Appendix to:

Legendre P, Gauthier O. 2014 Statistical methods for temporal and space-time analysis of community composition data. *Proc. R. Soc. B* **281**, xxx–xxx.

Appendix S1

Drift can generate a trend in community time series: a simulation study

Consider a quantitative data time series **y**. The series may have identifiable structure along the time axis caused by four different types of processes, which can be analysed separately:

• Induced temporal dependence – The response time series **y** may be related to explanatory variables (e.g. environmental) **X** that have a time structure of their own. The functional relationship between **y** and **X** may induce a transfer of that structure, which is then found in the time series of interest. Matrix **X** may include environmental (physical, chemical, geological, …) variables, or biotic variables not included in the response community under study (for example top-down influence of predators or bottom-up influence of other communities).

• In ecological data, community dynamics, including complex interactions among species (through predation, parasitism, competition, amensalism, mutualism, commensalism) or shifts in food resources, may generate temporal patterns in the community data series **y** of interest.

• Neutral processes, i.e. processes that are not functionally related to environmental conditions, can also generate spatial structures in the community composition data of interest. They include ecological drift (variation in species demography caused by random reproduction and survival of individuals due to competition, predator-prey interactions, etc.), random dispersal (migrations in animals, propagule dispersion in plants), and other interactions among species within the community of interest that are not under environmental control. The effect of neutral drift, which is cumulative along the series, is described in more detail in the present Appendix.

• Unexplained variation is caused by processes generating what statisticians call *random noise*. This is the sum of the non-explained variation of the series. For population or community data, the processes involved include the fact that the population or community targeted by sampling may have been moving around the area where sampling took place, the error of the sampling procedure, error in the lab measurements or counts, etc. That type of variation is not cumulative along the series.

We will now show that ecological [or neutral] drift along autocorrelated time series can, in many cases, generate trends. The MEM and AEM methods of analysis generate eigenfunctions that can model trends resulting from directional processes. That is why MEM and AEM analyses are appropriate to model the results of temporal processes that contain trends either induced by forcing environmental variables or biotic processes, or due to ecological drift. The MEM method can model trends although it was not originally designed for that purpose.

A simple simulation study was conducted to illustrate the fact that drift alone can generate trends in temporal data. Series of random normal deviates, $N(0,1)$, were generated using the

rnorm() function of R and summed cumulatively along the series to simulate neutral ecological drift. Several (100 or 1000) such series were generated with various numbers of observations, $n = \{25, 50, 75, 100\}$.

Most of the generated series (75 to 86%) had a significant slope with respect to time at significance level 0.05, hence a significant trend. The proportion increased slightly as the length of the series increased from 25 to 100; this was expected because tests of significance (here, the test of the regression coefficient) have higher power when n is larger. Any one simulated series could, however, have a positive or negative slope. About 50% of the series in each set of simulations had a positive slope, the other half had a negative slope. A set of 100 data series with length $n = 100$ is shown in Figure S1.1. Clark and McLachlan (2003, Fig. 1b) showed simulated species data displaying that behaviour. These authors also showed how neutral drift can generate increasing variance among sites (increasing beta diversity) as time goes, with or without competition among species.

These simple simulations show that the process of neutral drift is likely to cause a trend in data series. MEM and AEM modelling are appropriate to analyse data of that type.

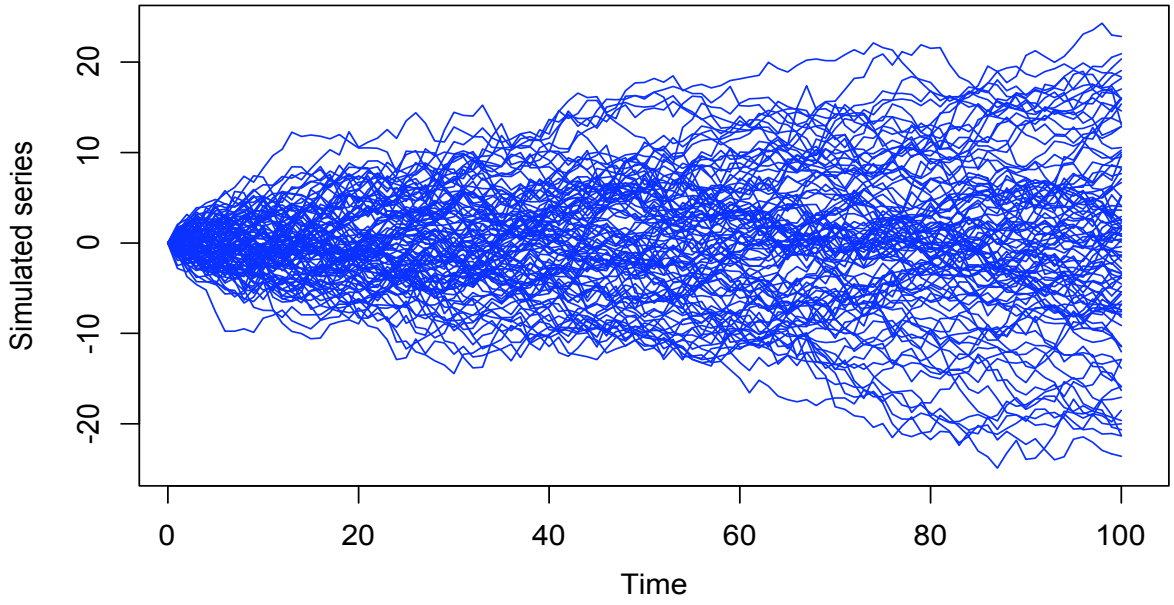

Figure S1.1. Example of simulated data series. There are 100 series plotted in the figure. Each one is the cumulative sum of 100 random normal deviates after the initial point at value 0 (left).

The following R function was used for these simulations.

==========

```
generate.drift <- function(n, p, plot.series=TRUE, print.res=TRUE, col="blue")
# This function generates series of cumulated random normal deviates,
# plots them, and analyses them by regression for linear trends.
#
```

```
# Arguments --
# n : length of each simulated series after the starting point of 0.
# p : number of series generated.
# plot.series=TRUE : produce a graph of the simulated series.
# print.res=TRUE : print regression results with stars for significance.
# col : colour of the lines in the plot. Default = "blue".
#
# Value (elements in the output list) --
# mat : matrix of simulated data. Beware: the first row contains zeros.
# lm.res : matrix of regression statistics.
# n.signif : proportion of significant regressions at 0.05 level.
# n.pos : proportion of regression lines with positive slopes.
#
# License: GPL-2
# Author:: Pierre Legendre, July 2013
#
{
a <- system.time({             # How much time for this set of simulations?
abscissa <- 0:n
mat <- rep(NA,n)
for(j in 1:p) mat <- cbind(mat, cumsum(rnorm(n)))
mat <- mat[,-1]
mat <- rbind(rep(0,p),mat)
colnames(mat) <- paste("Var",1:p,sep=".")
#
# Plot the simulated "cumsum" variables
if(plot.series) {
     range.mat <- range(mat)
     plot(abscissa, ylim=range.mat, type="n",xlab="Time",ylab="Simulated
series")
     for(j in 1:p) lines(abscissa,mat[,j], col=col)
     }
#
# Compute linear regressions for all generated "cumsum" variables
lm.res <- matrix(NA,p,4)
colnames(lm.res) <- c("reg.coeff","R.square","F","P.value")
rownames(lm.res) <- paste("Cumsum",1:p,sep=".")
for(j in 1:p) {
     lm.out <- lm(mat[,j] ~ abscissa)
     lm.res[j,1] <- summary(lm.out)$coefficients[2,1]
     lm.res[j,2] <- summary(lm.out)$r.squared
     F.stat <- summary(lm.out)$fstatistic
     lm.res[j,3] <- F.stat[1]
     lm.res[j,4] <- pf(F.stat[1],F.stat[2], F.stat[3], lower.tail=FALSE)
     }
n.signif <- length(which(lm.res[,4] <= 0.05))
n.pos <- length(which(lm.res[,1] >= 0))
#
})
a[3] <- sprintf("%2f",a[3])
cat("Computation time =",a[3]," sec",'\n')
# Output the results
if(print.res)
     printCoefmat(lm.res, P.values=TRUE, signif.stars=TRUE, has.Pvalue=TRUE)
list(mat=mat, lm.res=lm.res, n.signif=n.signif/p, n.pos=n.pos/p)
}
```

==========

MEM AND AEM MODELLING

Three of the simulated data series were analysed through AEM and dbMEM modelling. Using the function *generate.drift()*, 100 data series were generated with $n = 50$. Series 22 had the highest F-statistic (strongest trend) in that set of simulations (slope of regression against time = 0.3499, $F = 597.496$), series 25 had the median F-statistic for that set of simulations (slope = 0.1385, $F = 45.281$), and series 53 had the smallest F-statistic (slope = –0.0007, $F = 0.002$). They represent an appropriate diversity of situations to illustrate the calculation of the directional and nondirectional fractions of variability by temporal eigenfunction modelling.

In this exercise, the temporal models were computed using the 24 dbMEM and 24 AEM eigenfunctions with positive Moran's I indices because we were interested in modelling positive temporal autocorrelation in the data, as suggested by the data generation process. No selection of explanatory dbMEM or AEM variables was carried out. RV coefficients between the groups of MEM and AEM eigenfunctions were 0.9862 for the 24 functions modelling positive temporal correlation and 0.9867 for the 25 functions modelling negative correlation, indicating that the MEM and AEM sets of eigenvectors are highly similar and should have similar explanatory powers. The RV coefficient is a multivariate generalization of the Pearson correlation coefficients to compare two data sets (Escoufier 1973, Robert & Escoufier 1976).

The results (Table S1.1) show that MEM and AEM analyses successfully model the simulated data, which only contain drift, to the exclusion of any dependence induced by environmental variables. In Figure S1.2, the red lines (MEM models fitted to the original data) are always very close to the full black lines (original data), whereas the blue lines (MEM models fitted to the detrended data) are always very close to the dashed lines (detrended data). The AEM models fitted to the original data (not shown) are undistinguishable from the MEM models fitted to the original data (red lines). The two modelling methods are equally capable of modelling undetrended autocorrelated data series.

Although the trend in series 53 was not significant, we detrended the data because the series was produced through drift, which is known to generate trends in data. When analysing real data series in which the presence of a trend is not known from theoretical considerations, one has to rely on a test of significance of the trend to help decide whether or not to detrend the data prior to MEM analysis.

The interest of using temporal eigenfunctions for analysis lies in the fact that one can compute the proportion of variance (R^2) representing the directional and nondirectional components of variation. In these calculations, R^2 is used instead of R^2_{adj} because the explanatory matrix contains MEM or AEM, which are fixed factors. In Table S1.1,

• Line 4 shows the R^2 of the linear trend, obtained by regressing the data onto the vector of times, and line 5 is the R^2 of the residuals.

• Line 6 is the R^2 of the MEM model computed on the undetrended data and line 7 is the R^2 of the AEM model, also computed on the undetrended data.

• Line 8 shows the R^2 of the MEM model computed on the detrended (i.e. residual) data; that R^2 represents the proportion *of the detrended data* explained by the MEM model.

