Practical exercises –

Spatial, temporal, and space-time analysis of community and other oceanographic data for iAtlantic researchers

Pierre Legendre[1] and Olivier Gauthier[2], ©2021

[1]Département de sciences biologiques, Université de Montréal
[2]Observatoire Marin, Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale

This document gives calculation details in the R statistical language (R Core Team 2021) for spatial and temporal analysis of the *Chesapeake Bay Benthic Monitoring Program* data.

Table of Contents

0. INTRODUCTION TO THE PRACTICAL EXERCISES 2

1. R PACKAGES AND DATA FILES 3
1.1. REQUIRED R PACKAGES 3
1.2. THE DATA USED IN THESE EXERCISES 3
1.3. DESCRIPTION OF THE DATA FILES 4
1.4. PLOT A ROUGH MAP OF THE SITES IN CHESEAPEAKE BAY 6
1.5. PLOT A NICER MAP OF THE SITES, WITH CHESAPEAKE BAY BACKGROUND 6

2. LOAD THE NECESSARY PACKAGES, PREPARE THE DATA MATRICES 7
2.1. LOAD THE R PACKAGES NECESSARY FOR ANALYSIS 7
2.2. REMOVE FRESHWATER SITES #36 AND #79 7

3. PRELIMINARY SPATIAL ANALYSIS OF THE 1999 SURVEYS, SPRING AND FALL TOGETHER 8
3.1. PRINCIPAL COMPONENT ANALYSIS OF THE FAUNAL DATA, 25 SITES 9
3.2. TEST OF SPACE-TIME INTERACTION IN THE ABSENCE OF REPLICATION 9
3.3. TWO-WAY MANOVA OF FAUNAL DATA, 25 SITES, FACTORS SITE AND SEASON 10
3.4. PRINCIPAL COMPONENT ANALYSIS OF THE ENVIRONMENTAL DATA, 25 SITES 11

4. TIME SERIES ANALYSIS OF A SINGLE SITE AND SEASON — SITE 24, SPRING ONLY 12
4.1. SELECT DATA FROM SITE 24 IN THE CENTRE OF CHESAPEAKE BAY 12
4.1.1. Which time scale produced more variation: seasons or years? 12
4.1.2. Same analysis, *years* as a quantitative variable 13
4.2. TIME-CONSTRAINED CLUSTERING OF SITE 24 DATA, SPRING SURVEYS 14
4.2.1. Multivariate regression tree (MRT) analysis of the time series 14
4.2.2. Constrained hierarchical clustering of the time series 15
4.2.3. Search for a linear or polynomial trend in the faunal data 16
4.2.4. Explain the faunal variation by environmental variables with a RDA model 16
4.2.5. Explain the faunal variation by environmental variables with a MRT tree model 17

5. TIME SERIES ANALYSIS OF MULTIVARIATE DATA AT 5 SITES, SPRING SURVEYS, 13 YEARS 18
5.1. SELECT DATA OF SITES 22, 23, 201, 202 AND 203 18
5.2. THREE-WAY MANOVA OF THE FAUNA IN THE SUBSET OF 5 SITES 18
5.3. COMPARE 5 SITES DURING 13 YEARS: MULTIPLE FACTOR ANALYSIS (MFA) 18

6. SPACE-TIME ANALYSIS 21
6.1. CLUSTER THE 25 BRACKISH SITES USING SPACE-CONSTRAINED CLUSTERING 21
6.1.1. Partial RDA of the faunal data controlling for among-year variation 21
6.1.2. Construct a file of link edges describing how the sites are connected 22
6.1.3. Compute constrained hierarchical clustering 22
6.1.4. MRT interpretation of the classification groups using environmental variables 24
6.2. TEST SPACE-TIME INTERACTION FOR COMMUNITY DATA WITHOUT REPLICATION 24
6.3. LCBD ANALYSIS ON 25 BRACKISH SITES, SPRING AND FALL SEPARATELY 25
6.3.1. Compute LCDB indices 25
6.3.2. Compute species richness 26
6.3.3. Relate LCBD indices to environmental variables 26
6.4. TBI ANALYSIS AT THE 25 SITES: COMPARE THE SPRING AND FALL SURVEYS 27
6.4.1. Assemble separate faunal data frames for the spring and fall of 2003 27
6.4.2. TBI analysis of the 2003 data, Spring and Fall 27

REFERENCES 30

0. INTRODUCTION TO THE PRACTICAL EXERCISES

This document describes methods of analysis to answer questions of interest to the iAtlantic multidisciplinary research programme (https://www.iatlantic.eu/) about the changes in community composition through time that have occurred at specific sites or in study regions. In particular, (a) can we identify temporal trends in ecological communities? (b) Can we identify discontinuities (break points, tipping points) in community time series? (c) Are there important changes that took place at specific sites between surveys conducted at different times? These types of changes may be the signature of the effects of climate change on marine communities.

The *Chesapeake Bay Monitoring Program* data used in these exercises embed, at the scale of a marine bay, the types of spatial and temporal variation that iAtlantic research teams are studying over vast areas of the Atlantic Ocean. The data cover 25 brackish sites surveyed during 13 years at two seasons, spring and fall.

The Chesapeake Bay data were used by Legendre & Gauthier (2014) to describe other useful statistical methods of analysis for community composition data that make use of spatial and temporal eigenfunctions (MEM, AEM). These methods are useful to describe the spatial and temporal structures of community data at multiple scales, subjected to non-directional or directional physical processes. These other methods, not described in the present document, are still available for consultation and study in Appendix S2 ("Temporal eigenfunction methods – Practicals in R") of the Legendre & Gauthier paper cited in the References section. The present document reuses a portion (section 1) of that Appendix where the Chesapeake Bay data, reorganised into an RData file, were described.

1. R PACKAGES AND DATA FILES

1.1. REQUIRED R PACKAGES

The following packages, available on CRAN, will be used in these exercises:

```
install.packages(c("ade4","adespatial","ape","FactoMineR","leaflet","vegan"),dependencies=TRUE)

# Manually install mvpart from the GitHub repository; mvpart is no longer available from CRAN.
# On Windows machines, Rtools (4.0 and above) must be installed first. Go to:
# https://cran.r-project.org/bin/windows/Rtools/

# Following that:
install.packages("devtools")
library(devtools)
install_github("cran/mvpart", force = TRUE)

# If the "install_github" command returns an error about the namespace file (this may happen
# due to your computer platform and System version), copy or type the following commands:
assignInNamespace("version_info", c(devtools:::version_info,
                  list("4.0" = list(version_min = "3.3.0",
                  version_max = "99.99.99", path = "bin"))),"devtools")
install_github("cran/mvpart", force = TRUE)
```

1.2. THE DATA USED IN THESE EXERCISES

```
# The dataset used in these applications are taken from the Maryland Data Sets of the
# Chesapeake Bay Benthic Monitoring Program (http://www.baybenthos.versar.com/data.htm),
# a part of the Chesapeake Bay Program (http://www.chesapeakebay.net/). You will find
# detailed information about the sampling protocol on the web page. The whole dataset is made
# available online in numerous .txt files, one per group of variables and per year.

# Scientists studying the present practical exercises will be assumed to be familiar with the methods
of multivariate analysis used in the exercises. Some of these methods have been described recently.
They may want to revise the videos of the course or the references listed in the documentation files
of the functions before running the exercises.

# We compiled and formatted these files in an .Rdata file for immediate use in R. The 'reshape'
# R package (Wickham 2007) was most useful to accomplish this task.
# The file is called "ChesapeakeBay.Maryland.RData".
# Double-click on the RData file, or drag it onto the R icon or in the R console. Else, you can
# type load("ChesapeakeBay.Maryland.RData") if the R console working directory is set to the folder
# containing that file. Check this by typing getwd(). View the list of Chesapeake Bay data files:

ls()

# The Chesapeake Bay data were extracted from the http://www.baybenthos.versar.com/data.htm),
site by Dr Olivier Gauthier, Maître de Conférence (i.e., lecturer) in Numerical ecology and Benthic
ecology at Université de Bretagne Occidentale, France, for the Practical exercises in R published in
Appendix S2 of the Legendre & Gauthier (2014) paper.
```

1.3. DESCRIPTION OF THE DATA FILES

Refer to the *Maryland Dataset Data Dictionary* found on (http://www.baybenthos.versar.com/DOCS/DataDictionaryMD.pdf) for an in-depth description of the environmental variables and sampling protocols.

fauna (702x205) – Abundances of 205 benthic macrofaunal taxa in alphabetic order. This
includes all animals retained on a 0.5 mm sieve. Nearly all (*n* = 203) are invertebrates, but two
chordates (*Molgula manhattensis* and *Branchiostoma caribaeum*) are also encountered in the
retained samples.

