelt, Wald und Landschaft. Bern, Stuttgart, Wien. Haupt.

Braun-Blanquet, J. 1932. Plant Sociology: The Study of Plant Communities. (Translated by G.D. Fuller and H.S. Conard.) McGraw-Hill, New York and London.

Braun-Blanquet, J. 1964. Pflanzensoziologie. 3. Aufl. Wien, New York.

Bruelheide, H. 1995. Die Grünlandvegetation des Harzes und ihre Standortbedingungen: mit einem Beitrag zum Gliederungsprinzip auf der Basis von statistisch ermittelten Artengruppen. Dissertationes botanicae 244.

Bruelheide, H. 1997. Using formal logic to classify vegetation. *Folia Geobot. Phytotax.* **32**: 41–46.

Bruelheide, H. and Flintrop, T. 1994. Arranging phytosociological tables by species-relevé groups. *Journal of Vegetation Science* **5**: 311–316.

Brzeziecki, B., Kienast, F. and Wildi, O. 1993. A simulated map of the potential natural forest vegetation of Switzerland. *Journal of Vegetation Science* **4**: 499–508.

Clarke, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* **18**: 117–143.

Clements, F.E. 1916. Plant succession. An analysis of the development of vegetation. Carnegie Institute, Washington, Publication 242, Washington, DC.

Connell, H.J. and Slatyer, R.O. 1977. Mechanisms of succession in natural communities and their role in community stability and organisation. *American Naturalist* **111**: 1119–1144.

Dengler, J., Chytrý, M. and Ewald, J. 2008. Phytosociology. In: S.E. Jørgensen and B.D. Fath (eds.) General Ecology. Vol. 4 of Encyclopedia of Ecology, Elsevier, Oxford. 2767–2779.

Diamond, J. 1999. Guns, Germs and Steel: The Fates of Human Societies. W.W. Norton and Company, New York, London.

Digby, P.G.N. and Kempton, R.A. 1987. Multivariate Analysis of Ecological Communities. Chapman and Hall, London.

Dixon, P. 2003. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**: 927–930.

Ellenberg, H. 1956. Aufgaben und Methoden in der Vegetationskunde. In: H. Walter, Einführung in die Phytologie IV/1, Stuttgart.

Ellenberg, H. 1974. Zeigerwerte der Gefässpanzen Mitteleuropas. Scripta Geobotanica. Verlag Erich Goltze, Göttingen, DE.

Ellenberg, H. and Klötzli, F. 1972. Waldgesellschaften und Waldstandorte der Schweiz. *Mitt. Eidgenöss. Forsch.anst. Wald Schnee Landsch.* **48**(4): 587–930.

Elith, J., Graham, C.H., Anderson, P.R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.McC., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. and Zimmermann, N.E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**: 129–151.

Ewald, J. 2003. A critique for phytosociology. *Journal of Vegetation Science* **14**: 291–296.

Feldmeyer-Christe, E., Ecker, K., Küchler, M., Graf, U. and Waser, L. 2007. Improving predictive mapping in Swiss mire ecosystems through re-calibration of indicator values. *Applied Vegetation Science* **10**: 183–192.

Feoli, E. and Orlóci, L. 1979. Analysis of concentration and detection of underying factors in structured tables. *Vegetatio* **40**: 49–54.

Fisher, R.A. 1940. The precision of discriminant functions. *Annals of Eugenics* **10**: 422–429.

Floyd, R.W. 1962. Algorithm 97: shortest path. *Communications of the Association for Computing Machinery* **5**: 345.

Forrester, J.W. 1968. Principles of Systems. Wright-Allen Press, Cambridge, MA.

Gauch, H.G. 1982. Multivariate Analysis in Community Ecology. Cambridge Studies in Ecology. Cambridge University Press, Cambridge.

Ghosh, S. and Wildi, O. 2007. Statistical analysis of landscape data: space-for-time, probability surfaces and discovering species. In: F. Kienast, O. Wildi and S. Ghosh (eds.) A Changing World: Challenges for Landscape Research. Springer Landscape Series, Dordrecht. Vol. 8: 209–221.

Gilliam, F.S. and Saunders, N.E. 2003. Making more sense of the order: a review of Canoco for Windows 4.5, PC-ORD version 4 and SYN-TAX 2000. *Journal of Vegetation Science* **14**: 297–304.

