Use R!

Daniel Borcard
François Gillet
Pierre Legendre

Numerical Ecology with R

Second Edition

Springer

Use R!

Series Editors

Robert Gentleman Kurt Hornik Giovanni Parmigiani

More information about this series at http://www.springer.com/series/6991

Daniel Borcard • François Gillet • Pierre Legendre

Numerical Ecology with R

Second Edition

Springer

Daniel Borcard
Université de Montréal
Département de sciences biologiques
Montréal, Québec, Canada H3C 3J7

François Gillet
Université Bourgogne Franche-Comté
UMR Chrono-environnement
Besançon, France

Pierre Legendre
Université de Montréal
Département de sciences biologiques
Montréal, Québec, Canada H3C 3J7

ISSN 2197-5736 ISSN 2197-5744 (electronic)
Use R!
ISBN 978-3-319-71403-5 ISBN 978-3-319-71404-2 (eBook)
https://doi.org/10.1007/978-3-319-71404-2

Library of Congress Control Number: 2017961342

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Ecology is sexy. Teaching ecology is therefore the art of presenting a fascinating topic to well-predisposed audiences. It is not easy: the complexities of modern ecological science go well beyond the introductory chapters taught in high schools or the marvellous movies about ecosystems presented on TV. But well-predisposed audiences are ready to make the effort. *Numerical* ecology is another story. For some unclear reasons, a majority of ecology-oriented people are strangely reluctant when it comes to quantifying nature and using mathematical tools to help understand it. As if nature was inherently non-mathematical, which it is certainly not: mathematics is the common language of all sciences. Teachers of biostatistics and numerical ecology thus have to overcome this reluctance: before even beginning to teach the subject itself, they must convince their audience of the interest and necessity of it.

During many decades ecologists, be they students or researchers (in the academic, private or government spheres), used to plan their research and collect data with few, if any, statistical consideration, and then entrusted the "statistical" analyses of their results to a person hired especially for that purpose. That person may well have been a competent statistician, and indeed in many cases the progressive integration of statistics into the whole process of ecological research was triggered by such people. In other cases, however, the end product was a large amount of data summarized using a handful of basic statistics and tests of significance that were far from revealing all the richness of the structures hidden in the data tables. The separation of the ecological and statistical worlds presented many problems. The most important were that the ecologists were unaware of the array of methods available at the time, and the statisticians were unaware of the ecological hypotheses to be tested and the specific requirements of ecological data (the double-zero problem is a good example). Apart from preventing the data to be exploited properly, this double unawareness prevented the development of methods specifically tailored to ecological problems.

The answer to this situation is to form mathematically inclined ecologists. Fortunately, more and more such people have appeared during the recent decades. The result of their work is a huge development of statistical ecology, the availability

of several excellent textbooks, and the increasing awareness of the responsibility of ecologists with regard to the proper design and analysis of their research. This awareness makes the task easier for teachers as well.

Until the first years of this millennium, however, a critical ingredient was still missing for the teaching to be efficient and for the practice of statistics to become generalized among ecologists: a set of standard packages available to everyone, everywhere. A biostatistics or numerical ecology course means nothing without practical exercises. A course linked to commercial software is much better, but it is bound to restrict future applications if the researcher moves and loses access to the software that he or she knows. Furthermore, commercial packages are in most cases written for larger audiences than the community of ecologists and they may not include all the functions required for analysing ecological data. The **R** language resolved that issue, thanks to the dedication of the many researchers who created and freely contributed extensive, well-designed, and well-documented packages. Now the teacher no longer has to say: "this is the way PCA works... on paper;" she or he can say instead: "this is the way PCA works, now I will show you on-screen how to run one, and in a few minutes you will be able to run your own, and do it anywhere in the world on your own data!"

Another fundamental property of the **R** language is that it is meant as a self-learning environment. A book on **R** is therefore bound to follow that philosophy, and must provide the support necessary for anyone wishing to explore the subject by himself or herself. This book has been written to provide a bridge between the theory and practice of numerical ecology, that anyone can cross. Our dearest hope is that it will make many happy teachers and happy ecologists.

Since they are living entities, both the field of numerical ecology and the **R** language evolve. As a result, much has happened in both fields since the publication of the first edition of *Numerical Ecology with R* in 2011. Therefore, it was time not only to update the code provided in the first edition, but also to present new methods, provide more insight into existing ones, offer more examples and a wider array of applications of the major methods. We also took the opportunity to present the code in a more attractive way, generated by R Markdown in RStudio®, with different colours for functions, objects, arguments and comments.

Our dearest hope is that all this will make many more happy teachers and happy ecologists.