sampling (702x6)
STATION, SAMPLE_DATE, SAMP_TYPE, GMETHOD, YEAR, SEASON
STATION – A factor, ID tags 1 to 204 corresponding to 27 sites, each with 26 data rows.
SAMPLE_DATE – Sampling date, from 1996-05-06 to 2008-10-01.
SAMP_TYPE – A factor, FIXED or RANDOM sampling sites. Only the FIXED sites are
included in our RData file; see http://www.baybenthos.versar.com/data.htm for details.
GMETHOD – A factor, four gear types for sampling the benthic macrofauna.
Either "BC-PH" ("Post-Hole digger", 250 cm^2 surface area, *n* = 156), "BC-WC" (Wildco
box corer, 225 cm^2 surface area, *n* = 468), "PP" (Petite Ponar, 250 cm^2 surface area,
n = 26), or "VV-YM" (Van-Veen modified Young Grab, 440 cm^2 surface area, *n* = 52).
YEAR – 13 survey years, from 1996 to 2008.
SEASON – Season, a factor: Fall (*n* = 351) or Spring (*n* = 351).

summary(sampling, maxsum=27)

sediment (702x5)
MOIST, SAND, SILTCLAY, TC, TN
MOIST – Sediment moisture content in percent.
SAND – Sand content in percent by mass.
SILTCLAY – Silt-clay content in percent by mass.
TC – Total carbon content in percent.
TN – Total nitrogen content in percent.

waterquality (702x5)
CONDUCT, DO, PH, SALINITY, WTEMP
CONDUCT – Conductivity in mmho/cm, US equivalent to mS/cm in international notation.
DO – Dissolved oxygen in ppm, US equivalent to mg/L.
PH – pH of water sample.
SALINITY – In practical salinity units (PSU), equivalent to parts per thousand (‰).
WTEMP – Water temperature in Celsius (ºC).

xy (27x2)
LATITUDE and LONGITUDE in decimal degrees for each of the 27 sampling sites.
The original ID tags of the 27 sites are found in vector rownames(xy).

Please note the following decisions that were made in order to produce the data tables used in
the exercises that follow.

1) The *Chesapeake Bay Benthic Monitoring Program* includes both FIXED and RANDOM
sites. FIXED sites were sampled every year whereas RANDOM sites changed from year to
year. We only included the FIXED sites in our data tables.

2) While the monitoring program started in 1995 and is ongoing, we decided to restrict our
analyses to calendar years for which both a Spring (May) and Fall (late August to early
October) sampling were conducted. The dataset thus covers 13 years and 26 sampling
campaigns, spawning from Spring of 1996 to Fall of 2008.

3) Among the environmental variables available about the sediment, we removed Total
Inorganic Carbon (TIC) and Total Organic Carbon (TOC) from the **sediment** file because no
data were available for 1996.

4) Although the Data Dictionary states that SALINITY was measured in Practical Salinity
Units (PSU), data files for 1997 report SALINITY in Parts Per Thousand (PPT). We took this
to be a data entry error and merged the data accordingly.

5) Dissolved oxygen in the water column was available both in Parts Per Million (DO) and as
percent saturation (DO_PSAT). We elected to use only DO due to the fairly large number (16)
of missing values for variable DO_PSAT.

6) For one STATION/SAMPLE_DATE combination (Station 74 on 05/30/2000), the sum of
SAND and SILTCLAY granulometric fractions was greater than 100%. We rescaled these
values for their sum to be 100%.

7) A total of 8 measurements were missing in the environmental data tables: 5 for water quality
(all 5 measurements for site 68 on 05/17/2000) and 3 for sediment (MOIST for site 22
on 09/10/2007, and TC and TN for site 26 on 05/10/1999). In each case, we estimated the
empty cell using the mean value of the variable at the same site during the same season,
computed over the year interval (1996 to 2008) considered here.

8) Within the *Chesapeake Bay Benthic Monitoring Program*, three replicate faunal samples
were scheduled on each sampling occasion. In the fauna data frame, all available samples from
a given sampling occasion were summed. However, for some rare sampling occasions, only 2,
or even only 1, sample was available. This is not a big concern here as all analyses will be
conducted on Hellinger-transformed faunal abundances, where each faunal data vector is first
transformed into relative abundances, then the values are square-rooted. A total of 8 samples,
from an expected total of 2106, were missing due to various field or laboratory mishaps. These
are: samples number 2 and 3 for site 79, sample number 1 for site 68, and sample number 1 for
site 23 in the Spring of 1998; sample number 1 for site 1 in the Spring of 1999; sample number
3 for site 26 in the Fall of 1999; sample number 3 for site 79 in the Spring of 2001; and,
sample number 3 for site 79 in the Fall of 2008.

```
# 1.4. PLOT A ROUGH MAP OF THE SITES IN CHESEAPEAKE BAY

plot(xy[,c(2,1)], xlab="Longitude W", ylab="Latitude N", asp=1)
text(xy[,c(2,1)], labels=rownames(xy), pos=4)


# 1.5. PLOT A NICER MAP OF THE SITES, WITH CHESAPEAKE BAY BACKGROUND
#       using package leaflet (Appendix, Figure 1)

# install.packages("leaflet", dependencies=TRUE)      # If package not already installed
library(leaflet)

# The are many mapping options available in this package. See ?leaflet and ?tileOptions.

# Produce a map of the survey 27 sites in the Chesapeake Bay watershed

sites = paste("Site",rownames(xy),sep=".")
background <- addTiles(leaflet())
ChesapeakeMap <-
  addMarkers(
    background,
    lat = xy$LATITUDE,
    lng = xy$LONGITUDE,
    label = sites,
    labelOptions = labelOptions(noHide = TRUE, textOnly = TRUE)
  )
ChesapeakeMap

# The map will appear in a window. If you are working with the regular R console, it will appear in
your Web browser.
# You can move the map in the browser frame and change the viewing scale using the mouse.
Following that, export it to a pdf file.

# ========
```

2. LOAD THE NECESSARY PACKAGES, PREPARE THE DATA MATRICES

2.1. LOAD THE R PACKAGES USED IN THE FOLLOWING ANALYSES

```
library(ade4)
library(adespatial)
library(FactoMineR)
library(mvpart)
library(vegan)
```

2.2. REMOVE FRESHWATER SITES #36 AND #79

The freshwater benthic fauna is taxonomically very different from the brackish water fauna. Keep the 25 brackish sites sampled in the spring and fall surveys during 13 years: 25*2*13 = 650 rows in the data files.

The following data files will be used in sections 3 and 6 of this document.

```
freshwater <- which(sampling$STATION %in% c(36,79))
fauna.25 <- fauna[-freshwater,]                          # 650 x 205
sampling.25 <- sampling[-freshwater,]                    # 650 x 6
waterquality.25 <- waterquality[-freshwater,]            # 650 x 5
sediment.25 <- sediment[-freshwater,]                    # 650 x 5
xy.25 <- xy[-c(12,27),]                                  # 25 x 2
```

Remove unused factor levels from the sampling.25 data file
Function drop.levels() is found in folder “Chesapeake-Functions”, in the “Chesapeake Bay practical exercises” folder. Load this function.
A similar function with the same name (but different code) is available in R package {gdata}.

```
sampling.25 <- drop.levels(sampling.25)
```

========

3. PRELIMINARY SPATIAL ANALYSIS OF THE 1999 SURVEYS, SPRING AND FALL TOGETHER

The objective of this spatial analysis is to determine whether, and how, the study sites should be split up in order to produce useful temporal analyses in sections 4, 5 and 6 of this document.

Are the data homogeneous enough to be analysed as one group when looking for temporal structure (temporal trend or abrupt breaks in the series), or should they be divided into groups that should be analysed separately?