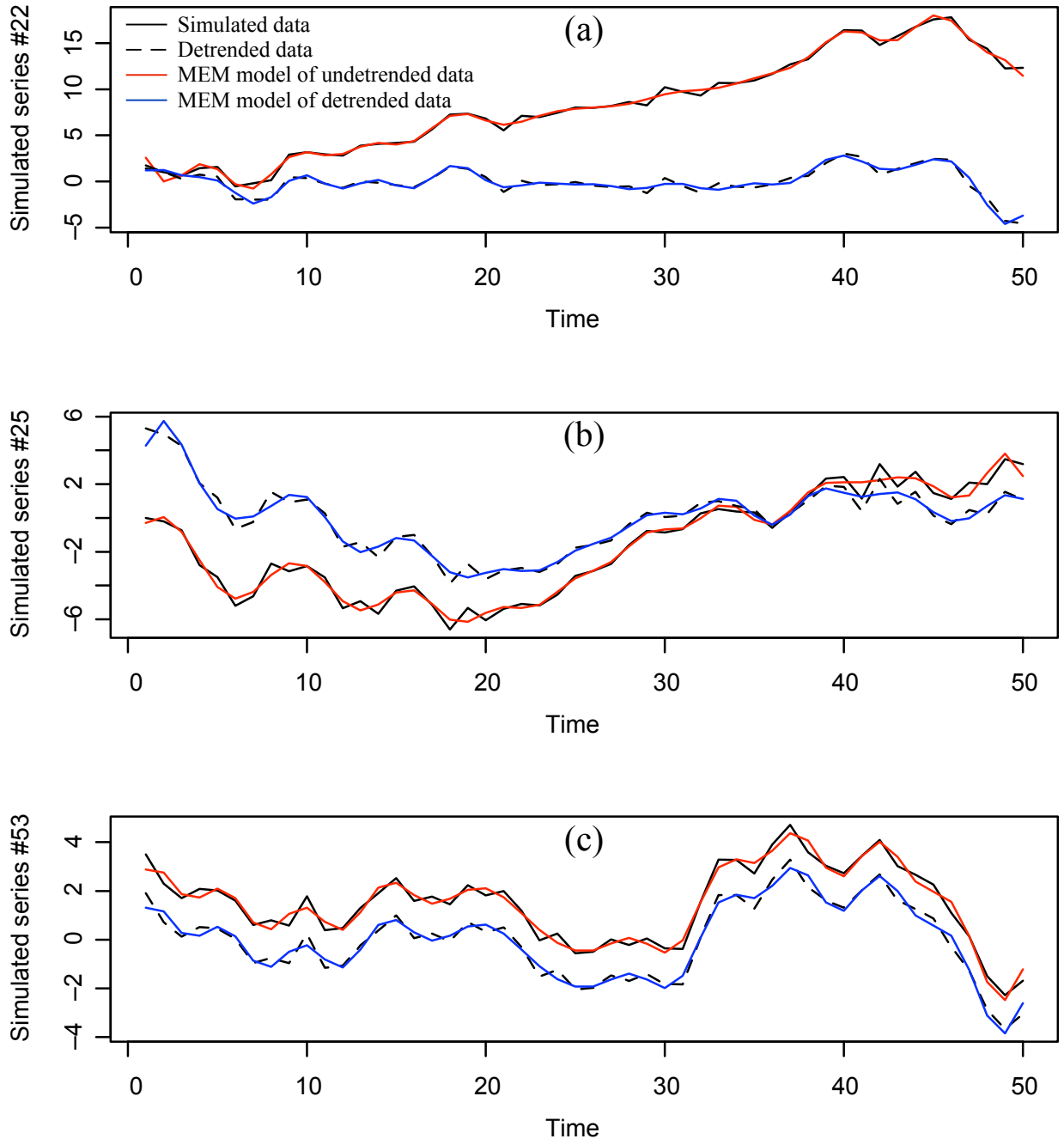

Figure S1.2. Results of MEM analysis of the three temporal data series. (a) Series 22 with strong linear trend, (b) series 25 with an intermediate trend, and (c) series 53 with a weak linear trend. The full black line represents the original simulated data whereas the dashed black line depicts the detrended data. The colour lines are the fitted values of the MEM models of the original (red) and detrended data (blue).

• Line 9 is the nondirectional R^2, computed as the product the R^2 on line 8 and the R^2 of the residuals on line 5.

• The directional R^2 shown on lines 10 and 11 are computed by subtracting the nondirectional R^2 (line 9) from the total R^2 of the MEM and AEM models (lines 6 and 7, respectively), computed on the undetrended data. These two estimates are very similar.

Table S1.1. Calculation of the nondirectional and directional components of variation (R^2) for the three selected time series.

	Series 22	Series 25	Series 53
1. Variance of original data	28.5174	8.9609	2.4611
2. Variance of trend	26.2618	4.8381	0.0040
3. Variance of residuals	2.2556	4.1228	2.4571
4. R^2 of trend	0.9209	0.5399	0.0016
5. R^2 of residuals	0.0791	0.4601	0.9984
6. R^2 of MEM model, undetrended data	0.9935	0.9805	0.9653
7. R^2 of AEM model	0.9962	0.9838	0.9739
8. R^2 of MEM model of the detrended data	0.9294	0.9538	0.9660
9. Nondirectional R^2 = line 8 × line 5	0.0735	0.4388	0.9644
10. Directional R^2 based on MEM model (undetrended) = line 6 – line 9	0.9200	0.5416	0.0008
11. Directional R^2 based on the AEM model = line 7 – line 9	0.9227	0.5450	0.0095

Examine the nondirectional and directional components of R^2 for the three selected variables at the bottom of Table S1.1. Series 22 had a strong temporal gradient and has a high directional R^2 value (line 4). For series 25, about half of its variation is nondirectional (line 5) and the other half is directional (line 4). Series 53, which had a weak trend, has nearly all its variation classified as nondirectional. As shown in these examples, MEM and AEM modelling can both be used to analyse the undetrended data and estimate the total component of variation explained by temporal eigenfunctions (lines 6 and 7). Following that, one can estimate the directional components of variation (lines 10 and 11) by subtracting the nondirectional components (line 9).

Reference

Clark JS, McLachlan JS. 2003 Stability of forest biodiversity. *Nature* **423**, 635–638.

Escoufier Y. 1973 Le traitement des variables vectorielles. *Biometrics* **29**, 751–760.

Robert P, Escoufier Y. 1976 A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Appl. Stat. – J. Roy. St. C* **25**, 257–265.

Appendix to:

Legendre P, Gauthier O. 2014 Statistical methods for temporal and space-time analysis of community composition data. *Proc. R. Soc. B* **281**, xxx–xxx.

Appendix S2

Temporal eigenfunction methods – Practicals in R

This Appendix gives the calculation details in the R statistical language (R Core Team 2013) for the case study developed in section 8 of the paper.

Table of Contents

0. REQUIRED R PACKAGES

The following packages, available on CRAN, will be used: 'ade4', 'mvpart', 'vegan'.

Download also the 'packfor', 'PCNM' and 'AEM' packages from the address http://r-forge.r-project.org/R/?group_id=195. Install them on your computer. Functions ordistep and ordiR2step of 'vegan' can be used instead of forward.sel of 'packfor' for selection of explanatory variables in the analysis of multivariate response data.

1. 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.

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>. It is available in Appendix S5.
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 your R console is set to the working
directory corresponding to the folder containing the file. Check this by typing getwd().

```
ls()
```

1.1. DESCRIPTION OF THE DATA FILES

Please 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 – Sampling years, 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 μmho/cm, US equivalent to μS/cm.
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 then 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.

# 1.2. 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)

# 1.3. PLOT A NICER MAP OF THE SITES WITH CHESAPEAKE BAY BACKGROUND
#       using RgoogleMaps (Appendix S3, Figure S3.2)

# install.packages("RgoogleMaps", dependencies=TRUE)
require(RgoogleMaps)

# Get the map background from the Google server.
# You will need to be connected to the Internet to fetch the map background from a
# Google server.  Choices for the type of background are:
# maptype = c("roadmap", "mobile", "satellite", "terrain", "hybrid",
#                   "mapmaker-roadmap", "mapmaker-hybrid")
MapBackground(lat=xy$LATITUDE, lon=xy$LONGITUDE, destfile= "bckg", maptype =
"terrain")

# Load the map to an R object
my.map <- ReadMapTile(destfile="bckg")

# Add points at the site coordinates
PlotOnStaticMap(my.map, lat=xy$LATITUDE, lon=xy$LONGITUDE, cex=1.2, col="red",
pch=19)

# Add site labels at positions offset from the site points
TextOnStaticMap(my.map, xy$LATITUDE+0.0005*abs(xy$LATITUDE),
xy$LONGITUDE+0.0005*abs(xy$LONGITUDE), labels=rownames(xy), add=T, col="black")

# Overlapping labels may be touched-up using a graphics editor such as Inkscape (free).

# Close the graphic window before proceeding, otherwise the next plots will be cropped.
# Either close the window manually, or type: dev.off() in the console.

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

2. LOAD THE NECESSARY PACKAGES, PREPARE THE DATA MATRICES

2.1. LOAD THE R PACKAGES NECESSARY FOR ANALYSIS

```
require(vegan)
require(ade4)
require(PCNM)
require(packfor)
require(AEM)
```

2.2. LOAD USEFUL FUNCTIONS

```
# Load file "R_functions_for_Practicals.txt" (Appendix S5), which contains three R functions
# written for these Practicals, using the "Source R Code..." menu for Windows clients
# or the "Source File..." menu for MacOS X clients.
```

2.3. IS THERE A LINEAR TREND IN TEMPERATURE AT ANY OF THE SITES?

```
# For each site, regress water temperature data against sampling dates.
# Function temperature.trend() is in file "R_functions_for_Practicals.txt" loaded in section 2.2.

(res <- temperature.trend(waterquality, rownames(xy), sampling))

# Are any of the regression coefficients significant?

# Because temperature is an important environmental variable in determining the structure of
# invertebrate communities, the presence of a temporal trend in temperature (over the sampling
# years) may suggest the presence of an induced temporal trend in the community data.

# The presence of such a trend in the faunal data can be checked by computing an RDA of the
# community data over time. This will be done in section 3.2 of these Practicals for the site 40
# data prior to MEM modelling.

# Environmental variables other than temperature may be driving community structure in
# ecological time series. Drift may also generate trends in time series (Appendix S1).

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

3. TIME SERIES ANALYSIS OF A SINGLE SITE: THE SITE 40 DATA

3.1. SELECT SITE 40 AMONG THE 27 SAMPLING SITES

```
# Site 40 is located in the upper (brackish) course of the Potomac River, which constitutes the
# border between Maryland and Virginia in the USA. Mean salinity at site 40 over the 13 years
# period was 2.3 PSU.

curr.site <- 40
sampling.40 <- sampling[sampling$STATION == curr.site, ]
sediment.40 <- sediment[sampling$STATION == curr.site, ]
waterquality.40 <- waterquality[sampling$STATION == curr.site, ]
fauna.40 <- fauna[sampling$STATION == curr.site, ]
fauna.40 <- fauna.40[ , colSums(fauna.40)!=0]       # (26 sampling units x 36 taxa)

# Quick analysis: identify the temporal structure of the sampling design
dates.40 <- sampling.40$SAMPLE_DATE
plot(dates.40, rep(1,26), yaxt="n", ylab="")
```

3.2. TEMPORAL EIGENFUNCTION ANALYSIS OF COMMUNITY DATA FROM SITE 40

```
# Hellinger transformation of the faunal data prior to analysis by RDA
# Transformation: square root of the relative abundances by rows
# See reference to Legendre & Gallagher (2001) in ?decostand
fauna.hel.40 <- decostand(fauna.40, method="hellinger")

# Is there a linear temporal trend in the response data?
fauna.trend <- rda(fauna.hel.40, sampling.40$SAMPLE_DATE)
anova(fauna.trend, step=10000, perm.max=10000)          # See note[1]
RsquareAdj(fauna.trend)
# Examine the R-square and the p-value to decide if there is a significant temporal trend.
# Detrend the response data (i.e. compute residuals) only if the trend is [highly] significant.
# The MEM and AEM methods are equally suitable to analyze the temporal structure of the data.
```

 # 3.2.1. dbMEM analysis

```
# Construct the dbMEM eigenfunctions. Generate all dbMEM eigenfunctions
# As used here, function dist() computes the number of days between sampling occasions
time.mem.40 <- PCNM(dist(dates.40), dbMEM=TRUE, moran=TRUE, all=TRUE)
summary(time.mem.40)
dim(time.mem.40$vectors)

# Which dbMEM model positive temporal correlation?
time.mem.40$Moran_I
```

[1] There are many functions called anova() in R. They do not compute an Analyse Of Variance. Instead, they compute different tests of significance and produce anova-like tables. R automatically selects the anova function that corresponds to the class of the object. The class associated with object "fauna.trend" is found with class(fauna.trend) [result: "rda" "cca"]. For this object, function anova.cca(), which performs a permutation test of the RDA statistic, was called by R and used for the analysis.