Montréal, QC, Canada — Daniel Borcard
Besançon, France — François Gillet
Montréal, QC, Canada — Pierre Legendre

Contents

About the Authors

Daniel Borcard is lecturer of Biostatistics and Ecology and researcher in Numerical Ecology at Université de Montréal, Québec, Canada. His research interests include Numerical Ecology, Ecology of communities, and Soil Ecology/Zoology.

François Gillet is professor of Community Ecology and Ecological Modelling at Université Bourgogne Franche-Comté, Besançon, France, and visiting professor at École Polytechnique Fédérale de Lausanne, Switzerland. His research deals with the structure, diversity, ecology and dynamics of plant communities.

Pierre Legendre is professor of Quantitative Biology and Ecology at Université de Montréal, fellow of the Royal Society of Canada, and Web of Science Highly Cited Researcher in Environment/Ecology. He is the founder of the field of numerical ecology.

Supplementary Material

All the necessary data files, the scripts used in the chapters, as well as the **R** functions and packages that are not available through the CRAN web site, can be downloaded from our web page http://adn.biol.umontreal.ca/~numericalecology/numecolR/.

Chapter 1
Introduction

1.1 Why Numerical Ecology?

Although multivariate analysis of ecological data already existed and was being actively developed in the 1960's, it really flourished in the years 1970 and later. Many textbooks were published during these years, among them the seminal *Écologie numérique* (Legendre and Legendre 1979), and its English translation *Numerical Ecology* (Legendre and Legendre 1983). The authors of these books unified under one single roof a very wide array of statistical and other numerical techniques and presented them in a comprehensive way, not only to help researchers understand the available methods of statistical analysis, but also to explain how to choose and apply them in an ordered, logical way to reach their research goals. Mathematical explanations were not absent from these books, and provided a precious insider look into the various techniques, which was appealing to readers wishing to go beyond the simple user level.

Since then, numerical ecology has become ubiquitous. Every serious researcher or practitioner has become aware of the tremendous interest of exploiting painfully acquired data as efficiently as possible. Other manuals have been published (e.g. Orlóci and Kenkel 1985; Jongman et al. 1995; McCune and Grace 2002; McGarigal et al. 2000; Zuur et al. 2007; Greenacre and Primicerio 2013; Wildi 2013). A second English edition of *Numerical Ecology* was published in 1998, followed by a third in 2012, broadening the perspective and introducing numerous methods that were unavailable at the times of the previous editions. The progress continues. In this book we present some of the developments that we consider most important, albeit in a more user-oriented way than in the abovementioned manuals, using the **R** language. For the most recent methods, we provide explanations at a more fundamental level when we consider it appropriate and helpful.

Not all existing methods of data analysis are addressed in this book, of course. Apart from the most widely used and fruitful methods, our choices are based on our own experience as quantitative community ecologists. However, small sections

D. Borcard et al., *Numerical Ecology with R*, Use R!,
https://doi.org/10.1007/978-3-319-71404-2_1

have sometimes been added to briefly describe other avenues than the main ones, without going into details.

1.2 Why R?

The **R** language has experienced such a tremendous development and reached such a wide array of users during the recent years that a justification of its application to numerical ecology is not required. Development also means that more and more domains of numerical ecology are now covered, up to the point where, computationally speaking, some of the most recent methods are actually only available through **R** packages.

This book is not intended as a primer in **R**, however. To find that kind of support, readers should consult the CRAN web page (http://www.R-project.org). The link to *Manuals* provides many free electronic documents, and the link to *Books* many references. Readers are expected to have a minimal working knowledge of the basics of the language, e.g. formatting data and importing them into **R**, awareness of the main classes of objects handled in this environment (vectors, matrices, data frames and factors), as well as the basic syntax necessary to manipulate, create and otherwise use objects within **R**. Nevertheless, Chap. 2 starts at an elementary level as far as multivariate objects are concerned, since these are the main targets of most analyses addressed throughout the book, while not necessarily being most familiar to many users.

The book is by far not exhaustive as to the array of functions devoted to any of the methods. Usually we present one or several variants, but often other functions serving similar purposes are available in **R**. Centring the book on a small number of well-integrated packages and adding some functions of our own when necessary helps users up the learning curve while keeping the amount of package-level idiosyncrasies at a reasonable level. Our choices should not suggest that other existing packages are inferior to the ones used in the book.