Sampling year 1999 was selected for analysis in this section. It is abbreviated to "1999' or '99' in the R code that follows. This analysis can be repeated with the data of any of the other 12 years.

```
year.1999 = which(sampling.25$YEAR == 1999)
length(year.1999)   # 25 sites * 2 seasons = 50 data vectors
# Check that the selection is correct:  ( tmp = sampling.25[year.1999,c(1,6)] )

# Edit the fauna data frame
fauna.25.1999 = fauna.25[year.1999,]
dim(fauna.25.1999)   # [50 205]

# Fauna: remove columns with colSums=0
fauna.25.99 = fauna.25.1999[ , colSums(fauna.25.1999)!=0]
dim(fauna.25.99)   # [50 84]; 84 species are kept

# Edit the sampling data frame, selecting the data rows for year 1999
sampling.25.99 = sampling.25[year.1999,]              # dim = [50 6]

# Copy the Spring and Fall surveys to separate data frames
spring = seq(1, 50, by=2)
fall = spring+1
fauna.25.99.S = fauna.25.99[spring,]
fauna.25.99.F = fauna.25.99[fall,]
# Shorten the row and column names before plots are produced
rownames(fauna.25.99.S) = paste("S",sampling.25.99[spring,1], sep=".")
rownames(fauna.25.99.F) = paste("F",sampling.25.99[fall,1], sep=".")
colnames(fauna.25.99.S) = paste("Sp",1:84,sep=".")
colnames(fauna.25.99.F) = paste("Sp",1:84,sep=".")
# Join the two fauna data frames, Spring and Fall, into a single data frame
fauna.25.99.SF = rbind(fauna.25.99.S,fauna.25.99.F)   # dim = 50 84

# In the combined data frame, the observation are divided into Spring and Fall surveys.
# Within each season, the sites are labelled in the same order as in the sampling data frame and its
derivative, file sampling.25.99.

# The Hellinger transformation is applied to the faunal data before the analyses by linear methods
(PCA and multivariate Manova), as in the Legendre & Gauthier (2014) paper.

fauna.25.99.hel = decostand(fauna.25.99.SF,"hellinger")
```

3.1. PRINCIPAL COMPONENT ANALYSIS OF THE FAUNAL DATA, 25 SITES

PCA ordination of the Hellinger-transformed faunal data, Spring and Fall together

```
pca.sp.out = rda(fauna.25.99.hel)
print(pca.sp.out)
```
Note the value of Total inertia. This is the total beta diversity, on a [0,1] scale, measured as the variance of the Hellinger-transformed community data for the Spring and Fall surveys together.

Plot the PCA results (Appendix, Figure 2)
```
site.sc1 = scores(pca.sp.out, display="sites", scaling=1)
gr = c(rep(1,25), rep(2,25))
p = plot(pca.sp.out, display="sites", scaling=1, type="n", main="PCA fauna, 25 sites 1999")
abline(v=0, h=0, lty=2, col="grey")

# Draw the points in two colours: spring = green, fall = red
mycol = c("green","red")
for(i in 1:2) {points(site.sc1[gr==i,], pch=(14+i), cex=2, col=mycol[i])}

# Add season-site labels
text(site.sc1,rownames(fauna.25.99.hel), cex=0.7, pos=2)

# Add legend interactively: click where you want the legend to be printed on the graph
surveys = c("Spring", "Fall")
legend(locator(1), surveys, pch=(14+c(1:2)), col=mycol[c(1:2)], pt.cex=2)
```
Are the spring and fall survey data points well mixed in the plot, or are they separated?

3.2. TEST OF SPACE-TIME INTERACTION IN THE ABSENCE OF REPLICATION

We will test **the site × season interaction** for the 1999 surveys in the absence of replication, using the function stimodels.R of package {adespatial}

The data are organised by season blocks (S and F seasons), with the 25 sites nested in each block.
The data are organized as in the documentation file of function stimodels.R, with *sites* nested
in *times*. Run the space-time interaction test using model "5".

In this run, we need to provide the spatial coordinates of the 25 sites, file xy.25 from Section 2.2

```
stimodels(fauna.25.99.hel, S=xy.25, Ti=2, nperm=9999, model="5")
```

Excerpt from the output file
```
Interaction test:   R2 = 0.0752   F = 1.0339   P( 9999 perm) = 0.4174
Space test:         R2 = 0.5942   F = 2.0425   P( 9999 perm) = 1e-04
Time test:          R2 = 0.1124   F = 9.2759   P( 9999 perm) = 1e-04
```

Model 5: the main factors space and time are coded using Helmert contrast variables, but the interaction is computed as the Hadamard product between dbMEM for Space and Helmert for Time.

The season × site interaction is not significant (p = 0.4174). Hence we can proceed with the interpretation of the contributions of the main factors, space = *sites* and time = *season*.

3.3. TWO-WAY MANOVA OF FAUNAL DATA, 25 SITES, FACTORS SITE AND SEASON

Repeat the analysis of the main factors, space = *sites* and time = *season*, using a more classical form of analysis.

Conduct a two-way Manova to determine if *season* in an important factor distinguishing the Spring and Fall faunas, as the PCA plot indicates. Use the adonis2() function of vegan, also with **permutation** tests. The analysis is without replication; hence the interaction between the factors cannot be tested by function adonis2, which is classical two-way or multiway manova, improved by the use of permutation tests.

Note – An example presenting a test for space-time interaction in the absence of replication will be demonstrated in section 5.3.

First, create simple factors for Sites and Seasons. The order of the site labels is that of the observations in the *fauna.25.99.SF* data frame.

```
tmp = as.character(sampling.25.99$STATION[spring])
site.f = factor(c(tmp, tmp))                    # Factor for sites
season.f = c(rep("S",25),  rep("F",25))         # Factor for seasons

( out = adonis2(fauna.25.99.hel ~ site.f + season.f, method="eucl", by="terms") )
```

Compare this output file to the output of function stimodels in section 3.2. Check that the R-squares (columns ‘R2’) for factors space = *sites* and time = *season* are identical in the two analyses.

The PCA ordination plot and the two-way Manova of the community composition data clearly indicate that the data from the spring and fall surveys do not belong to the same statistical population. Thus their temporal structures should be analysed separately.

3.4. PRINCIPAL COMPONENT ANALYSIS OF THE ENVIRONMENTAL DATA, 25 SITES, 1999 SURVEYS

```
# Prepare the data files for PCA
waterquality.25.99 = waterquality.25[year.1999,]
sediment.25.99 = sediment.25[year.1999,]

envir.25.99 = cbind(waterquality.25.99, sediment.25.99)     # dim = 50 10
envir.25.99.S = envir.25.99[spring,]
envir.25.99.F = envir.25.99[fall,]

# Before the PCA plot, simplify rownames and write variable names with small letters
rownames(envir.25.99.S) = paste("S",sampling.25.99[spring,1], sep=".")
rownames(envir.25.99.F) = paste("F",sampling.25.99[fall,1], sep=".")
colnames(envir.25.99.S) = colnames(envir.25.99.F) =
    c("Cond","DO","pH","Salinity","W.temp","Moist","Sand","SiltClay","TC","TN")

# Join the two fauna data frames, Spring and Fall, into a single data frame
envir.25.99.SF = rbind(envir.25.99.S,envir.25.99.F)   # dim = 50 10

# Standardize the environmental variables at the beginning of PCA: use argument "scale=TRUE "

pca.env.out = rda(envir.25.99.SF, scale=TRUE)      # scale=TRUE: standardize the variables
print(pca.env.out)

# Plot the PCA results (Appendix, Figure 3)
site.sc1 = scores(pca.env.out, display="sites", scaling=1)
gr = c(rep(1,25), rep(2,25))
p = plot(pca.env.out, display="sites", scaling=1,type="n", main="PCA Environment, 25 sites 1999")
abline(v=0, h=0, lty=2, col="grey")

# Draw the points in two colours: spring = green, fall = red
mycol = c("green","red")
for(i in 1:2) {points(site.sc1[gr==i,], pch=(14+i), cex=2, col=mycol[i])}

# Add season-site labels
text(site.sc1,rownames(fauna.25.99.hel), cex=0.7, pos=2)

# Add legend interactively: click where you want the legend to be printed on the graph
surveys = c("Spring", "Fall")
legend(locator(1), surveys, pch=(14+c(1:2)), col=mycol[c(1:2)], pt.cex=2)

# The PCA ordination of the environmental variables confirms that the data from the spring and fall
surveys do not belong to the same statistical population.

# Exercise –
# Can you compute a two-way Manova with function adonis2 without a line-by-line script?
# Using the example of section 3.2, compute a two-way Manova with the environmental variables as
response data, file envir.25.99.SF, and files site.f and season.f of section 3.2 as factors. Make sure
the response data envir.25.99.SF are standardized before the analysis.
```

4. TIME SERIES ANALYSIS OF A SINGLE SITE AND SEASON — SITE 24, SPRING ONLY

Site 24 is a brackish water site located in the centre of Chesapeake Bay.

4.1. SELECT DATA FROM SITE 24 IN THE CENTRE OF CHESAPEAKE BAY

Load function "select.site.R", found in folder “Chesapeake-Functions” in the course material.

Function "select.site.R" selects data for a specified site of the Chesapeake Bay data file and writes the data rows into separate files called "fauna", "sediment" and "waterqual", for either the spring survey (S), the fall survey (F), or the spring and fall surveys (SF), over the 13 years of the study. An additional output vector "Dates" shows the selected sampling dates.