```
# Compute the redundancy analysis (RDA) of the fauna by the dbMEM eigenvectors modelling
# positive temporal correlation
time.mem.40.pos <- as.data.frame(time.mem.40$vectors[,time.mem.40$Moran_I$Positive])
fauna.mem.40.pos <- rda(fauna.hel.40 ~ ., time.mem.40.pos)
anova(fauna.mem.40.pos, step=10000, perm.max=10000)
RsquareAdj(fauna.mem.40.pos)
anova(fauna.mem.40.pos, by="axis")
# Examine the R-square, the adjusted R-square and the p-value.
# The model based upon the 12 MEM modelling positive temporal correlation is significant.
# It produces two significant canonical axes.

# Compute the redundancy analysis (RDA) of the fauna by the dbMEM eigenvectors modelling
# negative temporal correlation
time.mem.40.neg <- as.data.frame(time.mem.40$vectors[,!time.mem.40$Moran_I$Positive])
fauna.mem.40.neg <- rda(fauna.hel.40 ~ ., time.mem.40.neg)
anova(fauna.mem.40.neg, step=10000, perm.max=10000)
RsquareAdj(fauna.mem.40.neg)
anova(fauna.mem.40.neg, by="axis")
# Examine the R-square, the adjusted R-square and the p-value
# The model containing the 13 MEM modelling negative temporal correlation is not globally
# significant. It produces a significant canonical axis, however, and that axis is worth looking at.

# The canonical axes (MEM models) are all orthogonal to one another.
# Property of orthogonal vectors: their cross-product is 0. Example:
tmp <- as.matrix(summary(fauna.mem.40.pos)$constraints)
round(t(tmp) %*% tmp, 4)

# Plot the RDA axes values of the three canonical models (Appendix S3, Figure S3.3)
par(mfrow=c(3,1))
# The positive ones
for(i in 1:2) {
      plot(dates.40, scores(fauna.mem.40.pos, display="lc", choice=i), type="b", pch=19, main =
paste("RDA axis", i, ", positive temporal correlation model"), xlab="Date", ylab="RDA axis")
      }
# The single significant negative axis
      plot(dates.40, scores(fauna.mem.40.neg, display="lc", choice=1), type="b", pch=19, main =
paste("RDA axis", 1, ", negative temporal correlation model"), xlab="Date", ylab="RDA axis",
col="red", col.main="red")
par(mfrow=c(1,1))

# --------

# Select the dbMEM that are useful for modelling
sel.mem.40 <- forward.sel(fauna.hel.40, time.mem.40$vectors, nperm=9999, alpha=0.10)
sel.mem.40
# Do not include variables with p-values that are much larger than 0.05

# Three models: all selected dbMEM, then those modelling positive and negative correlation
mem.select <- sort(sel.mem.40$order[sel.mem.40$pval<=0.05])
# This selection is the same as: mem.select <- c(2,3,5,6,8,11,21,25)
mem.select.pos <- c(2,3,5,6,8,11)
```

```
mem.select.neg <- c(21,25)

# Plot the selected dbMEM to see what they look like[2] (Appendix S3, Figure S3.4)
par(mfrow=c(4,2))
# The positive ones
for(i in 1:6) {
      plot(dates.40, time.mem.40$vectors[,mem.select.pos[i]], type="b", pch=19, main =
paste("Positive", mem.select.pos[i]), xlab="Date", ylab="dbMEM")
      }
# The negative ones
for(i in 1:2) {
      plot(dates.40, time.mem.40$vectors[,mem.select.neg[i]], type="b", pch=19, main =
paste("Negative", mem.select.neg[i]), xlab="Date", ylab="dbMEM", col="red", col.main="red")
}
par(mfrow=c(1,1))

# Analysis based on the selected dbMEM eigenvectors only
# Compute RDA of the fauna by the selected dbMEM in each group (positive, negative)
fauna.mem.40.pos.sel <- rda(fauna.hel.40 ~ .,
      as.data.frame(time.mem.40$vectors[,mem.select.pos]))
anova(fauna.mem.40.pos.sel, by="axis")

# Plot the RDA axes of the significant canonical dbMEM models produced by the selected MEM
# (Appendix S3, Figure S3.5)
par(mfrow=c(3,1))
# The positive ones
for(i in 1:2) {
      plot(dates.40, scores(fauna.mem.40.pos.sel, display= "lc", choices=i), type="b", pch=19,
main = paste("RDA axis", i, ", positive temporal correlation model"), xlab="Date", ylab="RDA
axis")
      }
# The single significant negative axis
      plot(dates.40, scores(fauna.mem.40.neg.sel, display= "lc", choices=1), type="b", pch=19,
main = paste("RDA axis", 1, ", negative temporal correlation model"), xlab="Date", ylab="RDA
axis", col="red", col.main="red")
par(mfrow=c(1,1))

# --------

# Variation along the significant MEM model axes can be interpreted by stepwise selection in
regression, using the available environmental variables assembled in file envir.40.

# Among the explanatory variables, we added annual Principal Component (PC)-based indices
# of the North Atlantic Oscillation (NAO), obtained from the National Center for Atmospheric
# Research Climate Data Guide (https://climatedataguide.ucar.edu/)

# Values of the North Atlantic Oscillation (NAO), yearly data from 1996 to 2008
```

[2] In eigenvectors (which include eigenfunctions) and ordination axes computed on different computers or different software, all signs may be reversed, producing graphs that are mirror images of graphs produced on other software or computers. Such inversions have no effect on the interpretation of ordination diagrams or eigenfunctions.

```
NAO <- c(-1.07, -1.07, 0.03, 0.03, -0.29, -0.29, 0.55, 0.55, 0.63, 0.63, -0.57, -0.57, 0.35, 0.35,
0.32, 0.32, 0.10, 0.10, -0.76, -0.76, -0.26, -0.26, 0.54, 0.54, 0.08, 0.08)

# Analyse MEM model 1 representing positive temporal correlation by stepwise regression
# (variable target.1)
envir.40 <- cbind(sediment.40, waterquality.40, sampling.40$SEASON, NAO)
target.1 <- summary(fauna.mem.40.pos.sel)$constraints[,1]
m0 <- lm(target.1 ~ 1, data= envir.40)
mtot <-
lm(target.1 ~ CONDUCT + DO + PH + SALINITY + WTEMP + MOIST + SAND +
SILTCLAY + TC + TN + NAO, sampling.40$SEASON, data= envir.40)
res.step.1 <- step(m0, scope=formula(mtot), direction="both", trace=-1)
summary(res.step.1)

# NAO and TN are selected and significant at the 0.05 level.
# Repeat the regression using only these two variables:
res.model1 <- lm(target.1 ~ NAO + TN, data= envir.40)
summary(res.model1)

# Analyse MEM model 2 representing positive temporal correlation by stepwise regression
# (variable target.2)
target.2 <- summary(fauna.mem.40.pos.sel)$constraints[,2]
m0 <- lm(target.2 ~ 1, data= envir.40)
mtot <-
lm(target.2 ~ CONDUCT + DO + PH + SALINITY + WTEMP + MOIST + SAND +
SILTCLAY + TC + TN + NAO + sampling.40$SEASON, data= envir.40)
res.step.2 <- step(m0, scope=formula(mtot), direction="both", trace=-1)
summary(res.step.2)   # TN is the only explanatory variable selected, but it is not significant.

# Analyse the single axis representing negative temporal correlation by stepwise regression
# (variable target.3)
target.3 <- summary(fauna.mem.40.neg.sel)$constraints[,1]
m0 <- lm(target.3 ~ 1, data= envir.40)
mtot <-
lm(target.3 ~ CONDUCT + DO + PH + SALINITY + WTEMP + MOIST + SAND +
SILTCLAY + TC + TN + NAO + sampling.40$SEASON, data= envir.40)
res.step.3 <- step(m0, scope=formula(mtot), direction="both", trace=-1)
summary(res.step.3)

# Sampling season is selected and significant at the 0.05 level.
# Repeat the regression using only that variable:
model.4 <- lm(target.3 ~ SEASON, data = sampling.40)
```

 # 3.2.2. AEM analysis

```
# Construct the AEM eigenfunctions. Generate all AEM eigenfunctions.

# Construct a vector of weigths for the edges, each representing the easiness of exchange
# between adjacent dates (nodes).
# The 'max.d' value used here to scale the distances through weighting function 1
# (?weight.time) is the smallest distance for which no significant autocorrelation is found in the
```

```
# multivariate Mantel correlogram (correlog.40, section 3.4 of this document; Fig. S3.8).
weights <- weight.time(dates.40, alpha=2, max.d=522)

# Construct the AEMs themselves
aem.40.out <- aem.time(26, w=weights, moran=TRUE, plot.moran=TRUE)
aem.40.out$Moran
# How many AEM have positive Moran's I > E(I) and model positive temporal correlation?

# Compute the redundancy analysis (RDA) of the fauna by the matrix of AEM eigenvectors
# modelling positive temporal correlation
fauna.aem.40.pos <- rda(fauna.hel.40, aem.40.out$aem[, aem.40.out$Moran$Positive])
anova(fauna.aem.40.pos)
RsquareAdj(fauna.aem.40.pos)
# Examine the R2, the R2adj and the p-value.
# You may want to recompute the AEM eigenfunctions under the assumption that the dates are
# equidistant, using aem.time with option 'w=NULL', and compare the R2 coefficients.

# Compute the redundancy analysis (RDA) of the fauna by the matrix of AEM eigenvectors
# modelling negative temporal correlation
fauna.aem.40.neg <- rda(fauna.hel.40, aem.40.out$aem[, !aem.40.out$Moran$Positive])
anova(fauna.aem.40.neg)
RsquareAdj(fauna.aem.40.neg)
# Examine the R2, the R2adj and the p-value.

# Select the AEM that are useful for modelling
sel.aem.40 <- forward.sel(fauna.hel.40, aem.40.out$aem, nperm=9999, alpha=0.10)
sel.aem.40
# Do not include selected variables with p-values that are much larger than 0.05
# Compare to the results of forward selection of AEM assuming equidistant observations.

# Three models: all AEM selected, then those modelling positive and negative correlation
aem.select <- sort(sel.aem.40$order[sel.aem.40$pval<=0.08])
# This selection is the same as: aem.select <- c(1,2,3,6,10,15,21,24,25)
aem.select.pos <- c(1,2,3,6,10)
aem.select.neg <- c(15,21,24,25)

# Plot the selected AEM to see what they look like (Appendix S3, Figure S3.6)
par(mfrow=c(3,3))
# The positive ones
for(i in 1:5) {
    plot(dates.40, aem.40.out$aem[,aem.select.pos[i]], type="b", pch=19, main =
paste("Positive", aem.select.pos[i]), xlab="Date", ylab="AEM")
    }
# The negative ones
for(i in 1:4) {
    plot(dates.40, aem.40.out$aem[,aem.select.neg[i]], type="b", pch=19, main =
paste("Negative", aem.select.neg[i]), xlab="Date", ylab="AEM", col="red", col.main="red")
    }
par(mfrow=c(1,1))

# Compute RDA of the fauna by the selected AEM in each group (positive, negative), p <= 0.08
fauna.aem.40.pos.5 <- rda(fauna.hel.40~ ., as.data.frame(aem.40.out$aem[,aem.select.pos]))
```

```
anova(fauna.aem.40.pos.5, by="axis")
RsquareAdj(fauna.aem.40.pos.5)
fauna.aem.40.neg.4 <- rda(fauna.hel.40~ ., as.data.frame(aem.40.out$aem[,aem.select.neg]))
anova(fauna.aem.40.neg.4, by="axis")
RsquareAdj(fauna.aem.40.neg.4)

# Plot the RDA axes of the significant AEM models produced by the selected AEM
# (Appendix S3, Figure S3.7)
par(mfrow=c(3,1))
# The positive ones
for(i in 1:2) {
      plot(dates.40, scores(fauna.aem.40.pos.5, display="lc", choices=i), type="b", pch=19, main
= paste("RDA axis", i, ", positive temporal correlation model"), xlab="Seasons within years
1996-2008", ylab="RDA axis")
      }
# The single significant negative axis
      plot(dates.40, scores(fauna.aem.40.neg.4, display="lc", choices=1), type="b", pch=19, main
= paste("RDA axis", 1, ", negative temporal correlation model"), xlab="Seasons within years
1996-2008", ylab="RDA axis", col="red", col.main="red")
par(mfrow=c(1,1))

# Compare the MEM and AEM models. If one looks like a mirror image of the other, see
# footnote 2 a few pages back. If it is necessary to produce a mirror-image plot for the
# comparison, add a minus sign in the plotting script in front of the model to be drawn.