1.3 Readership and Structure of the Book

The intended audience of this book is the researchers, practitioners, graduate students and teachers who already have a background in general and multivariate statistics and wish to apply their knowledge to their data using the **R** language, as well as people willing to accompany their learning of the discipline with practical applications. Although an important part of this book follows the organization and symbolism of Legendre and Legendre (2012) and many references to that book are made herein, readers may draw their training from other sources without problem.

Combining an application-oriented book such as this one with a detailed exposé of the methods used in numerical ecology would have led to an impossibly long and

cumbersome opus. However, all chapters start with a short introduction summarizing its subject matter, to ensure that readers are aware of the scope of the chapter and can appreciate the point of view from which the methods are addressed. Depending on the amount of documentation already existing in statistical textbooks, some introductions are longer than others.

Overall, the book guides readers through an applied exploration of the major methods of multivariate data analysis, as seen through the eye of an ecologist. Starting with some exploratory approaches (Chap. 2), it proceeds logically with the construction of the key building blocks of most techniques, i.e. association measures and matrices (Chap. 3), and then submits example data to three families of approaches: clustering (Chap. 4), ordination and canonical ordination (Chaps. 5 and 6), spatial analysis (Chap. 7), and finally community diversity (Chap. 8). The methods' aims thus range from descriptive to explanatory and to predictive and encompass a wide variety of approaches that should provide readers with an extensive toolbox that can address a wide palette of questions arising in contemporary multivariate ecological analysis.

1.4 How to Use This Book

The book is meant as a companion when working at the computer. The authors pictured a reader studying a chapter by reading the text and simultaneously executing the code. To fully understand the various methods, it is preferable to go through the chapters sequentially, since each builds upon the previous ones. At the beginning of each chapter, an empty **R** console is assumed to be open. All the necessary data files, the scripts used in the chapters, as well as the **R** functions and packages that are not available through the CRAN web site, can be downloaded from our web page (http://adn.biol.umontreal.ca/~numericalecology/numecolR/). Some of the homemade functions duplicate existing ones, providing alternative solutions (for instance different or expanded graphical outputs), while others have been written to streamline complex sequences of operations.

Although the code provided can be run in one single copy-and-paste shot within each chapter (with some rare exceptions for interactive functions), the best procedure is to proceed through the code slowly and explore each set of commands carefully. Although the use and meaning of some arguments is explained within the code or in the text, readers are warmly invited to use and abuse of the **R** documentation files (function name following a question mark) to learn about and explore the various options available. Our aim is not to describe all options of all functions, which would be an impossible and useless task. We are confident that an avid user, willing to go beyond the provided examples, will be kept busy for months exploring the options that he or she deems the most interesting.

Within each chapter, after the introduction, readers are invited to import the data as well as the **R** packages necessary for the exercises of the whole chapter. The **R** code used in each chapter is self-contained, i.e., it can usually be run in one step

even if some analyses are based on results produced in previous chapters. If such objects are needed, they are recomputed at the beginning of the chapter.

In everyday use, one generally does not produce an **R** object for every single operation, nor does one create and name a new graphical window for every plot. We do that in the book to provide readers with all the entities necessary to backtrack the procedures, compare results and explore variants. Therefore, after having run most of the code in a chapter, if one decides to explore another path using some intermediate result, the corresponding object will be available without need to re-compute it. This is particularly handy for results of computer-intensive methods (like some based on large numbers of random permutations).

In the code sections of the book, all calls to graphical windows have been deleted for brevity. They are found in the electronic code scripts, however. Furthermore, the book shows several, but not all, graphical outputs for reference.

Sometimes, readers are made aware of some special features of the code or of tricks used to obtain particular results, by means of hint boxes located at the bottom of code sections.

Although many methods are applied to the example data, ecological interpretation is not provided in all cases. Sometimes questions are left open to readers, as an incentive to verify if she or he has correctly understood the method, and hence its application and the numerical or graphical outputs.

Lastly, for some methods, programming-oriented readers are invited to write their own code. These incentives are placed in boxes called "code-it-yourself corners". When examples are provided, they are meant for pedagogical purposes and do not pretend at computational efficiency. The aim of these boxes is to help interested readers code in **R** the matrix algebra equations presented in Legendre and Legendre (2012) and obtain the main outputs that ready-made packages provide. The whole idea is of course to reach the deepest possible understanding of the mathematical working of some key methods.

1.5 The Data Sets

Apart from rare cases where *ad hoc* fictitious data are built for special purposes, the applications rely on two main data sets that are readily available in **R**. However, data provided in R packages can be modified over the years. Therefore we prefer to provide them also in the electronic material accompanying this book, because this ensures that the results obtained by the readers will be exactly the same as those presented in the book. The two data sets are briefly presented here. The first (Doubs) data set is explored in more detail in Chap. 2, and readers are encouraged to apply the same exploratory methods to the second one.