Load file "ChesapeakeBay.Maryland.RData", which contains the data frames "fauna", "sampling", "sediment" and "waterquality."
Load function "select.site.R"

Run the function to obtain data files for site #24, for the spring surveys (season="S") and all surveys (season="SF").

```
out.24.S = select.site(siID=24, season="S")
summary(out.24.S)
#                Length Class      Mode
# fauna.24.S     30     data.frame list    # dim = 13 30
# sediment.24.S   5     data.frame list
# waterqual.24.S  5     data.frame list
# Dates          13     Date       numeric

# Check the sampling dates incorporated in these data files
out.24.S$Dates

out.24.SF = select.site(siID=24, season="SF")
summary(out.24.SF)
#                 Length Class      Mode
# fauna.24.SF     36     data.frame list    # dim = 26 36
# sediment.24.SF   5     data.frame list
# waterqual.24.SF  5     data.frame list
# Dates           26     Date       numeric

# Check the sampling dates incorporated in these data files
out.24.SF$Dates
```

4.1.1. Which time scale produced more variation: seasons or years?

In the analysis of fauna.24.SF, we will have no degrees of freedom left to test the *seasons*years* interaction because factor *years* is a factor and there is no replication at the *seasons*years* level.

```
# Hellinger transformation of the faunal data as in the Legendre & Gauthier (2014) paper.
# Use vegan’s function decostand()
fauna.24.SF.hel = decostand(out.24.SF$fauna.24.SF, "hellinger")          # dim = 26 36
```

```
# The Hellinger transformation can also be computed as follows:
        row.sums = rowSums(out.24.SF$fauna.24.SF)
        faunal.data.hel = sqrt(sweep(out.24.SF$fauna.24.SF, 1, row.sums, "/"))
```

Manova by RDA of the faunal data against factors *Seasons* and *Years*, using function adonis2().

```
years = as.factor(rep(1996:2008, 2))
seasons = rep(c("S","F"), 13)
```

Manova by RDA of the faunal data against factors seasons and years, using function adonis2(). The interaction between seasons and years cannot be tested because there is no replication.

```
( res = adonis2(fauna.24.SF.hel ~ seasons+years, met="eucl", by="term") )
# Permutation test for adonis under reduced model
# Terms added sequentially (first to last)
# Permutation: free
# Number of permutations: 999
#
# adonis2(formula = fauna.24.SF.hel ~ seasons + years, method = "eucl", by
= "term")
#          Df SumOfSqs      R2      F Pr(>F)
# seasons   1   1.6189 0.20107 7.2103  0.001 ***
# years    12   3.7381 0.46428 1.3874  0.067 .
# Residual 12   2.6943 0.33464
# Total    25   8.0512 1.00000
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We find that the factor *seasons* is more significant than *years* for the fauna at site 24. Seasonal surveys are much more variable, with p = 0.001. Yearly differences are also significant, but at a lower significance level, with p = 0.067.

4.1.2. Same analysis, *years* as a quantitative variable, assuming a linear effect of years on fauna.
In this analysis, we will have degrees of freedom left to test the *seasons***years* interaction

```
years2 = rep(1996:2008, 2)
( res2 = adonis2(fauna.24.SF.hel ~ seasons*years2, met="eucl", by="term") )

# Permutation test for adonis under reduced model
# Terms added sequentially (first to last)
# Permutation: free
# Number of permutations: 999
#
# adonis2(formula = fauna.24.SF.hel ~ seasons * years2, method = "eucl", by
= "term")
#                Df SumOfSqs      R2      F Pr(>F)
# seasons         1   1.6189 0.20107 6.3600  0.001 ***
# years2          1   0.5219 0.06482 2.0504  0.048 *
# seasons:years2  1   0.3105 0.03857 1.2199  0.262
# Residual       22   5.5999 0.69554
# Total          25   8.0512 1.00000
# ---
```

The season*years interaction is not significant; hence we can interpret the results about the two main factors. Both *seasons* and *years2* are significant.

4.2. TIME-CONSTRAINED CLUSTERING OF SITE 24 DATA, SPRING SURVEYS

The objective of these analyses is to search for breakpoints in multivariate data series.

4.2.1. Multivariate regression tree (MRT) analysis of the time series
About MRT: refer to the course teaching document "Multivariate regression tree analysis.pdf"

Function mvpart() of package {mvpart} will be used.
Can we identify breaks in the data series? Use time-constrained clustering.

Multivariate regression tree analysis is an extension of Classification and Regression Tree (CART) analysis to multivariate response data. The analysis involes a response data matrix Y and a matrix X containing explanatory variables. Each matrix may contain univariate or multivariate data. The variables in X may be quantitative or factors.

The MRT method tries to identify a succession of breakpoints in the **Y** matrix that minimise the sum of the within-group sums-of-squares while being related to a breakpoint in one of the **X** variables. The result presents itself in the form of a tree where the objects are successively split in two groups. A cross-validation procedure is used to limit the size (i.e. the number of splits) of the tree.

The MRT method is used here as a form of space- or time-constrained clustering method, as proposed by Borcard et al. (2011, 2018). The constraining variable in **X** is a numerical variable indicating the position of the sites along a transect, or the position of surveys along time. The results will be the same if one uses real numbers to describe the geographic or temporal positions, or a sequence of integers. In this section, the **X** variable will be a series of integers (1:13) representing the 13 sampling years.

```
order = as.data.frame(1:13)
colnames(order) = 'order'

# Hellinger transformation of the faunal data, spring surveys, site 24
fauna.24.S.hel = decostand(out.24.S$fauna.24.S, "hellinger")                    # dim = 13 30

res.part = mvpart(data.matrix(fauna.24.S.hel) ~ order, data=order, xv="pick", xvmult=100)

# Two groups: surveys [1-4, 5-13], which corresponds to year clusters [1996-1999, 2000-2008]
# Pick two groups. Vector of constrained clustering results, two groups:
( MRT.24F.2gr = res.part$where )
# [1] 2 2 2 2 3 3 3 3 3 3 3 3 3

# Three groups: surveys [1-4, 5-9, 10-13], year clusters [1996-1999, 2000-2004, 2005-2008]
# Repeat the MRT analysis; pick three groups (Figure 4)
# Vector of constrained clustering results, three groups:
( MRT.24F.3gr = res.part$where )
# [1] 2 2 2 2 5 5 5 5 5 4 4 4 4
```

# Save the classification in 3 groups, to be drawn on top of the PCA ordination (next step)
MRT.24F.3gr = data.frame(MRT.24F.3gr)

Plot the PCA ordination of the 13 data years, Chesapeake site #24. Add convex hulls drawn around points of the MRT 3-group classification (Figure 5)

```
pca.out.24S = rda(fauna.24.S.hel)
plot(pca.out.24S, type="text", scaling="sites", display="sites")
pl <- with(MRT.24F.3gr, ordihull(pca.out.24S, MRT.24F.3gr,
scaling="sites", col=1:3, draw="polygon", label=TRUE))
```

Function ordihull draws convex hulls around groups of sites. It returns an object, called "pl" in the present script, where the positions of the surveys in the ordination are shown.
The summary of that object shows the centroid coordinates and areas of the three convex hulls.
summary(pl):

```
#                 2           4          5
# PC1  -0.29464701  0.20239599 0.08467674
# PC2   0.04345104 -0.09264720 0.07685918
# Area  0.01082039  0.01300107 0.11304220
```

The most important break in the "site 24" time series is between time #4 (1999) and time #5 (2000); see Figure 5. It corresponds to the classification in two groups. R-square = 1 – error = 1 – 0.677 = 0.323.

The classification in three groups has an R-square = 1 – error = 1 – 0.581 = 0.419; the second split added only 0.096 to the R-square. Hence the first split is the most important. Figure 5 shows that the group labelled 2, with years 1996–a999, is well separated from the other two groups.

4.2.2. Constrained hierarchical clustering of the time series

Function constr.hclust() of package {adespatial} will be used.
About this method: refer to the course document "Space-constrained hierarchical clustering.pdf"
Use again the faunal data (Hellinger transformed) of site 24, spring surveys, as in subsection 4.2.1

Compute Hellinger distance among years. This is the Euclidean *D* of Hellinger-transformed data

```
fauna.24.S.Dhel <- dist(fauna.24.S.hel)
grpWD2cst_fauna <- constr.hclust(fauna.24.S.Dhel, method="ward.D2",
    chron=TRUE, coords=1996:2008)
```

Plot the classifications into 2, 3 and 4 groups (Figure 6)