Gleason, H.A. 1926. The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club* **53**: 7–26.

Gleason, H.A. 1939. The individualistic concept of the plant association. *Amer. Midl. Nat.* **21**: 92–110.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**: 325–338.

Gower, J.C. and Ross, G.J.S. 1969. Minimum spanning tree and single linkage cluster analysis. *Appl. Stat.* **18**: 54–64.

Graf, U., Wildi, O., Feldmeyer-Christe, E. and Küchler, M. 2010. A phytosociological classification of Swiss mire vegetation. *Botanica Helvetica*. In press.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. Wiley-Interscience, New York, Chichester, Brisbane, Toronto.

Grünig, A., Steiner, G.M., Ginzler, C., Graf, U. and Küchler, M. 2005. Approaches to Swiss mire monitoring. *Stapfia* **85**: 435–452.

Guisan, A. and Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* **135**: 147–186.

Hennekens, S.M. and Schaminée, J.H.J. 2001. TOURBOVEG, a comprehensive data base management system for vegetation data. *Journal of Vegetation Science* **12**: 589–591.

Hill, M.O. 1973. Reciprocal averaging: an Eigenvector method of ordination. *J. Ecol.* **61**: 237–249.

Hill, M.O. 1979a. DECORANA: A FORTRAN Program for Detrended Correspondence Analysis and Reciprocal Averaging. Cornell University, Ithaca, NY.

Hill, M.O. 1979b. TWINSPAN: A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-way Table by Classification of the Individuals and Attributes. Cornell University, Ithaca, NY.

Hubbell, S.P. 2001. The unified theory of biodiversity and biogeography: a synopsis of the theory and some challenges ahead. In: J. Silvertown and J. Antonovics (eds.) Integrating ecology and evolution in a spatial context. Blackwell. 393–411.

Huisman, J., Olff, H. and Fresco, L.F.M. 1993. A hierarchical set of models for species response analysis. *J. Veg. Sci.* **4**: 37–46.

International Statistical Institute 2009. Multilingual Glossary of Statistical Terms. http://isi.cbs.nl/glossary.htm.

Jancey, R.C. 1979. Species ordering on a variance criterion. *Vegetatio* **39**: 59–63.

Jennings, M., Loucks, O., Peet, R., Faber-Langendoen, D., Glenn-Lewin, D., Grossmann, D., Damman, A., Barbour, M., Pfister, R., Walker, M., Talbot, S., Walker, J., Hartshorn, G., Waggoner, G., Abrams, M., Hill, A., Roberts, D., Tart, D. and Rejmanek, M. 2003. Guidelines for describing associations and alliances of the US National Vegetation Classification Panel. The Ecological Society of America Vegetation Classification Panel.

Jongman, R.H.G., ter Braak, C.J.F. and van Tongeren, O.F.R. 1995. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge.

Keller, W., Wohlgemuth, T., Kuhn, N., Schütz, M. and Wildi, O. 1998. Waldgesellschaften der Schweiz auf floristischer Grundlage. Mitteilungen der Eidgenössischen Forschungsanstalt für Wald, Schnee und Landschaft (WSL) 73, Vol. 2.

Kent, M. and Coker, P. 1992. Vegetation Description and Analysis. Belhaven Press, London.

Kienast, F., Wildi, O. and Ghosh, S. (eds.) 2007. A Changing World: Challenges for Landscape Research. Springer Landscape Series, Dordrecht. Vol. 8.

Kŏci, M., Chytrý, M. and Tichý, L. 2003. Formalized reproduction of an expert-based phytosociological classification: a case study of subalpine tall-forb vegetation. *J. Veg. Sci.* **14**: 601–610.

Kowarik, I. 1987. Kritische Anmerkungen zum theoretischen Konzept der potentiellen natürlichen Vegetation mit Anregungen zu einer zeitgemässen Modifikation. *Tüxenia* **7**: 53–67.

Küechler, A.W. 1988. Mapping dynamic vegetation. In: A.W. Küechler and I.S. Zonneveld (eds.) Vegetation mapping. Handbook of Vegetation Science **10**: 13–23.

Küchler, M. 2009. VEGEDAZ. http://www.wsl.ch/dienstleistungen/vegedaz/index_DE.