1.5.1 The Doubs Fish Data

In an important doctoral thesis, Verneaux (1973; see also Verneaux et al. 2003) proposed to use fish species to characterize ecological zones along European rivers and streams. He showed that fish communities were good biological indicators of these water bodies. Starting from the river source, Verneaux proposed a typology in four zones, and he named each one after a characteristic species: the trout zone (from the brown trout *Salmo trutta fario*), the grayling zone (from *Thymallus thymallus*), the barbel zone (from *Barbus barbus*) and the bream zone (from the common bream *Abramis brama*). The two upper zones are considered as the "Salmonid region" and the two lowermost ones form the "Cyprinid region". The corresponding ecological conditions, with much variation among rivers, range from relatively pristine, well oxygenated and oligotrophic to eutrophic and oxygen-deprived waters.

The Doubs data set that is used in the present book (**`Doubs.RData`**) consists of five data frames, three of them containing a portion of the data used by Verneaux for his studies. These data have been collected at 30 sites along the Doubs River, which runs near the France-Switzerland border in the Jura Mountains. The first matrix contains coded abundances of 27 fish species, the second matrix contains 11 environmental variables related to the hydrology, geomorphology and chemistry of the river, and the third matrix contains the geographical coordinates (Cartesian, X and Y in km) of the sites. The Cartesian coordinates have been obtained as follows. One of us (FG) returned to Verneaux's thesis to obtain more accurate positions of the sampling sites than available in existing databases. These new locations were coded in GPS angular coordinates (WGS84), and transformed into Cartesian coordinates by using function **`geoXY()`** of package **`SoDA`** in **R**. Earlier versions of these data have already served as test cases in the development of numerical techniques (Chessel et al. 1987). Two additional data frames are provided in the present book's material: **`latlong`** contains the latitudes and longitudes of the sampling sites, and **`fishtraits`** contains four quantitative variables and six binary variables describing the diet. Values are taken from various sources, mainly fishbase.org (Froese and Pauly 2017), checked and adapted to the regional context by François Degiorgi.[1]

Working with the original environmental data available in Verneaux's thesis, one of us (FG) made some corrections to the data available in **R** and restored the variables to their original units, which are presented in Table 1.1.

Since the fish species of this data set have well-defined ecological requirements that have been often exploited in ecological and applied environmental studies, it is useful to provide their full Latin and English names. This is done here in Table 1.2.

[1] Many thanks to Dr. Degiorgi for this precious work.

Table 1.1 Environmental variables of the Doubs data set used in this book and their units

Variable	Code	Units
Distance from the source	`dfs`	km
Elevation	`ele`	m a.s.l.
Slope	`slo`	‰
Mean minimum discharge	`dis`	$m^3 \cdot s^{-1}$
pH of water	`pH`	–
Hardness (Ca concentration)	`har`	$mg \cdot L^{-1}$
Phosphate concentration	`pho`	$mg \cdot L^{-1}$
Nitrate concentration	`nit`	$mg \cdot L^{-1}$
Ammonium concentration	`amm`	$mg \cdot L^{-1}$
Dissolved oxygen	`oxy`	$mg \cdot L^{-1}$
Biological oxygen demand	`bod`	$mg \cdot L^{-1}$

Table 1.2 Labels, Latin names, family and English names of the fish species of the Doubs dataset

Label	Latin name	Family	English name
Cogo	*Cottus gobio*	*Cottidae*	Bullhead
Satr	*Salmo trutta fario*	*Salmonidae*	Brown trout
Phph	*Phoxinus phoxinus*	*Cyprinidae*	Eurasian minnow
Babl	*Barbatula barbatula*	*Nemacheilidae*	Stone loach
Thth	*Thymallus thymallus*	*Salmonidae*	Grayling
Teso	*Telestes souffia*	*Cyprinidae*	Vairone
Chna	*Chondrostoma nasus*	*Cyprinidae*	Common nase
Pato	*Parachondrostoma toxostoma*	*Cyprinidae*	South-west European nase
Lele	*Leuciscus leuciscus*	*Cyprinidae*	Common dace
Sqce	*Squalius cephalus*	*Cyprinidae*	European chub
Baba	*Barbus barbus*	*Cyprinidae*	Barbel
Albi	*Alburnoides bipunctatus*	*Cyprinidae*	Schneider
Gogo	*Gobio gobio*	*Cyprinidae*	Gudgeon
Eslu	*Esox lucius*	*Esocidae*	Northern pike
Pefl	*Perca fluviatilis*	*Percidae*	European perch
Rham	*Rhodeus amarus*	*Cyprinidae*	European bitterling
Legi	*Lepomis gibbosus*	*Centrarchidae*	Pumpkinseed
Scer	*Scardinius erythrophtalmus*	*Cyprinidae*	Rudd
Cyca	*Cyprinus carpio*	*Cyprinidae*	Common carp
Titi	*Tinca tinca*	*Cyprinidae*	Tench
Abbr	*Abramis brama*	*Cyprinidae*	Freshwater bream
Icme	*Ameiurus melas*	*Ictaluridae*	Black bullhead
Gyce	*Gymnocephalus cernua*	*Percidae*	Ruffe
Ruru	*Rutilus rutilus*	*Cyprinidae*	Roach
Blbj	*Blicca bjoerkna*	*Cyprinidae*	White bream
Alal	*Alburnus alburnus*	*Cyprinidae*	Bleak
Anan	*Anguilla anguilla*	*Anguillidae*	European eel