```
par(mfrow=c(3,1))                  # If required, adjust these parameters to your screen size
for(k in 2:4) plot(grpWD2cst_fauna, k=k, las=1, xlab="Years",
    xlim=c(1996,2008), cex=3)
```

This method may not produce the exact same results as constrained partitioning by MRT in section 4.2.1 because the methods differ: MRT is a partitioning method (no hierarchy), whereas constr.hclust is a hierarchical clustering method. For this example, there is no difference in the outputs of the two methods, but there may be differences with other examples.

=> Another example of constrained clustering, concerning the Doubs River fish data (spatial series: abundances of 27 fish species at 29 sites along the Doubs River in eastern France), is provided in the documentation file of function constr.hclucst.R, second example –

• Carry out constrained clustering following the script in the documentation file of constr.hclust.
• Carry out space-constrained clustering of the Doubs River fish data using mvpart.R, as demonstrated in subsection 4.2.1.

4.2.3. Search for a linear or polynomial trend in the faunal data

A clustering method will always find clusters, even in continuous data. We cannot test statistically for the presence of 2, 3 or 4 groups in the faunal data because the groups have been obtained from an analysis of these data. – We can only test the conservative hypothesis that a polynomial trend through the multivariate data would represent the faunal variation better than a linear trend. With such a short time series (13 time points), it is not possible to test for the presence of more complex structures.

Construct a cubic polynomial of the year series 1 to 13. Function poly generates orthogonal polynomial function of the vector provided to the function, 1:13 in this example.

```
poly3 = poly(1:13, degree=3)
```

Forward selection of the monomials (polynomial terms) with respect to the multivariate faunal data, 13 years. Function forward.sel.R of adespatial performs variable selection through partial canonical analysis.
Alternative functions for selection would be ordistep.R and ordiR2step.R in package vegan.

```
( sel.res = forward.sel(fauna.24.S.hel, poly3) )
```

The only significant trend (p = 0.002) is linear (i.e. degree=1). R-square = 0.281, adjusted R-square = 0.216.

Conclusion: the variation in the multivariate faunal data is conservatively well modelled by a linear temporal trend.

4.2.4. Explain the faunal variation by environmental variables with a RDA model

We will use sediment and water quality in a linear RDA model. For illustration purpose, we will use a very large alpha value in order to select a variable. Alpha = 0.05 would select no variable.

(sel.res = forward.sel(fauna.24.S.hel, out.24.S$sediment.24.S, alpha=0.50))

Explain the faunal variation by the variation of the water quality variables

(sel.res = forward.sel(fauna.24.S.hel, out.24.S$waterqual.24.S, alpha=0.50))

The available environmental variables provided no significant linear predictors for the faunal data at $\alpha = 0.05$. An alternative hypothesis is that the observed variation is neutral with respect to the environmental variables. I may depend on random settling of invertebrate larvae among the years.

```
# 4.2.5. Explain the faunal variation by environmental variables with a MRT tree model

# We will use again sediment and water quality in a linear MRT tree model (function mvpart.R).
# The MRT method does not take temporal contiguity into account, except when it is explicitly used
as a form of time-constrained clustering method, as it was done in section 4.2.1.
# About MRT: refer to the course teaching document "Multivariate regression tree analysis.pdf"

X = cbind(out.24.S$sediment.24.S, out.24.S$waterqual.24.S)
res.part = mvpart(data.matrix(fauna.24.S.hel) ~ ., data=X, xv="pick", xvmult=100)

# In this example, the cross-validation minimum error is for 1 group. This indicates that the
available environmental variables are not capable of splitting the 13 surveys into statistically
meaningful groups.

# For illustration of the method, the mvpart tree with 3 groups is shown (Figure 7a).
# In addition, a PCA plot of the tree with 3 groups is produced (Figure 7b)
rpart.pca(res.part, interact=FALSE, wgt.ave=FALSE)

# This example is still useful as an illustration of how to compute a multivariate regression tree
analysis using a file of environmental variables as the explanatory data X.

# Examples where this analysis produced a meaningful explanatory model are provided in the
De'ath (2002) paper and in the Numerical ecology with R book (Borcard et al. 2018), section 4.12.2.
```

5. TIME SERIES ANALYSIS OF MULTIVARIATE DATA AT 5 SITES, SPRING SURVEYS, 13 YEARS

5.1. SELECT DATA OF SITES 22, 23, 201, 202 AND 203

```
curr.siteS <- c(201, 202, 203, 22, 23)
sampling.manova <- sampling[sampling$STATION %in% curr.siteS, ]     # 130 6
fauna.manova <- fauna[sampling$STATION %in% curr.siteS, ]           # 130 205

# Remove the "empty" taxa; Hellinger transformation
fauna.manova <- fauna.manova[ , colSums(fauna.manova)!=0]          # (130x70)
fauna.manova.hel <- decostand(fauna.manova, "hellinger")           # (130x70)
```

5.2. THREE-WAY MANOVA OF THE FAUNA IN THE SUBSET OF 5 SITES

Fauna versus sites, seasons, years. Use the adonis2() function of vegan, with permutation tests. The analysis is without replication; hence the interaction between the factors cannot be tested.

```
site.fac <- sampling.manova$STATION
season.fac <- sampling.manova$SEASON
year.fac <- as.factor(sampling.manova$YEAR)

( out = adonis2(fauna.manova.hel ~ site.fac+season.fac+year.fac, method="eucl", by="terms") )

Permutation test for adonis under reduced model
Terms added sequentially (first to last)
Permutation: free
Number of permutations: 999

adonis2(formula = fauna.manova.hel ~ site.fac + season.fac + year.fac,
method = "eucl", by = "terms")
            Df SumOfSqs      R2        F Pr(>F)
site.fac     4   22.779 0.28181 16.9898  0.001 ***
season.fac   1   12.430 0.15378 37.0846  0.001 ***
year.fac    12    8.082 0.09998  2.0093  0.001 ***
Residual   112   37.540 0.46443
Total      129   80.830 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova table shows that the three factors significantly influence the faunal data. It also shows that for the 5 sites located in the bay near Baltimore, Maryland, factor site (R2 = 0.28) has a greater influence than seasons (R2 = 0.15), which in turn has a greater influence than years (R2 = 0.10).

5.3. COMPARE 5 SITES DURING 13 YEARS: MULTIPLE FACTOR ANALYSIS (MFA)

Analyse Spring surveys only, 5 sites, 13 years, using function (MFA) of package {FactoMineR}
Prepare the data files

```
out.22.S = select.site(siID=22, season="S", trim.fauna=FALSE)
out.23.S = select.site(siID=23, season="S", trim.fauna =FALSE)
```

```
out.201.S = select.site(siID=201, season="S", trim.fauna =FALSE)
out.202.S = select.site(siID=202, season="S", trim.fauna=FALSE)
out.203.S = select.site(siID=203, season="S", trim.fauna=FALSE)

fauna.5sites.S = rbind(out.22.S$fauna.22.S, out.23.S$fauna.23.S, out.201.S$fauna.201.S,
out.202.S$fauna.202.S, out.203.S$fauna.203.S)
fauna.5sites.S = fauna.5sites.S [ ,colSums(fauna.5sites.S)!=0]     # 61 species kept in file
fauna.5sites.S.hel = decostand(fauna.5sites.S, "hellinger")      # Hellinger transformation
dim(fauna.5sites.S.hel)   # [65 61]

sediment.5sites.S = rbind(out.22.S$sediment.22.S, out.23.S$sediment.23.S,
out.201.S$sediment.201.S, out.202.S$sediment.202.S, out.203.S$sediment.203.S)
dim(sediment.5sites.S)   # [65 5]

waterqual.5sites.S = rbind(out.22.S$waterqual.22.S, out.23.S$waterqual.23.S,
out.201.S$waterqual.201.S, out.202.S$waterqual.202.S, out.203.S$waterqual.203.S)
dim(waterqual.5sites.S)   # [65 5]

# Regroup the three data tables
tab3 = data.frame(fauna.5sites.S.hel, sediment.5sites.S, waterqual.5sites.S)
dim(tab3)   # [65 71]

# Create vector indicating the number of variables in each group
( grn <- c(ncol(fauna.5sites.S.hel), ncol(sediment.5sites.S), ncol(waterqual.5sites.S)) )   # [61  5  5]

# Compute the MFA without multiple plots
t3.mfa <- MFA(
  tab3,
  group = grn,
  type = c("c", "s", "s"),              # "c": quantitative var.; "s": standardize these quantitative var.
  ncp = 2,
  name.group = c("Fauna", "Sediment", "Water quality"),
  graph = FALSE
)

t3.mfa

# Contents of file “t3.mfa”
# -----
**Results of the Multiple Factor Analysis (MFA)**
The analysis was performed on 65 individuals, described by 71 variables
*Results are available in the following objects :

  name                 description
1 "$eig"               "eigenvalues"
2 "$separate.analyses" "separate analyses for each group of variables"
3 "$group"             "results for all the groups"
4 "$partial.axes"      "results for the partial axes"
5 "$inertia.ratio"     "inertia ratio"
6 "$ind"               "results for the individuals"
```

```
7 "$quanti.var"         "results for the quantitative variables"
8 "$summary.quanti"     "summary for the quantitative variables"
9 "$global.pca"         "results for the global PCA"
# -----

# summary(t3.mfa)    # Eigenvalues, etc.
# t3.mfa$ind         # A long list containing all the results, separated in different matrices

# Plot the results
dev.new(title = "Partial axes", noRStudioGD = TRUE)
plot(t3.mfa,                        # Joint PCA, two canonical axes per data matrix (Figure 8a)
    choix = "axes",
    habillage = "group",
    shadowtext = TRUE)

dev.new(title = "Quantitative variables", noRStudioGD = TRUE)
plot(t3.mfa,                        # Joint PCA, all variables with correlation circle (Figure 8b)
    choix = "var",                  # Blow up the graph to see all variables more clearly
    habillage = "group",
    shadowtext = TRUE)
```