# --------

# Compare the MEM and AEM eigenfunctions through RV coefficients

# RV coefficient between MEM and AEM modeling positive temporal correlation
RV.pos <- RV.rtest(as.data.frame(time.mem.40$vectors[, time.mem.40$Moran_I$Positive]),
as.data.frame(aem.40.out$aem[, aem.40.out$Moran$Positive]))
RV.pos

# RV coefficient between MEM and AEM modeling negative temporal correlation
RV.neg <- RV.rtest(as.data.frame(time.mem.40$vectors[, !time.mem.40$Moran_I$Positive]),
as.data.frame(aem.40.out$aem[, !aem.40.out$Moran$Positive]))
RV.neg
```

 # 3.2.3. Scalogram of the dbMEM eigenfunctions

```
# Compute semipartial R2 for each dbMEM modelling positive temporal correlation, separately
# Function R2.by.variable() is found in file "R_functions_for_Practicals.txt" loaded in sect. 2.2.

R2.out <- R2.by.variable(fauna.hel.40, time.mem.40$vectors[,1:12], scale.Y=FALSE)
R2.out
# The output table lists the semipartial R2 computed separately for each dbMEM, the F-statistics
# and the p-values of the semipartial test. The semipartial R2 will be used in the scalogram. The
# test displayed in the table is very conservative because it tests the contribution of each MEM
# eigenfunction above and beyond all other MEM variables in the analysis.
```

```
# In the scalogram, the squares represent the semipartial R2 coefficients associated with each
# MEM eigenfunction. We chose to highlight (black squares) the dbMEM that significantly
# contribute sequentially to the explanation of the faunal response data, as in Legendre &
# Legendre (2012, Fig. 14.5). These significant dbMEM are found in the output list of the
# forward.sel() function. The scalogram produced here only concerns the dbMEM variables
# modelling positive temporal correlation.

# Plot the scalogram (Appendix S3, Figure S3.8)
# Open squares: pch=22; filled black or white squares: pch=15.
plot(1:12, R2.out[,1], type="o", pch=22, cex=1.0, xlim=c(1,12), ylim=c(0,0.25), xlab="dbMEM
1-12", ylab="R-square", main="Scalogram of positively correlated dbMEM, Chesapeake site
40")
points(1:12, R2.out[,1], pch=15, col="white", cex=0.8)
points(mem.select.pos, R2.out[mem.select.pos,1], pch=15, cex=1.0)


# 3.3. VARIATION PARTITIONING
#       involving environmental variables and dbMEM eigenfunctions (Figure 5 of the paper)

# Selection of the site 40 sediment variables
(res.sel1 <- forward.sel(fauna.hel.40, sediment.40, nperm=9999, alpha=0.10))
# You will receive an error message because no sediment variable was selected

# Selection of water quality variables
(res.sel2 <- forward.sel(fauna.hel.40, waterquality.40, nperm=9999, alpha=0.10))
# A single variable was selected: SALINITY (variable 4)

# Variation partitioning: waterquality, positive MEM, negative MEM
res.part <- varpart(fauna.hel.40, waterquality.40[,4], time.mem.40$vectors[,mem.select.pos],
time.mem.40$vectors[,mem.select.neg])
res.part
plot(res.part, digits=2)

# Example of partial test: test partial effect of waterquality while controlling for MEM
mod1 <- rda(fauna.hel.40, waterquality.40[,4], time.mem.40$vectors[,mem.select])
anova(mod1, step=1000, perm.max=1000)
RsquareAdj(mod1)
# Then, compare the adjusted R-square ($adj.r.squared) to line [a] of the varpart result.
# This illustrates the calculations done by function varpart().

# Example of partial test: test MEM.pos while controlling for (waterquality and MEM.neg)
mod2 <- rda(fauna.hel.40, time.mem.40$vectors[,mem.select.pos], cbind(waterquality.40[,4],
time.mem.40$vectors [,mem.select.neg]))
anova(mod2, step=1000, perm.max=1000)
RsquareAdj(mod2)

# Example of partial test: test MEM.neg while controlling for (waterquality and MEM.neg)
mod3 <- rda(fauna.hel.40, time.mem.40$vectors[,mem.select.neg], cbind(waterquality.40[,4],
time.mem.40$vectors[,mem.select.pos]))
anova(mod3, step=1000, perm.max=1000)
RsquareAdj(mod3)
```

```
# Many of the steps above are automatically computed by function quickPCNM.

# Fitted values, MEM modelling positive temporal correlation: plot maps using quickPCNM
res.pos = quickPCNM(fauna.hel.40, dates.40, method="none", myPCNM= time.mem.40$vectors
[,mem.select.pos], detrend=FALSE)
summary(res.pos)
res.pos$RDA_test
res.pos$RDA_axes_test
# Examine adjusted R2 of model and p-values of the axes. How many significant axes?
# Compare to the results in section 3.2.1 (model: fauna.mem.40.pos.sel)

# Fitted values, MEM modelling negative temporal correlation: plot maps using quickPCNM
res.neg = quickPCNM(fauna.hel.40, as.numeric(dates.40), method="none", myPCNM=
time.mem.40$vectors [,mem.select.neg], detrend=FALSE)
res.neg$RDA_test
res.neg$RDA_axes_test
# Examine adjusted R2 of model and p-values of the axes. How many significant axes?
# Compare to the results in section 3.2.1 (model: fauna.mem.40.neg.sel).


# 3.4. MULTIVARIATE CORRELOGRAM
#       See “Numerical ecology with R” (2011) and “Numerical ecology” (2012) for details.

# Multivariate Mantel correlogram  (Appendix S3, Figure S3.9)
# See "Numerical Ecology" (2012), Section 13.1.6
# See "Numerical Ecology with R" (2011), p. 235
correlog.40 <- mantel.correlog(dist(fauna.hel.40), XY=dates.40, n.class=26)
plot(correlog.40)


# 3.5. TIME-CONSTRAINED CLUSTERING: MULTIVARIATE REGRESSION TREE
#       See “Numerical ecology with R” (2011) and “Numerical ecology” (2012) for details.

# Use this method to identify one or several breakpoints in a data series
# See "Numerical Ecology" (2012), Section 12.6.4
# See "Numerical Ecology with R" (2011), Section 4.11.5, for details about graph interpretation
require(mvpart)

# Generate a "time.seq" variable containing the integers 1 to 26
time.seq = as.data.frame(1:26)          # Generate a "time.seq" variable containing integers 1 to 26
colnames(time.seq) = "time.seq"
part.res <- mvpart(data.matrix(fauna.hel.40) ~ time.seq, data=time.seq, xv="pick", xvmult=100)

# Click on the red dot of the graph to obtain the tree.
# Two groups represent the solution with the smallest CVRE (“best” in that sense).
# Find the observations in which group:
part.res$where

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

4. TWO-WAY MANOVA BY PARTIAL RDA

This section shows how to carry out a two-way multivariate ANOVA (called MANOVA) for community composition data. The two factors of interest are year and season.