Latin names after fishbase.org (Froese and Pauly 2017)

Table 1.3 Environmental variables of the oribatid mite data set used in this book and their units

Variable	Code	Units
Substrate density (dry matter)	`SubsDens`	$g \cdot dm^{-3}$
Water content	`WatrCont`	$g \cdot dm^{-3}$
Substrate	`Substrate`	7 unordered classes
Shrubs	`Shrub`	3 ordered classes
Microtopography	`Topo`	Blanket – Hummock

1.5.2 The Oribatid Mite Data

Oribatid mites (Acari: Oribatida) are a very diversified group of small (0.2 to 1.2 mm) soil-dwelling, mostly microphytophagous and detritivorous arthropods. A well-aerated soil or a complex substrate like *Sphagnum* mosses present in bogs and wet forests can harbour up to several hundred thousand (10^5) individuals per square metre. Local assemblages are sometimes composed of over a hundred species, including many rare ones. This diversity makes oribatid mites an interesting target group to study community-environment relationships at very local scales.

The example data set is composed of 70 cores of mostly *Sphagnum* mosses collected in a peat moss mat bordering a small lake (Lac Geai) on the territory of the *Station de biologie des Laurentides* of Université de Montréal, Québec, Canada in June 1989. The data were collected in order to test various ecological hypotheses about the relationships between living communities and their environment when the latter is spatially structured, and develop statistical techniques for the analysis of the spatial structure of living communities. It has since become a classical test data set, used in several publications (e.g. Borcard et al. 1992; Borcard and Legendre 1994; Borcard et al. 2004; Wagner 2004; Legendre 2005; Dray et al. 2006; Griffith and Peres-Neto 2006). These data are available in packages **`vegan`** and **`ade4`**.

The data set ("mite.RData") comprises three files that contain the abundances of 35 morphospecies, 5 substrate and microtopographic variables, and the X-Y Cartesian coordinates of the 70 cores (in cm). The environmental variables are the following (Table 1.3):

The cores have been sampled on a 10.0 m × 2.6 m strip of various substrates forming a transect between a mixed forest and the lake's free water on the shore of an acidic lake. Figure 1.1 shows the 70 soil cores and the types of substrate.

1.6 A Quick Reminder About Help Sources

The **R** language was designed to be a self-learning tool. So you can use and abuse of the various ways to ask questions, display code, run examples that are imbedded in the framework. Some important help tools are presented here (Table 1.4).