An alternative way to plot these results is described in file "chap6.R" of the book *Numerical ecology with R* (Borcard et al. 2018). On lines 1604–1643, the script shows how to plot *only* the variables, including species, that are significantly correlated to the first two ordination axes.

```
# Compute RV coefficients with tests; p-values are above the diagonal of the result matrix (below)
rv.p <- t3.mfa$group$RV
rv.p[1, 2] <- coeffRV(fauna.5sites.S.hel, scale(sediment.5sites.S))$p.value
rv.p[1, 3] <- coeffRV(fauna.5sites.S.hel, scale(waterqual.5sites.S))$p.value
rv.p[2, 3] <- coeffRV(scale(sediment.5sites.S), scale(waterqual.5sites.S))$p.value
round(rv.p[-4, -4], 6)

#                   Fauna Sediment Water quality
# Fauna          1.000000 0.000003       0.0e+00
# Sediment       0.227240 1.000000       2.2e-05
# Water quality 0.441901 0.211516       1.0e+00
```

The fauna is more strongly related (RV = 0.44) to water quality than to sediment characteristics.

```
# Eigenvalues, scree plot, broken stick model (Figure 8c). Load screestick.R from functions folder
ev <- t3.mfa$eig[, 1]
names(ev) <- paste("MFA", 1 : length(ev))
dev.new(
  title = "MFA eigenvalues and broken stick model",
  noRStudioGD = TRUE
)
screestick(ev, las = 2)
```

6. SPACE-TIME ANALYSIS

Data files were generated in section 2.1 containing the 25 brackish sites, after removing freshwater sites #36 and #79 that were also found in the Chesapeake Bay data sets.

6.1. CLUSTER THE 25 BRACKISH SITES USING SPACE-CONSTRAINED CLUSTERING

In this section, we will analyse the 25 brackish sites only. They are found in the following files prepared in section 2.2, where the two freshwater sites #36 and #79 were removed from the original data files.

Keep the 25 brackish sites sampled in spring and fall surveys, 13 years: 25*2*13 = 650 data rows

```
fauna.25                # 650 x 205
sampling.25             # 650 x 6
waterquality.25         # 650 x 5
sediment.25             # 650 x 5
xy.25                   # 25 x 2
```

Agenda for this section – Analysis of 25 sites, 13 years, 2 surveys per year

• We will first compute a partial RDA of the faunal data in order to obtain ordination axes for the 25 brackish sites while controlling for the among-year and seasonal variation.

• Then we will construct a file describing how are connected the sites that are neighbours on the Chesapeake map.

• We will use that file as the spatial contiguity constraint in clustering. The result of this analysis will be a map of the space-constrained geographic site groups based on analysis of the fauna.

6.1.1. Partial RDA of the faunal data controlling for among-year variation

```
# Hellinger transformation of the faunal data
fauna.25.hel = decostand(fauna.25, "hellinger")
```

Note – In file “sampling.25”, Station and Season are factors, but Year is not a factor. We will turn it into a factor in the call to the partial RDA.

```
(rda.out = rda(fauna.25.hel ~ STATION + Condition(factor(YEAR)+SEASON), data=sampling.25))
```

```
# How many canonical axes are there in the RDA output file? Look into the output object rda.out
rda.out$CCA$rank   # 24
```

rda.out output file: extract the file with 25 rows containing the centroid coordinates of the 25 sites (display = "cn"). Scaling=1: we are interested in the distances among sites in canonical space.

```
centroids.sites = scores(rda.out, scaling = 1, choices=1:24, display = "cn")          # [25 24]
dim(centroids.sites)                # dim = 25 24
```

```
# Function rda.R puts the site names in alphabetic order in the centroids.sites file.
# Rearrange the rows of file centroids.sites in such a way that the sites are in the same order as in the files fauna.25, xy.25, etc.

# To accomplish that, design a "reorder" vector
( reorder = c(1,3,22:25,4:14,2,15:21) )
 [1]  1  3 22 23 24 25  4  5  6  7  8  9 10 11 12 13 14  2 15 16 17 18 19 20 21

# Then, reorder the rows of the file of centroids

centroids.sites.o = centroids.sites[reorder,]                    # dim = 25 24

# Compare the order of the site names in the xy.25 file and in the reordered file of centroids
rownames(xy.25)
rownames(centroids.sites.o)
```

6.1.2. Construct a file of link edges describing how the sites are connected

```
# That file, called "links" in the function, has two columns, "From" and "To". It is described in the documentation file of function constr.hclust
?constr.hclust

# That file was written by hand in a text file because it should contain only edges (i.e. possible routes for the fauna) that are aquatic. Edges over land are not included in the file. (We don't know of any R software to remove these edges automatically.)
# The file is called E50. It contains 50 link edges. It is found in the "Chesapeake-Data" folder.

E50 = read.table(file.choose(), header=TRUE)                     # Read text file "E50.txt"
rownames(E50) = paste("Edge",1:50,sep=".")                       # Cosmetic change: add row names

# Examine the file of link edges
head(E50) ; tail(E50)

# The present version of function constr.hclust, found in package adespatial version 0.3-13, requires a file of geographic coordinates where Longitude is in the first column. That may be corrected in a later version of the function. Generate file "coo.25" with Longitude followed by Latitude:

coo.25 = xy.25[,c(2,1)]
```

6.1.3. Compute constrained hierarchical clustering

```
# Use function constr.hclust.R of package adespatial. This function implements hierarchical agglomerative clustering based on the Lance & Williams algorithm modified to use a constraint of geographic or temporal contiguity. See Legendre & Legendre (2012, section 8.5.9) for details. We will use option method="ward.D2".  The geographic constraints are provided by file E50.
```

```
c.hclust.out <-
constr.hclust(
     dist(centroids.sites.o),          # Response dissimilarity matrix
     method="ward.D2",                 # Clustering method
     links=E50,                        # File of link edges (constraint)
     coords=coo.25)                    # File of geographic coordinates
```

Ward's minimum variance method selected here (method="ward.D2") is a general-purpose clustering method optimizing Ward's (1963) objective least-squares criterion.

This analysis can produce maps of the space-constrained geographic site groups based on analysis of the fauna (Figure 9a). The following lines of code produce the map for k=5 space-constrained groups. The site labels are carefully positioned so that they do not overlap and can be read; this requires the following lines of code:

```
pos1 = c(8,15)                          # Write the labels down
pos2 = c(4,11,17,24)                    # Write the labels left
pos3 = c(2,5,6,7,9,10,12,14,18:23,25) # Write the labels up
pos4 = c(1,3,13,16)                     # Write the labels right

plot(c.hclust.out,k=5,links=FALSE,xlab="Easting",ylab="Northing",cex=1.5,
   main="Space-constrained clustering map of Chesapeake fauna")
text(coo.25[pos1,], labels=rownames(xy.25[pos1,]), cex=1, col= "blue", pos=1) #pos=bas
text(coo.25[pos2,], labels=rownames(xy.25[pos2,]), cex=1, col= "blue", pos=2) # pos=left
text(coo.25[pos3,], labels=rownames(xy.25[pos3,]), cex=1, col= "blue", pos=3) # pos=up
text(coo.25[pos4,], labels=rownames(xy.25[pos4,]), cex=1, col= "blue", pos=4)#pos=right
```

Run the plotting code (i.e. the 5 lines of code above) for other values of k to obtain maps with different numbers of groups, e.g. from k=2 to k=5.