4.1. TWO-WAY TEMPORAL MANOVA OF FIVE SITES

```
# Test whether the year and season factors can explain a significant fraction of the multivariate
# dispersion within a group of geographically close sites. Sites 22, 23, 201, 202 and 203 are
# considered replicates in this analysis. Distances between these sites ranges from 2.8 to 15.9 km
# (mean of 7.9 km); they are thus are far enough from one another that the faunal data should not
# be pseudoreplicated.

# Select a subset of five sites for multivariate MANOVA
curr.siteS <- c(201, 202, 203, 22, 23)

# These five sites are clustered in the northern portion of Chesapeake Bay near the city of
# Baltimore, Maryland. Run the code to see the map, which is not shown in Appendix S3
plot(xy[,c(2,1)], xlab="Longitude W", ylab="Latitude N", asp=1)
points(xy[rownames(xy) %in% curr.siteS,c(2,1)], pch=19, col= "red")
text(xy[,c(2,1)], labels=rownames(xy), pos=4)

# Extract sampling and fauna data for the sites in set curr.siteS
sampling.manova1 <- sampling[sampling$STATION %in% curr.siteS, ]
fauna.manova1 <- fauna[sampling$STATION %in% curr.siteS, ]

# Remove the "empty" taxa
fauna.manova1 <- fauna.manova1[ , colSums(fauna.manova1)!=0]   # (130x70)

# Remove unused levels from sampling data
# Function drop.levels() is found in file "R_functions_for_Practicals.txt" loaded in section 2.2.
sampling.manova1 <- drop.levels(sampling.manova1)

# What do we have on hand?
dim(fauna.manova1)
dim(sampling.manova1)
summary(sampling.manova1)

# Create a factor for YEAR by transforming vector sampling.manova1$YEAR
year.fac <- as.factor(sampling.manova1$YEAR)
year.fac

# Create a factor for SEASON
season.fac <- sampling.manova1$SEASON
season.fac

# Make sure that the factors are balanced, i.e. same number of observations in each cell
table(season.fac, year.fac)

# Create Helmert contrasts for the factors and their interaction.
# Explanation – In two-way anova, Helmert contrasts are used to code for the main factors A and
# B. Variables representing the interaction are generated by computing the products of all
```

```
# Helmert variables coding for A by all variables coding for B. As a result, the set of variables
# coding for the interaction is orthogonal to the set of variables coding for A and for B. The
# fractions of variation explained by A, B and the interaction are thus linearly independent of
# one another.

year.season.helm <- model.matrix(~ season.fac * year.fac,
    contrasts=list(season.fac="contr.helmert", year.fac="contr.helmert"))
# year.season.helm   # If you want to look at the Helmert contrasts

# Property 1 of Helmert contrasts: all variables should sum to 0
apply(year.season.helm[, 2:ncol(year.season.helm)], 2, sum)

# Property 2: the cross products (scalar products) of the Helmert contrasts should all be 0,
# showing that they are orthogonal to one another
res <- t(year.season.helm[,-1]) %*% year.season.helm[,-1]
head(res)   # Check that the non-diagonal terms of matrix "res" are all 0.

# Transform the species abundance data using the Hellinger transformation
fauna.hel.manova1 <- decostand(fauna.manova1, method="hellinger")

# --------

# Before proceeding with MANOVA, we must test whether there is homogeneity of the
# multivariate within-group covariance matrices

# Compute the Hellinger distance matrix from the transformed data
fauna.hel.manova1.D1 <- dist(fauna.hel.manova1)

# We cross the season & year factors to create the groups
year.season.fac <- as.factor(paste(year.fac, season.fac, sep="."))
year.season.fac

# Test of homogeneity of the multivariate within-group covariance matrices
fauna.hel.manova1.MHV <- betadisper(fauna.hel.manova1.D1, year.season.fac)
permutest(fauna.hel.manova1.MHV)

# If p < 0.05, the test rejects H0: the multivariate within-group covariance matrices are
# homogeneous. If H0 is not rejected, we can proceed with the analysis of variance.

# --------

# Multivariate analysis of variance

# Step 1. Check if there is a significant interaction between sampling YEAR and SEASON.
# We use the interaction terms as explanatory variables, the factors themselves as covariables.
# Do not use column 1 (Intercept). Look at colnames(year.season.helm) to find out which
# columns represent the different terms (factors and interactions). It is important to choose the
# correct columns as explanatory variables and covariables in the three analyses that follow.
season.year.rda1 <- rda(fauna.hel.manova1, year.season.helm[, 15:26], year.season.helm[, 2:14])
anova(season.year.rda1, step=1000, perm.max=1000, model="direct")
RsquareAdj(season.year.rda1)
```

```
# Is the interaction significant? If it is, MANOVAs of the YEAR factor should be computed for
# each season separately, and conversely.

# Step 2. Can factor SEASON explain a significant portion of the multivariate dispersion?
# Sampling YEAR and interaction are used as covariables.
season.rda1 <- rda(fauna.hel.manova1, year.season.helm[, 2], year.season.helm[, 3:26])
anova(season.rda1, step=1000, perm.max=1000, strata=year.fac, model="direct")
RsquareAdj(season.rda1)      # Measure of effect size

# Step 3. Can factor sampling YEAR explain a significant portion of the multivariate dispersion?
# SEASON and interaction are used as covariables.
year.rda1 <- rda(fauna.hel.manova1, year.season.helm[, 3:14], year.season.helm[, c(2, 15:26)])
anova(year.rda1, step=1000, perm.max=1000, strata=season.fac, model="direct")
RsquareAdj(year.rda1)         # Measure of effect size

# Which factor explains the largest fraction of the multivariate faunal dispersion?

# ========

# 4.2. SPACE-TIME STUDY: VARIABILITY AMONG YEARS AND REGIONS, ONE SEASON AT A TIME

# Sites 43, 44 and 47 are in the Potomac River estuary in the south-west of Chesapeake Bay,
# sites 201, 202 and 203 are in the inlet near Baltimore in the north. The 6 sites form two groups
# with three replicates each. We will test whether the year and site.group factors can explain the
# multivariate dispersion between two groups of geographically distant sites during each season.

# Select a subset of sites for multivariate MANOVA
curr.siteS <- c(43, 44, 47, 201, 202, 203)

# Plot sites on a simple map; run the code to see the map
plot(xy[,c(2,1)], xlab="Longitude W", ylab="Latitude N", asp=1)
points(xy[rownames(xy) %in% curr.siteS,c(2,1)], pch=19, col= "red")
text(xy[,c(2,1)], labels=rownames(xy), pos=4)

# Extract sampling and fauna for curr.siteS
fauna.6sites <- fauna[sampling$STATION %in% curr.siteS, ]
sampling.6sites <- sampling[sampling$STATION %in% curr.siteS, ]


# 4.2.1. MANOVA of the spring surveys
#         sites 43, 44, 47, 201, 202 and 203

# Select the spring sampling units.
fauna.manova2 <- fauna.6sites[sampling.6sites$SEASON == "Spring", ]
sampling.manova2 <- sampling.6sites[sampling.6sites$SEASON == "Spring", ]

# Remove the "empty" taxa
fauna.manova2 <- fauna.manova2[ , colSums(fauna.manova2)!=0]

# Remove unused levels from sampling data
# Function drop.levels() is found in file “R_functions_for_Practicals.txt” loaded in section 2.2.
sampling.manova2 <- drop.levels(sampling.manova2)
```

```
# What do we have on hand?
dim(fauna.manova2)
dim(sampling.manova2)
summary(sampling.manova2)

# Create a factor for YEAR (identical to year.fac generated in 4.1)
year.fac <- as.factor(sampling.manova2$YEAR)
year.fac

# Create a factor describing the two groups of sites.
site.vec <- as.character(sampling.manova2$STATION)
site.vec[site.vec %in% c("43","44","47")] <- "43.44.47"
site.vec[site.vec %in% c("201","202","203")] <- "201.202.203"
site.group.fac <- as.factor(site.vec)
site.group.fac

# Make sure that the factors are balanced, i.e. same number of observations in each cell
table(site.group.fac, year.fac)

# Helmert contrasts for the factors and interaction
year.group.helm <- model.matrix(~ site.group.fac * year.fac,
    contrasts=list(site.group.fac="contr.helmert", year.fac="contr.helmert"))
# year.group.helm   # If you want to look at the Helmert contrasts

# Property 1 of Helmert contrasts: all variables should sum to 0
apply(year.group.helm[, 2:ncol(year.group.helm)], 2, sum)

# Property 2: the cross products (scalar products) of the Helmert contrasts should all be 0,
# showing that they are orthogonal to one another
res <- t(year.group.helm[,-1]) %*% year.group.helm[,-1]
head(res)   # Check that the non-diagonal terms of matrix "res" are all 0.

# Transform the species abundance data using the Hellinger transformation
fauna.hel.manova2 <- decostand(fauna.manova2, method="hellinger")

# --------

# Test homogeneity of the multivariate within-group covariance matrices

# Compute the Hellinger distance matrix from the transformed data
fauna.hel.manova2.D1 <- dist(fauna.hel.manova2)

# We cross the year and site.group factors to create the groups
year.site.group.fac <- as.factor(paste(year.fac, site.group.fac, sep="."))

# Test of homogeneity of the multivariate within-group covariance matrices
fauna.hel.manova2.MHV <- betadisper(fauna.hel.manova2.D1, year.site.group.fac)
permutest(fauna.hel.manova2.MHV)

# --------
```

```
# Multivariate analysis of variance of the spring surveys

# Step 1. Check if there is a significant interaction between the site groups and YEAR.
# We use the interaction terms as explanatory variables, the factors themselves as covariables.
# Do not use column 1 (Intercept). Look at colnames(year.group.helm) to find out which
# columns represent the different terms (factors and interactions). It is important to choose the
# correct columns as explanatory variables and covariables in the three analyses that follow.
site.year.rda2 <- rda(fauna.hel.manova2, year.group.helm[, 15:26], year.group.helm[, 2:14])
anova(site.year.rda2, step=1000, perm.max=1000, model="direct")
RsquareAdj(site.year.rda2)
# Is the interaction significant? If it is, MANOVAs of the YEAR factor should be computed for
# each site group separately, and conversely.

# Step 2. Can factor site.group.fac explain a significant portion of the multivariate dispersion?
# Sampling YEAR and interaction are used as covariables.
site.rda2 <- rda(fauna.hel.manova2, year.group.helm[, 2], year.group.helm[, 3:26])
anova(site.rda2, step=1000, perm.max=1000, strata=year.fac, model="direct")
RsquareAdj(site.rda2)

# Step 3. Can factor sampling YEAR explain a significant portion of the multivariate dispersion?
# Factor site.group.fac and interaction are used as covariables.
year.rda2 <- rda(fauna.hel.manova2, year.group.helm[, 3:14], year.group.helm[, c(2, 15:26)])
anova(year.rda2, step=1000, perm.max=1000, strata=site.group.fac, model="direct")
RsquareAdj(year.rda2)


# 4.2.2. MANOVA of the fall surveys
#        sites 43, 44, 47, 201, 202 and 203

# Select the fall sampling units.
fauna.manova3 <- fauna.6sites[sampling.6sites$SEASON == "Fall", ]
sampling.manova3 <- sampling.6sites[sampling.6sites$SEASON == "Fall", ]

# Remove the "empty" taxa
fauna.manova3 <- fauna.manova3[ , colSums(fauna.manova3)!=0]

# Remove unused levels from sampling data
# Function drop.levels() is found in file “R_functions_for_Practicals.txt” loaded in section 2.2.
sampling.manova3 <- drop.levels(sampling.manova3)

# What do we have on hand?
dim(fauna.manova3)
dim(sampling.manova3)
summary(sampling.manova3)

# Create a factor for YEAR (identical to year.fac generated in 4.1)
year.fac <- as.factor(sampling.manova3$YEAR)
year.fac

# Create a factor describing the two groups of sites (identical to site.group.fac generated in 4.2.1)
site.vec <- as.character(sampling.manova3$STATION)
site.vec[site.vec %in% c("43","44","47")] <- "43.44.47"
```

```
site.vec[site.vec %in% c("201","202","203")] <- "201.202.203"
site.group.fac <- as.factor(site.vec)
site.group.fac

# Make sure that the factors are balanced, i.e. same number of observations in each cell
table(site.group.fac, year.fac)

# Helmert contrasts for the factors and their interaction (identical to year.group.helm in 4.2.1)
year.group.helm <- model.matrix(~ site.group.fac * year.fac,
    contrasts=list(site.group.fac="contr.helmert", year.fac="contr.helmert"))
# year.group.helm   # If you want to look at the Helmert contrasts

# Property 1 of Helmert contrasts: all variables should sum to 0
apply(year.group.helm[, 2:ncol(year.group.helm)], 2, sum)

# Property 2: the cross products (scalar products) of the Helmert contrasts should all be 0,
# showing that they are orthogonal to one another
res <- t(year.group.helm[,-1]) %*% year.group.helm[,-1]
head(res)   # Check that the non-diagonal terms of matrix "res" are all 0.

# Transform the species abundance data using the Hellinger transformation
fauna.hel.manova3 <- decostand(fauna.manova3, method="hellinger")

# --------

# Test homogeneity of the multivariate within-group covariance matrices

# Compute the Hellinger distance matrix from the transformed data
fauna.hel.manova3.D1 <- dist(fauna.hel.manova3)

# We cross the year and site.group factors to create the groups
year.site.group.fac <- as.factor(paste(year.fac, site.group.fac, sep="."))

# Test of homogeneity of the multivariate within-group covariance matrices
fauna.hel.manova3.MHV <- betadisper(fauna.hel.manova3.D1, year.site.group.fac)
permutest(fauna.hel.manova3.MHV)

# --------

# Multivariate analysis of variance of the fall surveys

# Step 1. Check if there is a significant interaction between the site groups and YEAR.
# We use the interaction terms as explanatory variables, the factors themselves as covariables.
# Do not use column 1 (Intercept). Look at colnames(year.group.helm) to find out which
# columns represent the different terms (factors and interactions). It is important to choose the
# correct columns as explanatory variables and covariables in the three analyses that follow.
site.year.rda3 <- rda(fauna.hel.manova3, year.group.helm[, 15:26], year.group.helm[, 2:14])
anova(site.year.rda3, step=1000, perm.max=1000, model="direct")
RsquareAdj(site.year.rda3)
# Is the interaction significant? If it is, MANOVAs of the YEAR factor should be computed for
# each site group separately, and conversely.
```

```
# Step 2. Can factor site.group.fac explain a significant portion of the multivariate dispersion?
# Sampling YEAR and interaction are used as covariables.
site.group.rda3 <- rda(fauna.hel.manova3, year.group.helm[, 2], year.group.helm[, 3:26])
anova(site.group.rda3, step=1000, perm.max=1000, strata=year.fac, model="direct")
RsquareAdj(site.group.rda3)

# Step 3. Can factor sampling YEAR explain a significant portion of the multivariate dispersion?
# Factor site.group.fac and interaction are used as covariables.
year.rda3 <- rda(fauna.hel.manova3, year.group.helm[, 3:14], year.group.helm[, c(2, 15:26)])
anova(year.rda3, step=1000, perm.max=1000, strata=site.group.fac, model="direct")
RsquareAdj(year.rda3)

# Which factor explains the largest fraction of the multivariate faunal dispersion?

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

5. SPACE-TIME ANALYSIS: LOCAL CONTRIBUTIONS TO BETA DIVERSITY (LCBD)

```
# Local Contributions to Beta Diversity (LCBD indices) are comparative indicators of the
# ecological uniqueness of the sampling units. The LCBD values indicate how much each
# observation contributes to beta diversity compared to a site with average species composition,
# which would have a LCBD value of 0. Sampling units may have large LCBD values for
# different reasons, as explained in the paper.