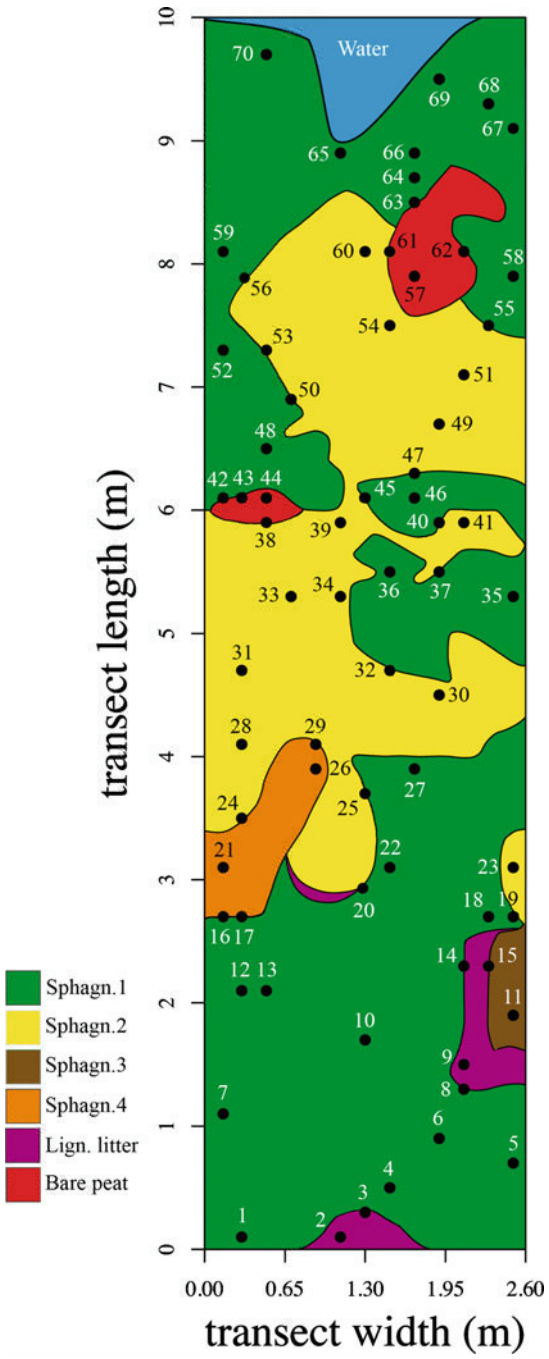

Fig. 1.1 Map of the mite data sampling area, showing the location of the 70 cores and the type of substrate (Details: see Borcard and Legendre 1994)

Table 1.4 Several help resources in **R**

Action	Use	Example	Remarks
?(question mark)	Obtain information about a function	`?decostand`	The package to which the function belongs must be active
`??` (double question mark)	Obtain information on the basis of a keyword	`??diversity`	The search is done in all packages installed in the computer
Type the name of a function	Display the code of the function on-screen	`diversity`	Not all functions can be displayed fully; some contain compiled code
`help (package="....")`	Display information on the package, including a list of all functions and data.	`help (package="ade4")`	
`data (package="...")`	List the datasets contained in a package	`data (package="vegan")`	
http://cran.r-project.org/	Broader search than above; access to discussion lists	Search on the CRAN web site: click on the "Search" link and choose one of the links	Outside the **R** master webserver.
http://www.rseek.org/	Seek any **R** function	Search "line plot"	Outside the **R** console

1.7 Now It Is Time...

... to get your hands full of code, numerical outputs and plots. Revise the basics of the methods, explore the code, analyse it, change it, try to apply it to your data and interpret your results. Above all, we hope to show that doing numerical ecology in **R** is fun!

Chapter 2
Exploratory Data Analysis

2.1 Objectives

Nowadays, most ecological research is done with hypothesis testing and modelling in mind. However, Exploratory Data Analysis (**EDA**), with its visualization tools and simple statistics, is still required at the beginning of the statistical analysis of multidimensional data, in order to:

- get an overview of the data;
- transform or recode some variables;
- orient further analyses.

As a worked example, we will explore the classical Doubs River dataset to introduce some techniques of EDA using **R** functions found in standard packages. In this chapter you will:

- learn or revise some bases of the **R** language;
- learn some EDA techniques applied to multidimensional ecological data;
- explore the Doubs dataset in hydrobiology as a first worked example.

2.2 Data Exploration

2.2.1 Data Extraction

The Doubs data used here are available in a .RData file found among the files provided with the book; see Chap. 1.

D. Borcard et al., *Numerical Ecology with R*, Use R!,
https://doi.org/10.1007/978-3-319-71404-2_2

```
# Load required packages
library(vegan)
library(RgoogleMaps)
library(googleVis)
library(labdsv)

# Source additional functions that will be used later in this
# Chapter. Our scripts assume that files to be read are in
# the working directory.
source("panelutils.R")

# Load the data. File Doubs.Rdata is assumed to be
# in the working directory
load("Doubs.RData")

# The file Doubs.RData contains the following objects:
#    spe: species (community) data frame (fish abundances)
#    env: environmental data frame
#    spa: spatial data frame - cartesian coordinates
#    fishtraits: functional traits of fish species
#    latlong: spatial data frame - latitude and longitude
```

Hints *At the beginning of a session, make sure to place all necessary data files and scripts in a single folder and define this folder as your working directory, either through the menu or by using function* **`setwd()`**.

Although it is not necessary, we strongly recommend that you use RStudio as script manager, which adds many interesting features to standard text editors. The R code in the companion materials of this book is optimized for RStudio and complies with the R Core Team's guidelines for good practices in R programming. Once all necessary files are placed in the same folder and RStudio is configured to run R scripts, just double-click on an R script file and the corresponding folder will be automatically defined as the current working directory.

