Data Analysis in Vegetation Ecology

Otto Wildi

WILEY-BLACKWELL

Contents

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

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.

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 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 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.

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

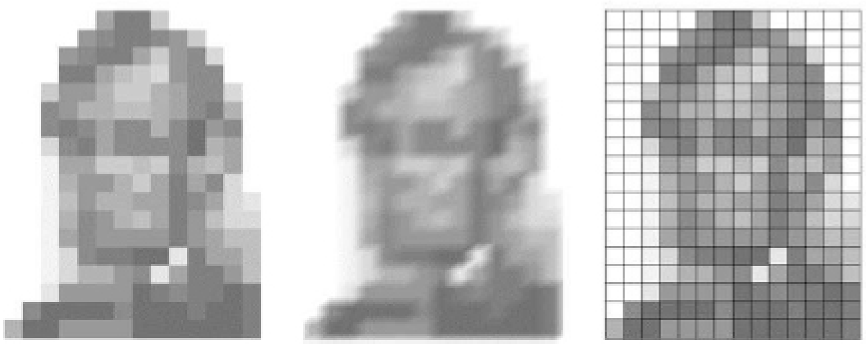
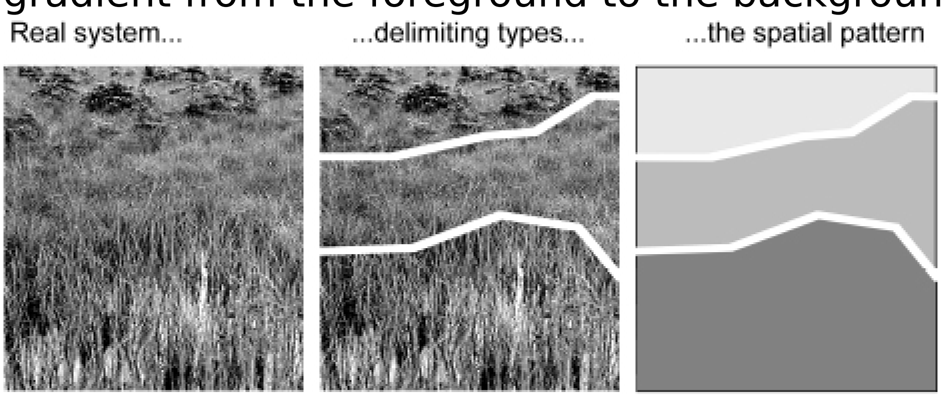


Figure 2.2 Vegetation mapping as a method for establishing a pattern (bog vegetation with a wetness gradient from the foreground to the background).



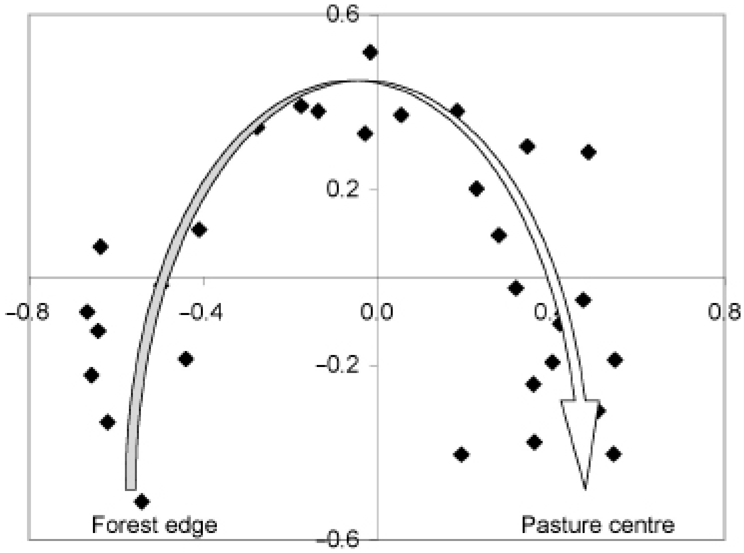
Patterns are often obscured not just by overlay, but by random variation (sometimes referred to as statistical noise) hiding the regularities. Methods are needed to divide the total variation into two components, one containing the regularity and one representing randomness.

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 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.

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.



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.