# Download function beta.div() from Appendix S4 of the Legendre & De Cáceres (2013) paper
# at http://onlinelibrary.wiley.com/doi/10.1111/ele.12141/suppinfo (Supporting information).
# The function itself is in file ele12141-sup-0005-AppendixS4.R, whereas the documentation
# is found in file ele12141-sup-0004-AppendixS4.pdf.
```

5.1. PREPARE A DATA FILE

```
# Create a simple factor for seasons for the 27 sites, 13 years and 2 seasons, length = 702
season27 <- sampling$SEASON
```

5.2. COMPUTE LCDB INDICES OVER THE 27 SITES, SPRING AND FALL SEPARATELY

```
beta.out.27.spring <- beta.div(fauna[season27=="Spring",], method="hellinger", nperm=499)
beta.out.27.fall <- beta.div(fauna[season27=="Fall",], method="hellinger", nperm=499)
# Warning – It is better to use 999 permutations. Permutation tests take a bit of time.
# The LCBD indices are available in the $LCBD vector, the permutational p-values in $p.LCBD
signif.27.spring <- which(beta.out.27.spring$p.LCBD <= 0.05)
signif.27.fall <- which(beta.out.27.fall$p.LCBD <= 0.05)

# Copy the LCBD computed for each season to matrices (27 sites x 13 years)
LCBD.27.spring <- matrix(beta.out.27.spring$LCBD, 27, 13, byrow=TRUE)
LCBD.27.fall <- matrix(beta.out.27.fall$LCBD, 27, 13, byrow=TRUE)
rownames(LCBD.27.spring) <- rownames(LCBD.27.fall) <- rownames(xy)
colnames(LCBD.27.spring) <- colnames(LCBD.27.fall) <- 1996:2008

# Compute LCBD values per site, summed over the years, for each season separately
LCBD.27.per.site.spring <- apply(LCBD.27.spring, 1, sum)
LCBD.27.per.site.fall <- apply(LCBD.27.fall, 1, sum)

# Compute LCBD values per year, summed over the sites, for each season separately
LCBD.27.per.year.spring <- apply(LCBD.27.spring, 2, sum)
LCBD.27.per.year.fall <- apply(LCBD.27.fall, 2, sum)

# --------

# Maps of the LCBD values per site, summed over the years, for each season separately.
# Note – In this plot and onwards, the square roots of the LCBD values are used in the argument
# value of parameter "cex" in order to plot bubbles whose areas (rather than radii) are
# proportional to the LCBD values.

# Map of LCBD values per site, summed over the years, spring; Appendix S3, Figure S3.10
plot(xy[,c(2,1)], asp=1, type="n", xlab="Latitude", ylab="Longitude", main="LCBD indices,
Chesapeake sites, spring, all years", xlim=c(-77.3,-75.7))
points(xy[,c(2,1)],pch=21, col="white", bg="steelblue2", cex=15*sqrt(LCBD.27.per.site.spring))
```

```
text(xy[,c(2,1)], labels=rownames(xy), pos=4)

# Map of LCBD values per site, summed over the years, fall; run the code to see the map
plot(xy[,c(2,1)], asp=1, type="n", xlab="Latitude", ylab="Longitude", main=" LCBD indices,
Chesapeake sites, fall, all years ", xlim=c(-77.3,-75.7))
points(xy[,c(2,1)], pch=21, col="white", bg="steelblue2", cex=15*sqrt(LCBD.27.per.site.fall))
text(xy[,c(2,1)], labels=rownames(xy), pos=4)

# Time maps of LCBD per year, summed over sites, for each season; App. S3, Figure S3.11
par(mfrow=c(2,1))
plot(1996:2008, rep(0, 13), type="p", xlab="Sampling years", ylab="", main="LCBD indices
along years, Chesapeake, spring", ylim=c(-1,1), pch=21, col="white", bg="steelblue2",
cex=20*sqrt(LCBD.27.per.year.spring))
plot(1996:2008, rep(0, 13), type="p", xlab="Sampling years", ylab="", main="LCBD indices
along years, Chesapeake, fall", ylim=c(-1,1), pch=21, col="white", bg="steelblue2",
cex=20*sqrt(LCBD.27.per.year.fall))
par(mfrow=c(1,1))

# --------

# Plot space-time maps of LCBD, spring and fall data; Appendix S3, Figure S3.12
par(mfrow=c(1,2))
seq.X.27 <- rep(1996:2008, 27)
seq.Y.27 <- rep(1:27, each=13)

plot(seq.X.27, seq.Y.27, asp=1, type="n", ylab="Sites", xlab="Years", main="Space-time map,
LCBD, spring", ylim=c(1,27), xlim=c(1996,2008), yaxt="n", cex.axis=0.8)
points(seq.X.27, seq.Y.27, pch=21, col="white", bg="steelblue2",
cex=30*sqrt(beta.out.27.spring$LCBD))
points(seq.X.27[signif.27.spring], seq.Y.27[signif.27.spring], pch=21, col="black",
bg="steelblue2", cex=30*sqrt(beta.out.27.spring$LCBD[signif.27.spring]))
axis(side=2, 1:27, labels=rownames(xy), las=1, cex.axis=0.8)

plot(seq.X.27, seq.Y.27, asp=1, type="n", ylab="Sites", xlab="Years", main="Space-time map,
LCBD, fall", ylim=c(1,27), xlim=c(1996,2008), yaxt="n", cex.axis=0.8)
points(seq.X.27, seq.Y.27, pch=21, col="white", bg="steelblue2",
cex=30*sqrt(beta.out.27.fall$LCBD))
points(seq.X.27[signif.27.fall], seq.Y.27[signif.27.fall], pch=21, col="black", bg="steelblue2",
cex=30*sqrt(beta.out.27.fall$LCBD[signif.27.fall]))
axis(side=2, 1:27, labels=rownames(xy), las=1, cex.axis=0.8)
par(mfrow=c(1,1))


# 5.3. REPEAT THE LCBD ANALYSES ON 25 BRACKISH SITES, SPRING AND FALL SEPARATELY
#       after excluding sites #36 and #79, located in freshwater


# 5.3.1. Prepare the data files

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 7
```

```
sediment.25 <- sediment[-freshwater,]                                    # 650 x 7
xy.25 <- xy[-c(12,27),]                                                  # 27 x 2

# Create a simple factor for seasons for the 25 sites, 13 years and 2 seasons, length = 650
season25 <- sampling.25$SEASON

# Assemble separate sampling data frames for the spring and fall data
temp = as.matrix(sampling.25)
sampling.25.spring <- as.data.frame(temp[temp[,6]=="Spring", ])          # 325 x 6
sampling.25.fall <- as.data.frame(temp[temp[,6]=="Fall", ])              # 325 x 6
```

 # 5.3.2. Compute LCDB indices

```
beta.out.25 <- beta.div(fauna.25, method="hellinger", nperm=0)           ### Used in section 5.3.3
beta.out.25.spring <- beta.div(fauna.25[season25=="Spring",], method="hellinger", nperm=499)
beta.out.25.fall <- beta.div(fauna.25[season25=="Fall",], method="hellinger", nperm=499)
# Warning – It is better to use 999 permutations. Permutation tests take a bit of time.
# 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)

# --------

# Plot space-time maps of LCBD, spring and fall data; Appendix S3, Figure S3.13
# 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))

# --------
```

```
# Analysis of variance of the LCBD indices, 25 sites, factors site and year

# Two-way anova for LCBD computed for 325 spring data only
tmp.25.1 <- sampling.25[sampling.25$SEASON=="Spring",]
tmp.25.1$YEAR <- as.factor(tmp.25.1$YEAR)
aov.LCBD.25.spring.out <- aov(beta.out.25.spring$LCBD ~ STATION+YEAR, data= tmp.25.1)
summary(aov.LCBD.25.spring.out)

# Two-way anova for LCBD computed for 325 fall data only
tmp.25.2 <- sampling.25[sampling.25$SEASON=="Fall",]
tmp.25.2$YEAR <- as.factor(tmp.25.2$YEAR)
aov.LCBD.25.fall.out <- aov(beta.out.25.fall$LCBD ~ STATION+YEAR, data= tmp.25.2)
summary(aov.LCBD.25.fall.out)
```

 # 5.3.3. Compare spring and fall LCBD: paired t*-test*

```
# The data in vector beta.out.25$LCBD subjected to the t-test were computed at the beginning
# of section 5.3.2 for the spring and fall faunal data together
(t.test.spring.fall.res <- t.test(beta.out.25$LCBD ~ rep(c(1,2), 325), paired=TRUE))

# LCBD values in a study always sum to 1
(sum.spring <- sum(beta.out.25$LCBD[seq(1, 650, 2)]))
(sum.fall <- sum(beta.out.25$LCBD[seq(2, 650, 2)]))
sum.spring + sum.fall
```

5.4. ENVIRONMENTAL VARIABLES EXPLAINING LCBD VARIATION ACROSS YEARS, 25 SITES

```
# Values of North Atlantic Oscillation (NAO), yearly data, from 1996 to 2008. See section 3.2.1.
NAO <- c(-1.07, -1.07, 0.03, 0.03, -0.29, -0.29, 0.55, 0.55, 0.63, 0.63, -0.57, -0.57, 0.35, 0.35,
0.32, 0.32, 0.10, 0.10, -0.76, -0.76, -0.26, -0.26, 0.54, 0.54, 0.08, 0.08)
NAO <- rep(NAO, 25)

site.names.25 <- rep(rownames(xy)[-c(12,27)], each=26)

# Recode factor site using Helmert contrast variables
sites.helmert <- model.matrix(~as.factor(site.names.25), contrasts="contr.helmert")[,-1]
colnames(sites.helmert) <- rownames(xy)[-c(1,12,27)]       # Site 1 has no Helmert variable
explanatory.25bis <- cbind(waterquality.25, sediment.25, NAO, sites.helmert,
sampling.25$SEASON)
```

 # 5.4.1. Spring only

```
# Vector listing only the spring obervations
spring.obs <- which(sampling.25$SEASON == "Spring")

# Selection of explanatory variables will be done using partial regression controlling for factor
# site, which is likely to account for a large fraction of the variation in the data. Partial linear
# regression is linear regression computed after residualizing the response and explanatory
# variables onto the variables or factors one wants to control for, here factor site.
# See Legendre & Legendre (2012, section 10.3.5).
```

```
# Compute the two regression models using lm() and obtain residuals
# 1. Residualize the LCBD data on factor site
LCBD.spring.resid <- resid(lm(beta.out.25.spring$LCBD ~ ., explanatory.25bis[spring.obs,
12:35]))

# 2. Residualize the explanatory data on factor site
envir.spring.resid <- resid(lm(as.matrix(explanatory.25bis[spring.obs, 1:11]) ~ .,
explanatory.25bis[spring.obs, 12:35]))

# Selection of environmental variables using function forward.sel() of package 'packfor'
(sel.spring.res <- forward.sel(LCBD.spring.resid, envir.spring.resid))

# For partitioning, the response data will be the original (not residualized) LCBD values.
# For explanatory, use CONDUCT (var. 1), DO (var. 2), SALINITY (var. 4) and NAO (var. 11)
# and the Helmert contrasts representing factor site.
(part.25.spring.res <- varpart(beta.out.25.spring$LCBD, explanatory.25bis[spring.obs,
c(1,2,4,11)], explanatory.25bis[spring.obs, 12:35]))

# Test of the unique contribution of the 4 environmental variables, controlling for site
test.4var <- rda(beta.out.25.spring$LCBD, explanatory.25bis[spring.obs, c(1,2,4,11)],
explanatory.25bis[spring.obs, 12:35])
anova(test.4var)

# Test of the unique contribution of factor site, controlling for the 4 environmental
test.site <- rda(beta.out.25.spring$LCBD, explanatory.25bis[spring.obs, 12:35],
explanatory.25bis[spring.obs, c(1,2,4,11)])
anova(test.site)
```

 # 5.4.2. Fall only

```
# Vector listing only the fall obervations
fall.obs <- which(sampling.25$SEASON == "Fall")

# Compute the two regression models using lm() and obtain residuals
# 1. Residualize the LCBD data on factor site
LCBD.fall.resid <- resid(lm(beta.out.25.fall$LCBD ~ ., explanatory.25bis[fall.obs, 12:35]))

# 2. Residualize the explanatory data on factor site
envir.fall.resid <- resid(lm(as.matrix(explanatory.25bis[fall.obs, 1:11]) ~ .,
explanatory.25bis[fall.obs, 12:35]))

# Selection of environmental variables using function forward.sel() of package 'packfor'
 (sel.fall.res <- forward.sel(LCBD.fall.resid, envir.fall.resid, alpha = 0.08))

# --------

# For partitioning, the response data will be the original (not residualized) LCBD values.
# For explanatory, use CONDUCT (var. 1), PH (var 3), SALINITY (var. 4) and NAO (var. 11)
# and the Helmert contrasts representing factor site.
```

```
(part.25.fall.res <- varpart(beta.out.25.fall$LCBD, explanatory.25bis[fall.obs, c(1,3,4,11)],
explanatory.25bis[fall.obs, 12:35]))

# Test of the unique contribution of the 4 environmental variables, controlling for site
test.4var <- rda(beta.out.25.fall$LCBD, explanatory.25bis[fall.obs, c(1,3,4,11)],
explanatory.25bis[fall.obs, 12:35])
anova(test.4var)

# Test of the unique contribution of factor site, controlling for the 4 environmental
test.site <- rda(beta.out.25.fall$LCBD, explanatory.25bis[fall.obs, 12:35],
explanatory.25bis[fall.obs, c(1,3,4,11)])
anova(test.site)


# 5.5. CHANGES IN SPECIES COMPOSITION RELATED TO CHANGES IN LCBD AMONG YEARS

# This analysis will examine among-year variation at site 40, which was also used in section 3.1.
# This site has 36 taxa and exhibits strong variability among years, in both the spring and fall.