Users of the standard ***R*** *console can use the* ***R*** *built-in text editor to write* ***R*** *code and run any selected portion using easy keyboard commands (<Control+Return> or <Command+Return> depending on the machine you are using). To open a new file, click on the File menu, then click on New script. Dragging an* ***R*** *script, for example our file "chap2.R", onto the R icon, will automatically open it in a new file managed by the* ***R*** *text editor.*

If you are uncertain of the class of an object, type **`class`**`(object_name)`.

2.2.2 Species Data: First Contact

We can start data exploration, which will first focus on the community data (object **`spe`** loaded as an element of the Doubs.RData file above). Verneaux used a semi-quantitative, species-specific, abundance scale (0–5), so that comparisons between

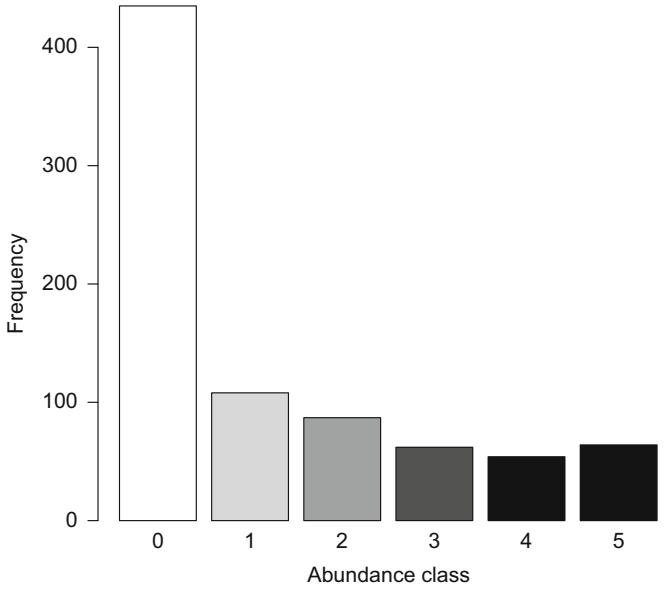

Fig. 2.1 Barplot of abundance classes

species abundances will make sense. The maximum value, 5, corresponds to the class with the maximum number of individuals captured by electrical fishing in the Doubs River and its tributaries (i.e. not only in this data set) by Verneaux. Therefore, species-specific codes cannot be understood as unbiased estimates of the true abundances (number or density of individuals) or biomasses at the sites.

We will first apply some basic **R** functions and draw a barplot (Fig. 2.1):

```
## Exploration of a data frame using basic R functions
spe                          # Display the whole data frame in the
                             # console
                             # Not recommended for large datasets!
spe[1:5, 1:10]               # Display only 5 lines and 10 columns
head(spe)                    # Display only the first 6 lines
tail(spe)                    # Display only the last 6 rows
nrow(spe)                    # Number of rows (sites)
ncol(spe)                    # Number of columns (species)
dim(spe)                     # Dimensions of the data frame (rows,
                             # columns)
```

```r
colnames(spe)              # Column labels (descriptors = species)
rownames(spe)              # Row labels (objects = sites)
summary(spe)               # Descriptive statistics for columns

## Overall distribution of abundances (dominance codes)
# Minimum and maximum of abundance values in the whole data set
range(spe)
# Minimum and maximum value for each species
apply(spe, 2, range)
# Count the cases for each abundance class
(ab <- table(unlist(spe)))
# Barplot of the distribution, all species confounded
barplot(ab,
  las = 1,
  xlab = "Abundance class",
  ylab = "Frequency",
  col = gray(5 : 0 / 5)
)
# Number of absences
sum(spe == 0)
# Proportion of zeros in the community data set
sum(spe == 0) / (nrow(spe) * ncol(spe))
```

Hint *Observe how the shades of grey of the bars have been defined in the function* **`barplot()`**. *The argument* `col = gray(5 : 0 / 5)` *means "I want five shades of grey with levels ranging from 5/5 (i.e., white) to 0/5 (black)".*

Look at the barplot of abundance classes. How do you interpret the high frequency of zeros (absences) in the data frame?