Krüsi, B.O., Schütz, M., Bigler, C., Grämiger, H. and Achermann, G. 1998. Huftiere und Vegetation im Schweizerischen Nationalpark von 1917 bis 1997. Teil 1: Einfluss auf die botanische Vielfalt der subalpinen Weiden; Teil 2: Einfluss auf das Wald-Freilandverhältnis. In: R. Cornelius and R. Hofmann (eds.) Extensive Haltung robuster Haustierrassen, Wildtiermanagement, Multi-Spezies-Projekte – Neue Wege in Naturschutz und Landschaftspflege? Inst. Zoo-Wildtierforsch., Berlin. 62–74.

Landolt, E. 1977. Oekologische Zeigerwerte zur Schweizer Flora. *Veröff. Geobot. Inst. ETH, Stiftung Rübel* 64.

Legendre, P. and Anderson, M.J. 1999. Distance-based redundancy analysis: testing multi-species responses in multi-factorial experiments. *Ecological Monographs* **69**: 1–24.

Legendre, P. and Legendre, L. 1998. Numerical Ecology. 2nd ed. Elsevier, Amsterdam.

Legendre, P. and Fortin, M.-J. 1989. Spatial analysis and ecological modelling. *Vegetatio* **80**: 107–138.

Lepš, J. and Šmilauer, P. 2003. Multivariate Analysis of Ecological Data using CANOCO. Cambridge University Press, Cambridge.

Lindacher, R. 1996. Verifikation der potentiellen natürlichen Vegetation mittels Vegetationssimulation am Beispiel der TK 6434 'Hersbruck'. *Hoppea, Denkschr. Regensb. Bot. Ges.* **57**: 5–143.

Lippe, E., De Smitt, J.T. and Glenn-Lewin, D.C. 1985. Markov models and succession: a test from a heathland in the Netherlands. *Journal of Ecology* **73**: 775–791.

Lischke, H. 2005. Modeling tree species migration in the Alps during the Holocene: what creates complexity? *Ecological Complexity* **2**: 159–174.

Lischke, H., Lotter, A.F. and Fischlin, A. 2002. Untangling a Holocene pollen record with forest model simulations and independent climate data. *Ecological Modelling* **150**: 1–21.

Loehle, C. 1983. Evaluation of theories and calculation tools in ecology. *Ecological Modelling* **19**: 239–247.

Lotter, A.F. 1999. Late-glacial and Holocene vegetation history and dynamics as shown by pollen macrofossil analyses in annually laminated sediments from Soppensee, central Switzerland. *Vegetation History and Archaeobotany* **8**: 165–184.

Maarel, van der, E. 1979. Transformation of cover-abundance values in phytosociology and its effects on community similarity. *Vegetatio* **39**: 97–114.

Maarel, van der, E. 2005. Vegetation Ecology. Blackwell Publishing, Malden, Oxford, Victoria.

Maarel, van der, E., Janssen, J.G.M. and Louppen, J.M.W. 1978. TABORD, a program for structuring phytosociological tables. *Vegetatio* **38**: 143–156.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**: 209–220.

Mátyás, G. and Sperisen, C. 2001. Chloroplast DNA polymorphisms provide evidence for postglacial re-colonisation of oaks (*Quercus ssp*.) *across the Swiss Alps. Theor. Appl. Genet.* **102**: 12–20.

Maynard Smith, J. 1974. Models in Ecology. Cambridge University Press, London, New York, Melburne.

Meadows, D.H., Meadows, D.L. and Randers, J. 1972. The Limits to Growth. Universe Books.

Mucina, L. 1997. Classification of vegetation: past, present and future. *J. Veg. Sci.* **8**: 751–760.

Mueller-Dombois, D. and Ellenberg, H. 1974. Aims and Methods of Vegetation Ecology. John Wiley & Sons, New York, Chichester, Brisbane, Toronto.

Neuhäusl, R. 1984. Umweltgemässe natürliche Vegetation, ihre Kartierung und Nutzung für den Umweltschutz. *Preslia* **56**: 205–212.

Nishisato, S. 1980. Analysis of Categorial Data: Dual Scaling and its Applications. Mathematical Expositions No. 24. University of Toronto Press, Toronto.

Orlóci, L. 1967. An agglomerative method for classification of plant communities. *J. Ecol.* **55**: 193–206.

Orlóci, L. 1978. Multivariate Analysis in Vegetation Research. 2nd ed. Junk, The Hague.