# Extract sampling and fauna data for site 40, spring and fall separately
sampling.40 <- sampling[sampling$STATION==40, ]
fauna.40 <- fauna[sampling$STATION==40, ]
fauna.40.spring <- fauna.40[sampling.40$SEASON=="Spring",]
fauna.40.fall <- fauna.40[sampling.40$SEASON=="Fall",]

# Remove the absent taxa and apply a Hellinger transformation
fauna.40.spring <- fauna.40.spring[ , colSums(fauna.40.spring)!=0]
dim(fauna.40.spring)     # 13 rows x 19 taxa
fauna.40.spring.hel <- decostand(fauna.40.spring, method="hellinger")
rownames(fauna.40.spring.hel) <- 1996:2008

fauna.40.fall <- fauna.40.fall[ , colSums(fauna.40.fall)!=0]
dim(fauna.40.fall)     # 13 rows x 19 taxa
fauna.40.fall.hel <- decostand(fauna.40.fall, method="hellinger")
rownames(fauna.40.fall.hel) <- 1996:2008

# Compute the LCBD values of site 40 for spring and fall separately
beta.out.spring.40 <- beta.div(fauna.40.spring, method="hellinger", nperm=999)
LCBD.spring.40 <- beta.out.spring.40$LCBD
LCBD.spring.40
beta.out.spring.40$p.LCBD

beta.out.fall.40 <- beta.div(fauna.40.fall, method="hellinger", nperm=999)
LCBD.fall.40 <- beta.out.fall.40$LCBD
LCBD.fall.40
beta.out.fall.40$p.LCBD

# Which LCDB values are significant at the alpha=0.05 level ?
signif.40.spring <- which(beta.out.spring.40$p.LCBD <= 0.05)
signif.40.fall <- which(beta.out.fall.40$p.LCBD <= 0.05)

# Plot LCBD values along the years at site 40, spring and fall; Appendix S3, Figure S3.14
```

```
# Significant LCBD values at the 0.05 level are plotted with a black rim
par(mfrow=c(2,1))

plot(1996:2008, rep(0, 13), type="p", xlab="Sampling years", ylab="", main="LCBD indices
along years, Site 40, spring", ylim=c(-1,1), pch=21, col="white", bg="steelblue2",
cex=20*sqrt(LCBD.spring.40))
points((1996:2008)[signif.40.spring], rep(0, 13)[signif.40.spring], pch=21, col="black",
bg="steelblue2", cex=20*sqrt(beta.out.spring.40$LCBD[signif.40.spring]))

plot(1996:2008, rep(0, 13), type="p", xlab="Sampling years", ylab="", main="LCBD indices
along years, Site 40, fall", ylim=c(-1,1), pch=21, col="white", bg="steelblue2",
cex=20*sqrt(LCBD.fall.40))
points((1996:2008)[signif.40.fall], rep(0, 13)[signif.40.fall], pch=21, col="black",
bg="steelblue2", cex=20*sqrt(beta.out.fall.40$LCBD[signif.40.fall]))
par(mfrow=c(1,1))

# Compute correlations between taxon abundances and LCDB values. Only display the taxa for
# which the absolute values of the correlations are greater than or equal to 0.5 (arbitrarily chosen
# value). Do not perform tests of significance: the LCBD values and taxon abundances are not
# independent of each other.

# Spring
cor.tax.LCBD.spring.40 <- cor(fauna.40.spring.hel, LCBD.spring.40)
as.matrix(cor.tax.LCBD.spring.40[abs(cor.tax.LCBD.spring.40)>=0.5,])

# Fall
cor.tax.LCBD.fall.40 <- cor(fauna.40.fall.hel, LCBD.fall.40)
as.matrix(cor.tax.LCBD.fall.40[abs(cor.tax.LCBD.fall.40)>=0.5,])


# 5.6. REPEAT THE GRAPHICAL SPACE-TIME ANALYSIS FOR TAXONOMIC RICHNESS, 25 SITES

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

# Put the richness data in a matrix with rows = sites and columns = sampling times
rich.mat <- matrix(rich, 25, 26, byrow=TRUE)
rownames(rich.mat) <- rownames(xy.25)
rich.spring.mat <- rich.mat[, seq(from=1, to=25, by=2)]
rich.fall.mat <- rich.mat[ , seq(from=2, to=26, by=2)]
colnames(rich.spring.mat) <- 1996:2008
colnames(rich.fall.mat) <- 1996:2008

rich.per.year.spring <- apply(rich.spring.mat, 2, mean)
rich.per.year.fall <- apply(rich.fall.mat, 2, mean)

# --------
```

```
# Plot the taxonomic richness values along the years
par(mfrow=c(2,1))

plot(1996:2008, rep(0, 13), type="p", xlab="Sampling years", ylab="", main="Richness along
years, Chesapeake, spring", ylim=c(-1,1), pch=21, col="white", bg="steelblue2",
cex=sqrt(rich.per.year.spring))

plot(1996:2008, rep(0, 13), type="p", xlab="Sampling years", ylab="", main=" Richness along
years, Chesapeake, fall", ylim=c(-1,1), pch=21, col="white", bg="steelblue2",
cex=sqrt(rich.per.year.fall))
par(mfrow=c(1,1))

# --------

# Plot space-time maps of richness, spring and fall data; Appendix S3, Figure S3.15

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

# Alternative way: transform the rich.spring.mat and rich.fall.mat matrices (18 lines above) into
# vectors. Take into account the fact that the sites for each year form a column in these matrices
# rich.spring <- as.vector(t(rich.spring.mat))
# rich.fall <- as.vector(t(rich.fall.mat))

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))

# --------

# Is there a significant difference in mean between the spring and fall richness data?

t.test.res <- t.test(rich.spring, rich.fall, paired=TRUE)

# --------

# Relationship between LCBD and richness

# Spring only
cor.test(rich.spring, beta.out.25.spring$LCBD)
```

```
# Fall only
cor.test(rich.fall, beta.out.25.fall$LCBD)

# Note – The correlations between richness and LCDB could be recomputed based upon rarefied
# estimates. On the one hand, use vegan's rarefy() function to obtain rarefied richness estimates
# for a standard sampling effort. On the other hand, use function beta.div() to compute LCBD
# based upon one of Chao et al. (2006) indices, which account for unseen species in the survey
# data. Compute correlations between the resulting richness and LCBD indices for the spring and
# fall surveys separately.

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

REFERENCES NOT FOUND IN THE MAIN PAPER

Chao A, Chazdon RL, Colwell RK, Shen TJ. 2006 Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**, 361–371.

R Core Team. 2013 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

Appendix to:

Legendre P, Gauthier O. 2014 Statistical methods for temporal and space-time analysis of community composition data. *Proc. R. Soc. B* **281**, xxx–xxx.

Appendix S3

Figures from the Practicals in R

This appendix contains the figures produced during the Practical exercises of Appendix S2 that are not included in the paper. In addition, Figure S3.1, referred to in the main paper, is presented in this appendix.

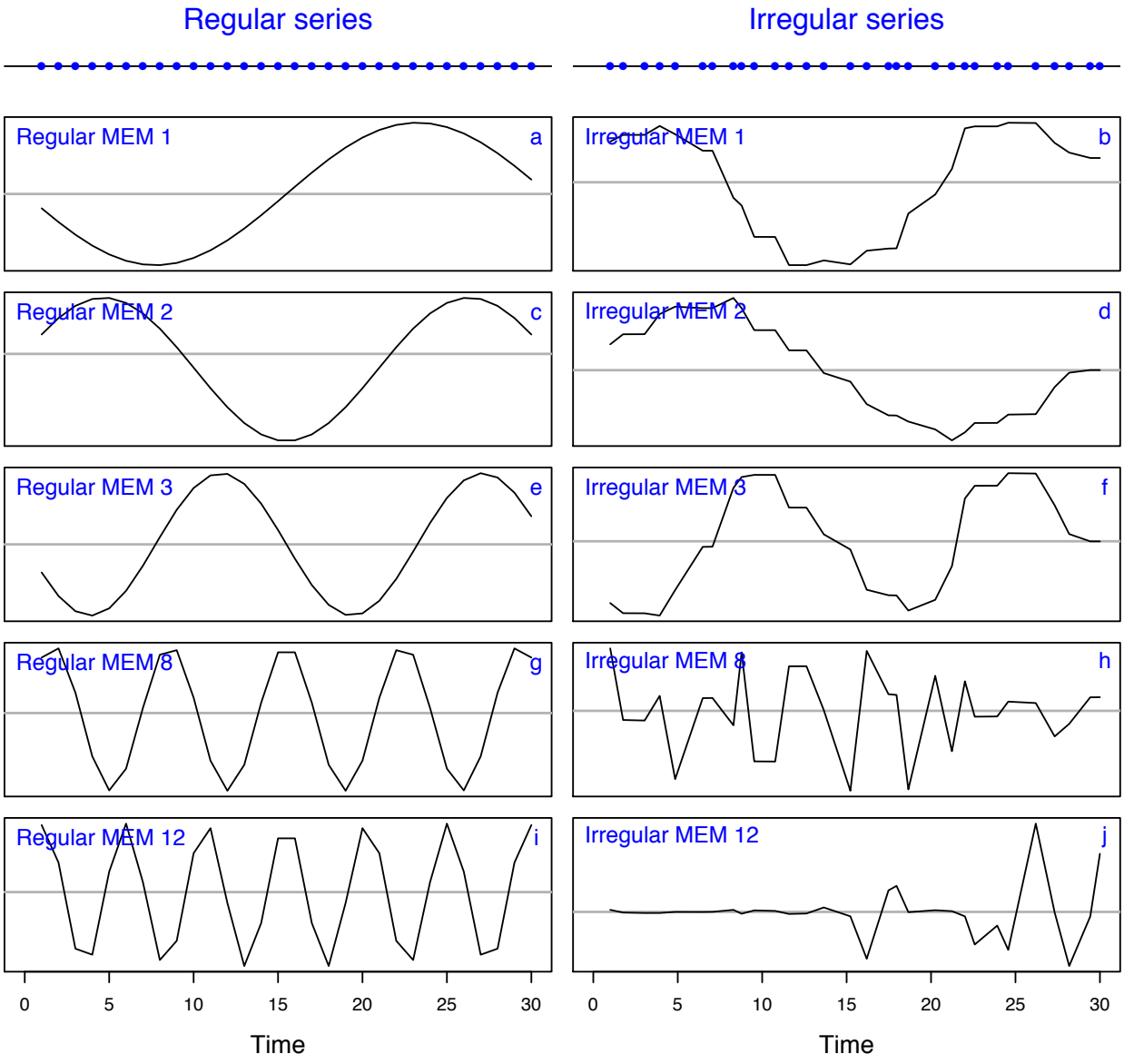

Figure S3.1. A selection of dbMEM eigenfunctions computed for a series of 30 observations. Left: regularly-spaced observations, truncation distance = 1. Right: irregularly-spaced observations, truncation distance = 1.6. In each case, there were 29 eigenfunctions in total, 14 of which modelled positive temporal correlation. As the irregularity of the observations along the series increases, there may be fewer functions modelling positive temporal correlation, and the non-stationary character of the last ones increases. Hence a regular sampling design may produce models with higher R^2 and tests with more power. Top: position of the observations in each series. See Blanchet *et al*. (2011, Fig. E1 and E2) for similar pictures drawn for 100 points.