2.2.3 Species Data: A Closer Look

The commands above give an idea of the data structure. But codes and numbers are not very attractive or inspiring, so let us illustrate some features. We will first create a map of the sites (Fig. 2.2):

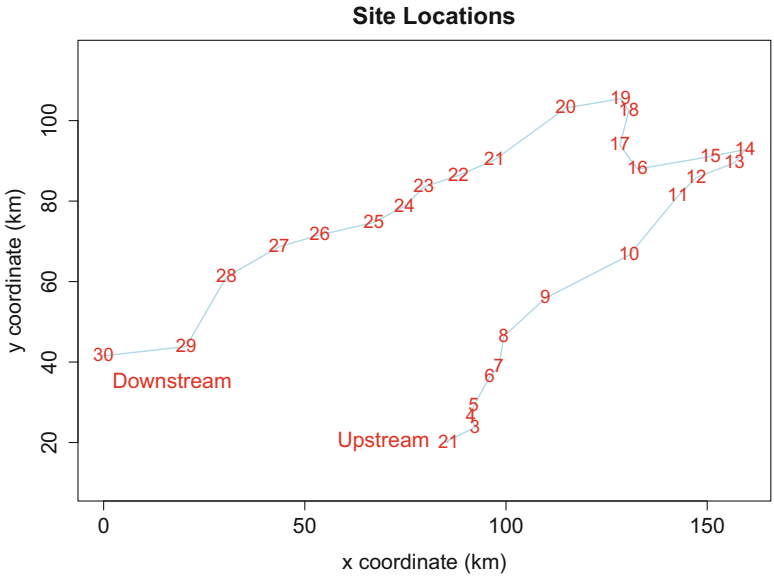

Fig. 2.2 Map of the 30 sampling sites along the Doubs River. Sites 1 and 2 are very close to each other

```
## Map of the locations of the sites
# Geographic coordinates x and y from the spa data frame
plot(spa,
  asp = 1,
  type = "n",
  main = "Site Locations",
  xlab = "x coordinate (km)",
  ylab = "y coordinate (km)"
)
# Add a blue line connecting the sites along the Doubs River
lines(spa, col = "light blue")
# Add the site labels
text(spa, row.names(spa), cex = 0.8, col = "red")
# Add text blocks
text(68, 20, "Upstream", cex = 1.2, col = "red")
text(15, 35, "Downstream", cex = 1.2, col = "red")
```

When the data set covers a sufficiently large area, it is possible to project the sites onto a Google Maps® map:

```
## Sites projected onto a Google Maps® background
# By default the plot method of the googleVis package uses
# the standard browser to display its output.
nom <- latlong$Site
latlong2 <- paste(latlong$LatitudeN, latlong$LongitudeE, sep = ":")
df <- data.frame(latlong2, nom, stringsAsFactors = FALSE)

mymap1 <- gvisMap(df,
  locationvar = "latlong2",
  tipvar = "nom",
  options = list(showTip = TRUE)
)
plot(mymap1)
```

Now the river looks more real, but where are the fish? To show the distributions and abundances of the four species used to characterize ecological zones in European rivers (Fig. 2.3), one can type:

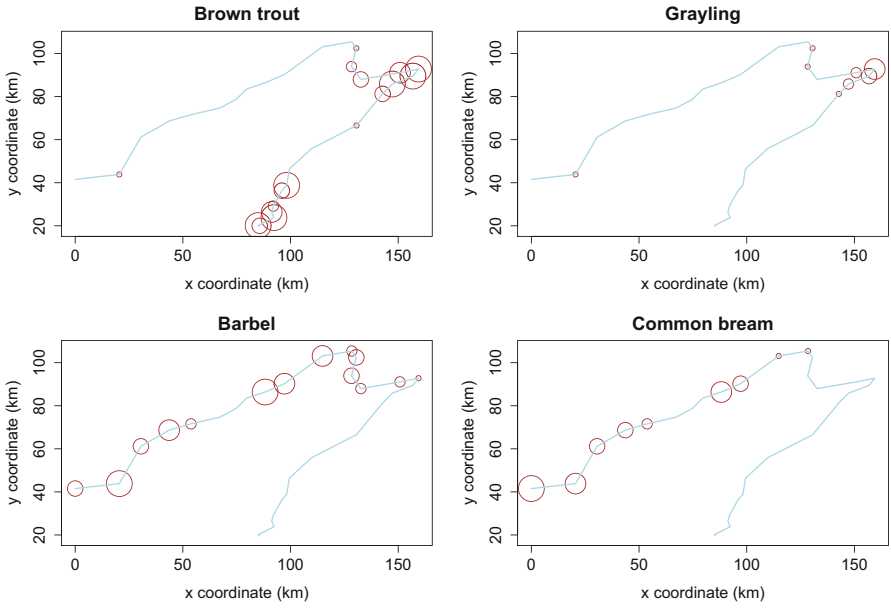

Fig. 2.3 Bubble maps of the abundances of four fish species