A map showing the sites *and* the link edges can be produced as follows (map not shown in the file "Figures for Practical exercises")

```
plot(c.hclust.out,k=5,links=TRUE,xlab="Easting",ylab="Northing",cex=1.5
   main="Space-constrained clustering map of Chesapeake fauna")
```

Add the site identifiers to this map using the 4 lines above (the code lines beginning with "text").

The list of sites in each constrained group identified by constr.hclust can be obtained as follows:

```
cutree(c.hclust.out, k=5)
```

The following code produces a numeric vector describing the partition of sites into k groups:

```
k5.groups = as.numeric(cutree(c.hclust.out, k=5))
```

A dendrogram of the space-constrained hierarchical clustering results can be produced as follows (Figure 9b). The dendrogram contains reversals due to by imposition of the constraint during the hierarchical agglomeration procedure. Reversals do not impair the interpretation of the space-constrained groups produced by the method.

```
stats:::plot.hclust(c.hclust.out, hang=-1)
```

6.1.4. MRT interpretation of the classification groups using environmental variables

A classification can be interpreted by explanatory variables using a regression tree or an RDA.
Here we will use the function varpart.R (method MRT), already used in section 4.2.1 and 4.2.5, to identify possible explanatory variables for the classification into 3 groups. Here varpart.R is **not** used as a constrained clustering method, contrary to section 4.2.1 where it was.
About MRT: refer to the course teaching document "Multivariate regression tree analysis.pdf"

For the sake of the demonstration, we will use as explanatory variables the sediment data collected in year 2003, Spring survey. This will be assumed to be representative of the spatial variation of the sediment throughout the 25 brackish sites of Chesapeake Bay.

Classification into 3 groups produced by constrained.hclust.R

```
k3.groups = as.numeric(cutree(c.hclust.out, k=3))
#  [1] 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 1 3 1 3 2 2 2 3 3 3
```

For MRT analysis, we have to turn this numeric vector into a factor in a data.frame
```
k3.gr.df = data.frame(as.factor(k3.groups))
colnames(k3.gr.df) = "k3.gr"
```

head(`k3.gr.df`) ; tail(`k3.gr.df`)

# Use the environmental data of year 2003, spring
tmp = sampling.25
sediment.2003 = sediment.25[tmp[,5]=="2003" & tmp[,6]=="Spring",]

The variables SAND and SILTCLAY are perfectly collinear, with cor(SAND,SILTCLAY) = -1.00
Remove the variable SAND from the analysis.
Note: function mvpart would have eliminated SAND automatically, had we not done it here.

# Compute the regression tree analysis
res.part = mvpart(data.matrix(`k3.gr.df`) ~ . , data= sediment.2003[,-2], xv="pick", xvmult=100)

The only significant split is into 2 groups, according to the cross-validation results. That split separates the group of data rows (1,2,16,18) from the other 21 data rows.
That group corresponds to sites #1, #6, #15, #51 found in the centre of Chesapeake Bay. They are represented by red dots in Figure 9a.
These 4 sites had less than 3.8% of SILTCLAY in the sediment, whereas the other sites had larger values. Sediment deposition is controlled by hydrodynamics, which seems to differ in these 4 sites compared to the other 21 sites.

6.2. TEST SPACE-TIME INTERACTION FOR COMMUNITY DATA WITHOUT REPLICATION

Exercise –
A test of space-time interaction in the absence of replication was carried out in section 3.2 for 25 brackish sites, sampling year 1999, spring and fall surveys.
Can you repeat this analysis now for the same 25 sites across the 13 sampling years, for the spring surveys only [or the fall surveys only, as you wish], using function stimodels.R of {adespatial}?
How should you code the 13 years? Check argument S in the documentation file of stimodels.

```
# 6.3. LCBD ANALYSIS ON 25 BRACKISH SITES, SPRING AND FALL SEPARATELY

# In this section, we will use the method of beta diversity analysis described by Legendre & De
Cáceres (1993) and the data files for the 25 brackish sites prepared in section 2.2.

# Objective of this analysis – Produce two space-time LCBD maps, one for Spring and one for Fall

# Fauna: remove columns with colSums=0. The other species were freshwater species.
fauna.25 = fauna.25[ , colSums(fauna.25)!=0]
dim(fauna.25)                          # [650 155]; 155 species were identified at the brackish sites

# Create a simple factor for seasons for the 25 brackish sites, 13 years and 2 seasons; length = 650

season25 <- sampling.25$SEASON

# 6.3.1. Compute LCDB indices

# Compute LCBD indices for Spring and Fall separately, 325 data rows in each analysis

beta.out.25.spring <- beta.div(fauna.25[season25=="Spring",], method="hellinger", nperm=999)
beta.out.25.fall <- beta.div(fauna.25[season25=="Fall",], method="hellinger", nperm=999)

# Code for computing LCBD indices for all 650 data rows, 650 Space-Time points
# beta.out.25 <- beta.div(fauna.25, method="hellinger", nperm=0)

# The LCBD indices are available in the $LCBD vector, the permutational p-values in $p.LCBD

signif.25.spring <- which(beta.out.25.spring$p.LCBD <= 0.05)
signif.25.fall <- which(beta.out.25.fall$p.LCBD <= 0.05)
length(signif.25.spring)       # 39 LCBD indices are significant (alpha = 0.05) over 325 S-T points
length(signif.25.fall)         # 48 LCBD indices are significant (alpha = 0.05) over 325 S-T points
# Permutation tests: the number of significant LCBD indices may vary among computer runs

# Plot space-time maps of LCBD, spring and fall data separately; Figure 10a
# Significant LCBD values at the 0.05 level are plotted with a black rim

par(mfrow=c(1,2))
seq.X.25 <- rep(1996:2008, 25)
seq.Y.25 <- rep(1:25, each=13)

plot(seq.X.25, seq.Y.25, asp=1, type="n", ylab="Sites", xlab="Years", main="Space-time map,
   LCBD, spring", ylim=c(1,25), xlim=c(1996,2008), yaxt="n", cex.axis=0.8)
points(seq.X.25, seq.Y.25, pch=21, col="white", bg="steelblue2",
   cex=30*sqrt(beta.out.25.spring$LCBD))
points(seq.X.25[signif.25.spring], seq.Y.25[signif.25.spring], pch=21, col="black", bg="steelblue2",
   cex=30*sqrt(beta.out.25.spring$LCBD[signif.25.spring]))      # Significant LCBD values, spring
axis(side=2, 1:25, labels=rownames(xy.25), las=1, cex.axis=0.8)
```

```
plot(seq.X.25, seq.Y.25, asp=1, type="n", ylab="Sites", xlab="Years", main="Space-time map,
    LCBD, fall", ylim=c(1,25), xlim=c(1996,2008), yaxt="n", cex.axis=0.8)
points(seq.X.25, seq.Y.25, pch=21, col="white", bg="steelblue2",
    cex=30*sqrt(beta.out.25.fall$LCBD))
points(seq.X.25[signif.25.fall], seq.Y.25[signif.25.fall], pch=21, col="black", bg="steelblue2",
    cex=30*sqrt(beta.out.25.fall$LCBD[signif.25.fall]))          # Significant LCBD values, fall
axis(side=2, 1:25, labels=rownames(xy.25), las=1, cex.axis=0.8)
par(mfrow=c(1,1))


# 6.3.2. Compute species richness (written to vector "rich")
rich <- apply(decostand(fauna.25, method="pa"), 1, sum)

# Plot space-time maps of richness, spring and fall data; Appendix A3, Figure 10b

rich.spring <- rich[seq(from=1, to=649, by=2)]
rich.fall <- rich[seq(from=2, to=650, by=2)]

par(mfrow=c(1,2))
seq.X.25 <- rep(1996:2008, 25)
seq.Y.25 <- rep(1:25, each=13)

plot(seq.X.25, seq.Y.25, asp=1, type="n", ylab="Sites", xlab="Years", main="Space-time map,
Richness, spring", ylim=c(1,25), xlim=c(1996,2008), yaxt="n", cex.axis=0.8)
points(seq.X.25, seq.Y.25, pch=21, col="white", bg="steelblue2", cex=0.5*sqrt(rich.spring))
axis(side=2, 1:25, labels=rownames(xy.25), las=1, cex.axis=0.8)

plot(seq.X.25, seq.Y.25, asp=1, type="n", ylab="Sites", xlab="Years", main="Space-time map,
Richness, fall", ylim=c(1,25), xlim=c(1996,2008), yaxt="n", cex.axis=0.8)
points(seq.X.25, seq.Y.25, pch=21, col="white", bg="steelblue2", cex=0.5*sqrt(rich.fall))
axis(side=2, 1:25, labels=rownames(xy.25), las=1, cex.axis=0.8)
par(mfrow=c(1,1))


# 6.3.3. Relate LCBD indices to environmental variables

# Exercise –
# LCBD form a new data vector, which can be analysed by methods for univariate data, like any
other variable. The spatial and temporal variation in LCBD indices can be interpreted by linear
regression or regression tree analysis against environmental variables. Can you do it?
```

6.4. TBI ANALYSIS AT THE 25 SITES: COMPARE THE SPRING AND FALL SURVEYS

The method of temporal beta diversity analysis (TBI, Legendre 2019) only allows, in its present form, the comparison *of two different* surveys conducted at the same set of sites (several sites).