Orlóci, L. 1991a. CONAPACK: A Program for Canonical Analysis of Classification Tables. Ecological Computations Series: Vol. 4. SPB Academic Publishing, The Hague.

Orlóci, L. 1991b. On character-based plant community analysis: choice, arrangement, comparison. *Coenoses* **6**: 103–107.

Orlóci, L. 1993. The complexities and scenarios of ecosystem analysis. In: G.P. Patil and C.R. Rao (eds.) Multivariate Environmental Statistics. Elsevier Scientific, New York. 423–432.

Orlóci, L. 2000. From order to causes: a personal view, concerning the principles of syndynamics. http://sites.netscape.net/lorloci.

Orlóci, L. 2001. Prospects and expectations: reflections on a science in change. *Community Ecology* **2**: 187–196.

Orlóci, L. and Kenkel, N. 1985. Introduction to Data Analysis. International Co-operative Publ. House, Burtonsville, MD.

Orlóci, L. and Orlóci, M. 1985. Comparison of communities without the use of species: model and example. *Ann. Bot. (Roma)* **43**: 275–285.

Orlóci, L. and Pillar, V. de Patta 1989. On sample size optimality in ecosystems survey. *Biometrie-Praximetrie* **29**: 173–184.

Orlóci, L., Anand, M. and He, X. 1993. Markov chain: a realistic model for temporal coenosere? *Biom. Praxim* **33**: 7–26.

Orlóci, L., Pillar, V. de Patta, Anand, M., Behling, H. 2002. Some interesting characteristics of the vegetation process. *Community Ecology* **3**(2): 125–146.

Pantke, R. 2003. Pflanzengesellschaften der Schweiz. http://pages.unibas.ch/vegetation-ch/.

Parker, V.T. and Pickett, S.T.A. 1998. Historical contingency and multiple scales of dynamics within plant communities. In: D.L. Peterson and V.T. Parker (eds.) Ecological Scale: Theory and Applications. Columbia University Press. 171–191.

Peres-Neto, P.R., Jackson, D.A. and Somers, K.M. 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis* **49**: 974–997.

Pickett, T.A. 1989. Space-for-time substitution as an alternative to long-term studies. In: E. Likens (ed.), Long-term Studies in Ecology: Approaches and Alternatives. Springer, New York. 110–135.

Pielou, E.C. 1984. The Interpretation of Ecological Data. John Wiley & Sons, New York, London, Sidney, Toronto.

Pillar, V. de Patta, Duarte, L.S., Sosinski, E.E. and Joner, F. 2009. Discriminating trait-convergence and trait-divergence assembly patterns in ecological community gradients. *Journal of Vegetation Science* **20**: 334–348.

Podani, J. and Feoli, E. 1991. A general strategy for the simultaneous classification of variables and objects in ecological data tables. *Journal of Vegetation Science* **2**: 435–444.

Poore, M.E.D. 1955. The use of phytosociological methods in ecological investigations. I–III. *J. Ecol.* **43**: 226–244, 245–269, 606–651.

Poore, M.E.D. 1962. The method of successive approximation in descriptive ecology. *Advances in Ecological Research* **1**: 35–68.

Rao, C.R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhyaá, Ser. A* **26**: 329–358.

Raunkiaer, C. 1937. The Life Forms of Plants. Oxford University Press, Oxford. (Translated from the original, published in Danish, 1907)

Rényi, A. 1961. On measures of entropy and information. In: J. Neyman (ed.) Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Univeristy of California Press, Berkeley. 547–561.

Roberts, D.W. 1986. Ordination on the basis of fuzzy set theory. *Vegetatio* **66**: 123–131.

Rogers, P.C., Moore, K.D. and Ryel, R.J. 2009. Aspen succession and nitrogen loading: a case for epiphytic lichens as bioindicators in the Rocky Mountains, USA. *Journal of Vegetation Science* **20**: 498–510.

Roweis, T.S. and Saul, L.K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**: 2323–2326.

Sampford, M.R. 1962. An Introcuction to Sampling Theory with Applications to Agriculture. Oliver and Boyd, Endinburgh.

Shepard, R.N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika* **27**: 125–139.

Sneath, P.H.A. and Sokal, R.R. 1973. Numerical Taxonomy: The Principles and Practice of Numerical Classification. W.H. Freeman, San Francisco.