This section carries out a TBI analysis for the Spring and Fall surveys of 2003 at the 25 brackish sites.

6.4.1. Assemble separate faunal data frames for the spring and fall of 2003

```
tmp = sampling.25
fauna.2003.S = fauna.25[tmp[,5]=="2003" & tmp[,6]=="Spring", ]        # dim = 25 155
fauna.2003.F = fauna.25[tmp[,5]=="2003" & tmp[,6]=="Fall", ]          # dim = 25 155
```

Note: file fauna.25 has been purged from the absent species in section 6.3. The absent species would be of no use for TBI analysis, so we can use file fauna.25 here. Make sure, however, that the files for time 1 (T1) and time 2 (T2) contain the same list of species in the same order.

Fauna: DO NOT remove columns with colSums=0 from the separate Spring and Fall matrices
For TBI analysis, the two matrices must contain the same list of species in the same order

Do NOT pre-transform the faunal data. A dissimilarity function, chosen by the user, will be computed within the TBI function.

6.4.2. TBI analysis of the 2003 data, Spring and Fall

Comparing the T1 = Spring and T2 = Fall surveys at the 25 sites
Run enough permutations that some p-values remain significant after correction for multiple tests
Extensive calculations for tests of significance with nperm=9999; waiting time about 1 min

```
( res.TBI = TBI(fauna.2003.S, fauna.2003.F, method="%difference",
    nperm=9999) )
```

Examine the file of results –

1. $t.test_B.C – Is the overall paired *t*-test of differences between B (losses) and C (gains) significant? Examine columns “p.param” and “p<=0.05”.
If so, examine element $BCD.summary of the output file. What is the dominant sign of the changes in community composition at individual sites (column “Change”)?

2. $p.adj – Are there sites with significant TBI indices after correction for multiple tests? Look for p.adj values ≤ 0.05.
For these sites, what is the sign in column “Change” in community composition?

3. Produce a B-C plot showing the detailed differences with function plot.TBI.R (Figure 11a)

```
s.names = as.numeric(rownames(xy.25))
plot(res.TBI, s.names=s.names, xlim=c(0.0,1.1), main = "B-C plot, 25 Chesapeake
    Bay sites")
```

Interpretation of the B-C plot from columns B and C in element $BCD.mat of output file; B+C=D

• Species losses (statistic B) are in abscissa, species gains (statistic C) in ordinate.

• The symbols are drawn to sizes representing the values of the D = (B+C) statistics where D is the selected dissimilarity measure (here: the *percentage difference* dissimilarity, aka Bray-Curtis index). Square for sites with overall species gains, circles for species losses, shown by the "Change" column in $BCD.mat. Going up the green line, symbol sizes increase, meaning that D values increase.

• The green line is drawn at 45º from the origin of the plot. It separates the upper zone where gains dominate from the lower zone where losses dominate.

• The red line is parallel to the green line. It passes through the centre of mass of the points. In the present graph, the red line lower than the green indicates that the changes in species composition are dominated by species losses from the Spring to the Fall surveys. A red line above the green line would indicate the opposite, i.e. gains would dominate losses over the sites under study.

4. Draw separate B-C plots for the 3 constrained clustering groups of sites (k=3) from section 6.1

The B-C plots are drawn from segment $BCD.mat of the output file. We will edit the output file to create three separate files with the correct list of sites in the respective $BCD.mat elements.

One could also create a script that will carry out the file edition and the production of several B-C plots on a single page.

```
# Membership of the k=3 clusters obtained by constr.hclust, section 6.1.3
gr1 = c(3:11,20:22)               # Northern cluster, n=12
gr2 = c(1,2,16,18)                # Central cluster, n=4
gr3 = c(12:15,17,19,23:25)        # Southern cluster, n=9

# Choice of colours to plot the sites belonging to the three clusters
col.vec = c("gold1", "cadetblue2", "coral2")

# Select data rows of "res.TBI$BCD.mat" corresponding to the northern cluster (gr1)
res.TBI.gr1 = res.TBI
res.TBI.gr1$BCD.mat = res.TBI.gr1$BCD.mat[gr1,]

# Select data rows of "res.TBI$BCD.mat" corresponding to the central cluster (gr2)
res.TBI.gr2 = res.TBI
res.TBI.gr2$BCD.mat = res.TBI.gr2$BCD.mat[gr2,]

# Select data rows of "res.TBI$BCD.mat" corresponding to the southern cluster (gr3)
res.TBI.gr3 = res.TBI
res.TBI.gr3$BCD.mat = res.TBI.gr3$BCD.mat[gr3,]
```

Print the B-C plots corresponding to the three site clusters (Figure 11b)

```
par(mfrow=c(2,2))
plot(res.TBI.gr1, s.names=s.names[gr1], xlim=c(0,1.1), ylim=c(0,0.7),  col.bg =
     col.vec[1], main="B-C plot, northern cluster")
#
plot(res.TBI.gr2, s.names=s.names[gr2], xlim=c(0,1.1), ylim=c(0,0.7), col.bg =
     col.vec[2], main="B-C plot, central cluster")
#
plot(res.TBI.gr3, s.names=s.names[gr3], xlim=c(0,1.1), ylim=c(0,0.7), col.bg =
     col.vec[3], main="B-C plot, southern cluster")
```

The figure shows that losses of abundances-per-species have occurred between the Spring and Fall surveys in the northern and southern clusters. This phenomenon does not seem to have taken place in the central cluster, insofar as a firm conclusion can be drawn from the observation of 4 sites only.

The same TBI analysis, repeated with the Sørensen coefficient (method="sorensen") which only takes species presence-absence into account, led to a similar conclusion: losses of species have occurred between the Spring and Fall surveys in the northern and southern clusters. The importance of the losses is, however, less marked with numbers of species data than with abundances-per-species (Figure 11b).

This example illustrates the fact that it is often interesting, for ecological interpretation, to carry out TBI analysis using both species abundance and presence-absence data.

Separate TBI analyses can be conducted for separate species groups found in ecological communities. Divide the species into groups according to size classes (e.g. for trees) or traits and compute separate TBI analyses for the different groups. For example, Brice et al. (2019) divided the tree community found in > 6000 forest plots into boreal, pioneer and temperate species, and computed TBI analysis for each group separately. They examined the species gains and losses in these three species groups from South to North along a latitude gradient.

REFERENCES

Borcard, D., F. Gillet & P. Legendre. 2018. Numerical ecology with R, 2nd edition. Use R! series, Springer International Publishing AG. xv + 435 pp.

Brice, M.-H., K. Cazelles, P. Legendre & M.-J. Fortin. 2019. Disturbances amplify tree community responses to climate change in the temperate-boreal ecotone. *Global Ecology & Biogeography* 28:1668–1681.

De'ath, G. 2002. Multivariate regression trees: a new technique for modeling speciesenvironment relationships. *Ecology* 83: 1105-1117.

Legendre, P. 2019. A temporal beta-diversity index to identify sites that have changed in exceptional ways in space-time surveys. *Ecology and Evolution* 9: 3500–3514. https://doi.org/10.1002/ece3.4984.

Legendre, P. & M. De Cáceres. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters* 16: 951-963. https://dx.doi.org/10.1111/ele.12141.

Legendre, P. & O. Gauthier. 2014. Statistical methods for temporal and space-time analysis of community composition data. *Proceedings of the Royal Society B - Biological Sciences* 281: 20132728. PDF available on http://adn.biol.umontreal.ca/~numericalecology/Reprints/

Legendre, P. & L. Legendre. 2012. *Numerical ecology, 3rd English edition.* Elsevier Science BV, Amsterdam. xvi + 990 pp.

R Core Team. 2021. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58: 236-244.

Wickham, H. 2007. Reshaping data with the reshape package. *Journal of Statistical Software* 21: 1–20.