Stromberg, J.C., Rychener, T.J. and Dixon, M.D. 2009. Return of fire to a free-flowing desert river: effects of vegetation. *Restoration Ecology* **17**: 327–338.

Tenenbaum J.B., de Silva V. and Langford J.C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**: 2323–2326.

ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**: 1167–1179.

Tichý, L. 2002. JUICE, software for vegetation classification. *Journal of Vegetation Science* **13**: 451–453.

Tüxen, R. 1956. Die heutige potentielle Vegetation von Oberfranken. *Angew. Pflanzensoz. (Stolzenau)* **13**: 5–42.

Usher, M.B. 1981. Modelling ecological succession, with particular reference to Markovian model. *Vegetatio* **46**: 11–18.

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T.J.C., Fromentin, J.-M., Hoegh-Guldberg, O. and Bairlein, F. 2002. Ecological responses to recent climate change. *Nature* **416**, 389–395.

Wagner, H.H. 2004. Direct multi-scale ordination with canonical correspondence analysis. *Ecology* **85**: 342–351.

Ward, J.H. 1963. Hierarchical grouping to optimise an objective function. *J. Amer. Statist. Assoc.* **58**(301): 236–244.

Wildi, O. 1976. Untersuchung von Vegetationsgrenzen mit Hilfe dynamischer Modelle. *Ber. Deutsch. Bot. Ges. Bd.* **89**: 365–370.

Wildi, O. 1977. Beschreibung exzentrischer Hochmoore mit Hilfe quantitativer Methoden. *Veröff. Geobot. Inst. ETH, Stiftung Rübel* **60**: 128S.

Wildi, O. 1984. Species selection by interactive ranking. *Vegetatio* **56**: 161–166.

Wildi, O. 1989. A new numerical solution to traditional phytosociological tabular classification. *Vegetatio* **81**: 95–106.

Wildi, O. 1990. A multiple scale sampling design for long term monitoring. Proceedings of the International Conference and Workshop, Vol. 2. Bethesda, MD. 975–982.

Wildi, O. 2001. Statistical design and analysis in long term vegetation monitoring. In: C.A. Burga and A. Kratochwil (eds.) Biomonitoring: General and Applied Aspects on Regional and Global Scales. Kluver, Dordecht. Tasks for Vegetation Science. Vol. 35: 17–39.

Wildi, O. 2002. Modelling succession from pasture to forest in time and space. *Community Ecology* **3**(2): 181–189.

Wildi, O. and Orlóci, L. 1991. Flexible gradient analysis: a note on ideas and an application. In: E. Feoli and L. Orlóci (eds.) Computer Assisted Vegetation Analysis. Kluver, Dortrecht. 265–271.

Wildi, O. and Orlóci, L. 1996. Numerical Exploration of Community Patterns. 2nd ed. SPB Academic Publishing, The Hague.

Wildi, O. and Schütz, M. 2000. Reconstruction of a long-term recovery process from pasture to forest. *Community Ecology* **1**: 25–32.

Wildi, O., Feldmeyer-Christe, E., Ghosh, S. and Zimmermann, N.E. 2004. Comments on vegetation monitoring approaches. *Community Ecology* **5**: 1–5.

Wildi, O. and Orlóci, L. 2007. Essay on the Study of the Vegetation Process. In: F. Kienast, O. Wildi and S. Ghosh (eds.) A Changing World: Challenges for Landscape Research. Springer Landscape Series, Dordrecht. Vol. 8: 195–207.

Wildi, O. and Schütz, M. 2007. Scale sensitivity of synthetic long-term vegetation time series derived through overlay of short-term field records. *J. Veg. Sci.* **18**: 471–478.

Wissel, C. 1989. Theoretische Ökologie. Springer-Verlag, Berlin.

Wohlgemuth, T., Moser B., Brändli, U.-B., Kull, P. and Schütz, M. 2008. Diversity of forest plant species at the community and landscape scales in Switzerland. *Plant Biosystems* **142**: 604–613.

Zimmermann, N.E. and Kienast, F. 1999. Predictive mapping of alpine grassland in Switzerland: species versus community approach. *Journal of Vegetation Science* **10**: 469–482.

Index

Data Analysis in Vegetation Ecology Otto Wildi

Index compiled by Alison Waggitt