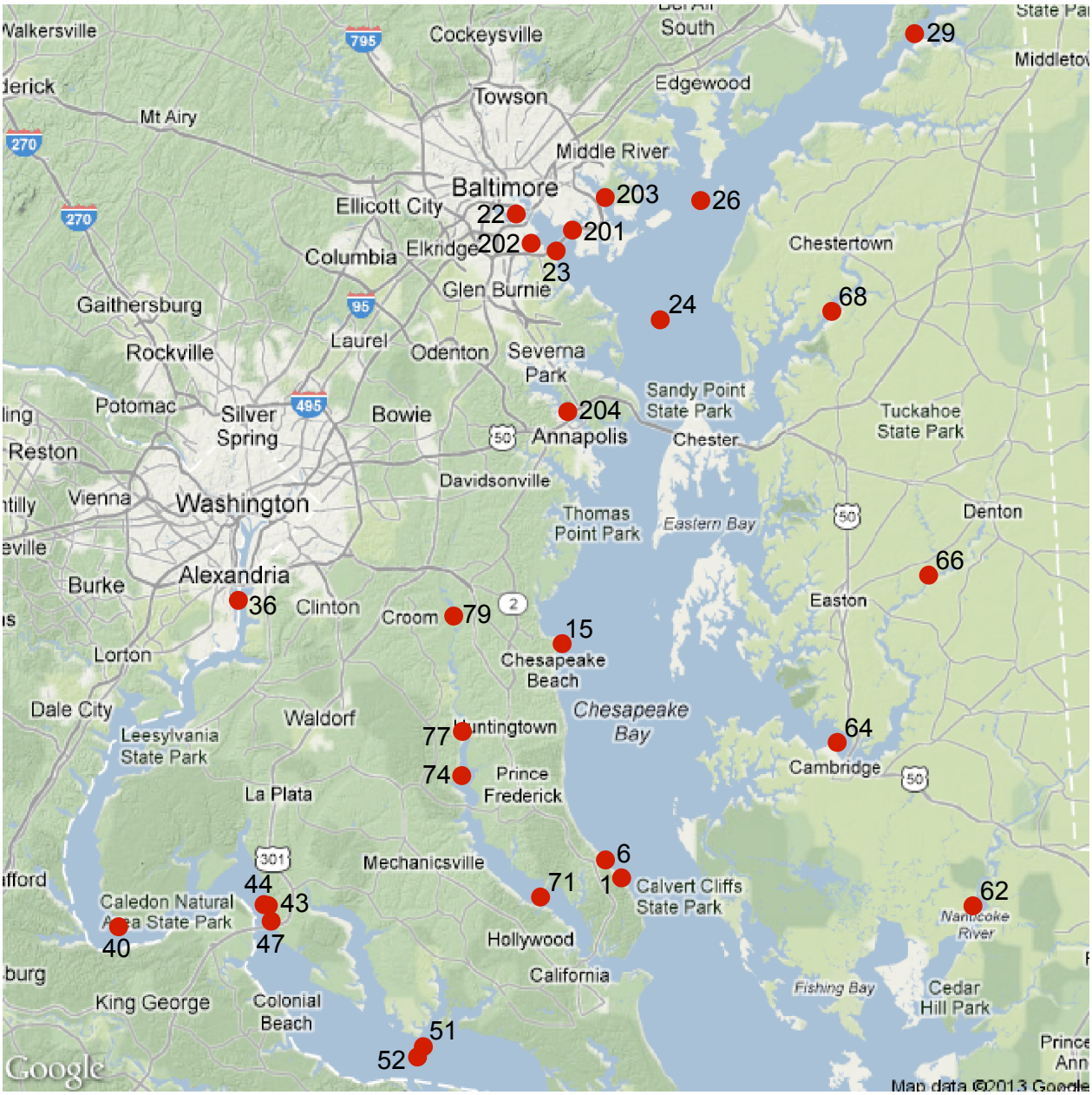

Figure S3.2. Map of the 27 Chesapeake Bay ecological survey fixed sampling sites.

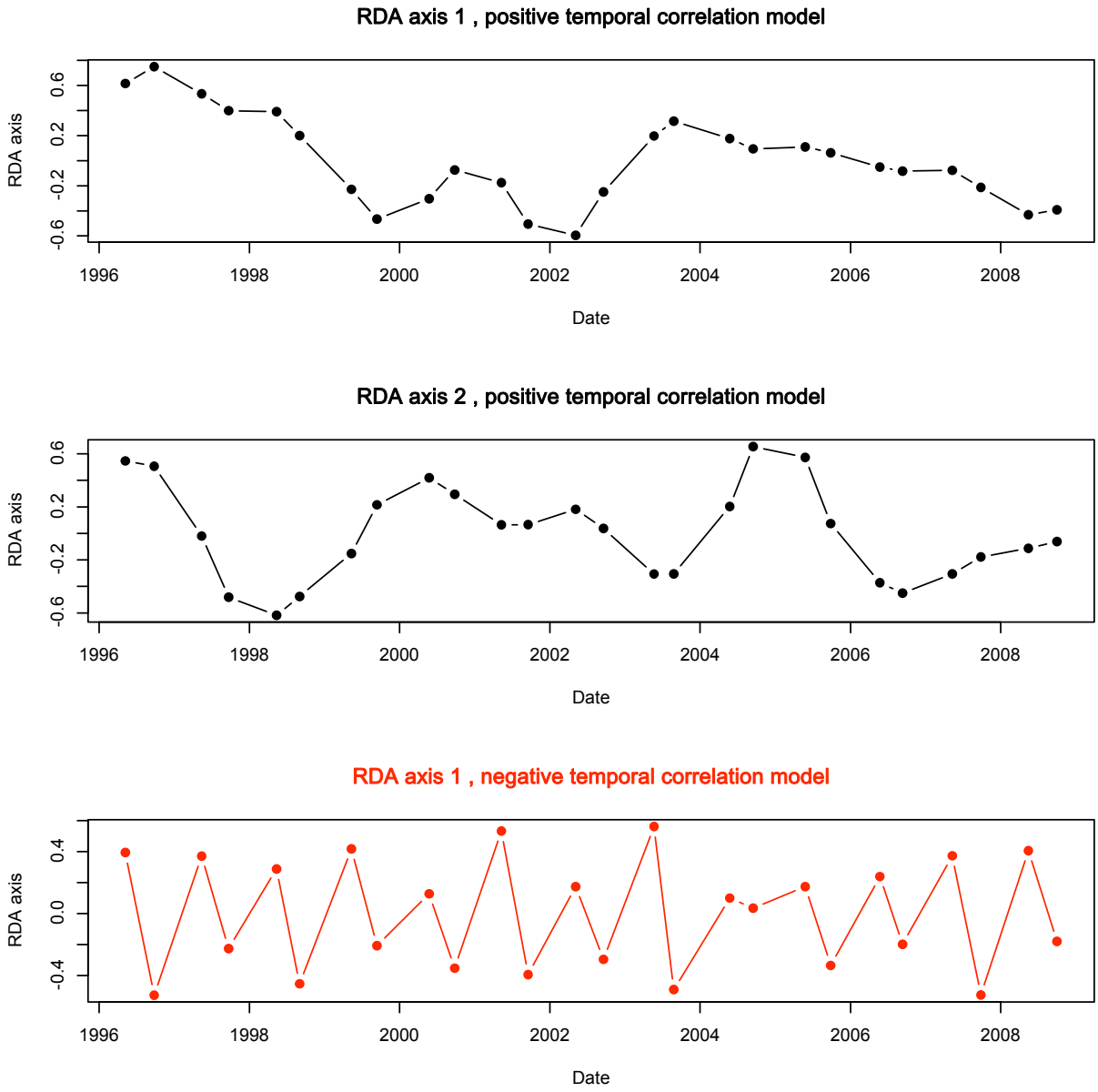

Figure S3.3. Temporal variation of values along the significant ($p < 0.05$) canonical axes of dbMEM models of the site 40 Hellinger-transformed macrofaunal data constructed using all positive (top and middle panels) and negative (bottom panel) dbMEM. Along the abscissa, year labels are shown on January 1st of the year.

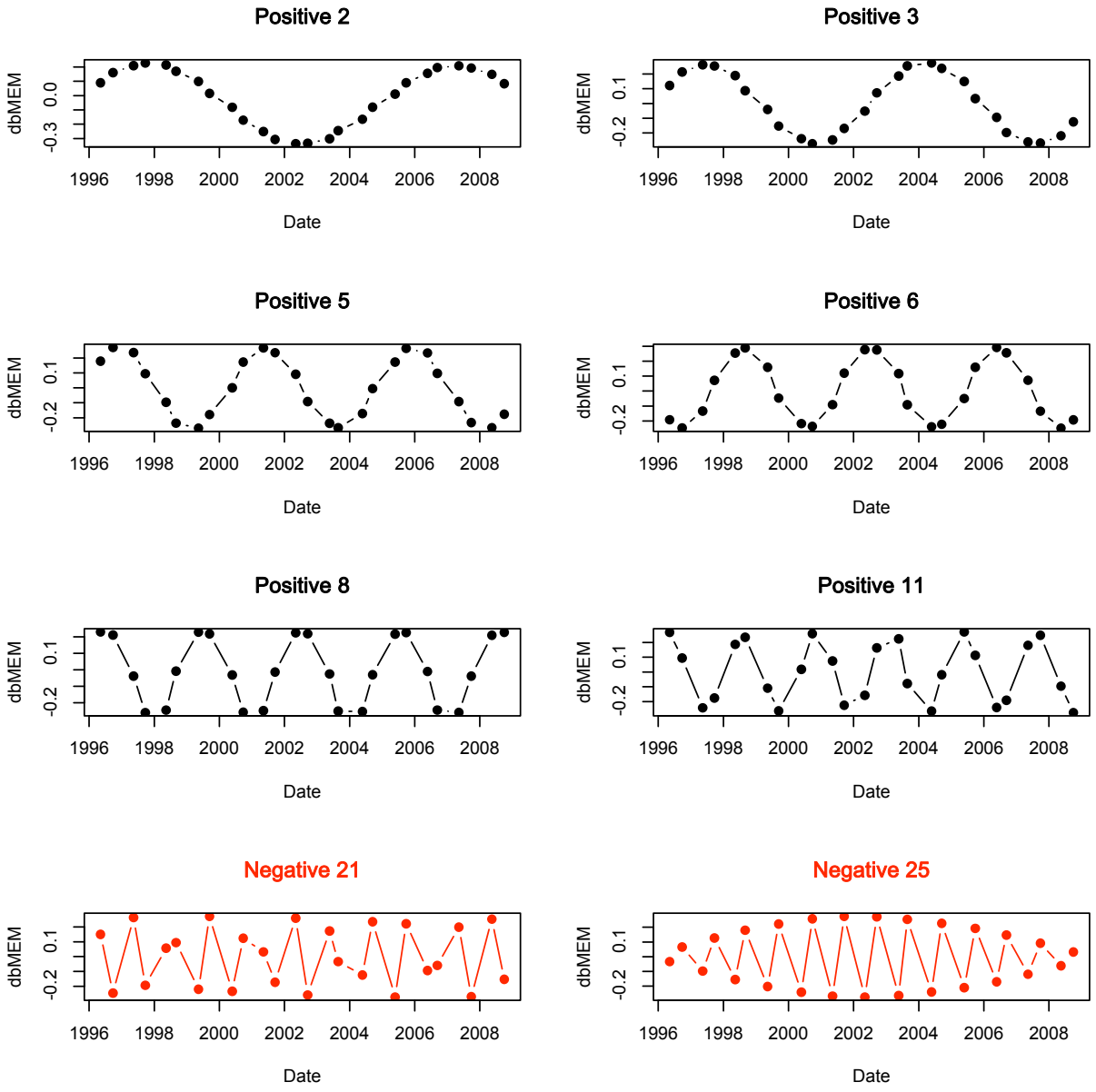

Figure S3.4. The 8 dbMEM eigenfunctions selected by forward selection ($p < 0.05$) from the pool of all available dbMEM for the site 40 data. dbMEM 2, 3, 5, 6, 8 and 11, which model positive temporal correlation, are plotted in black, while dbMEM 21 and 25, which correspond to negative temporal correlation, are plotted in red. Along the abscissa, year labels are shown on January 1st of the year. Note that the signs of the eigenfunctions may be inverted when the calculations are done on different computers or using different software; that is of no consequence.

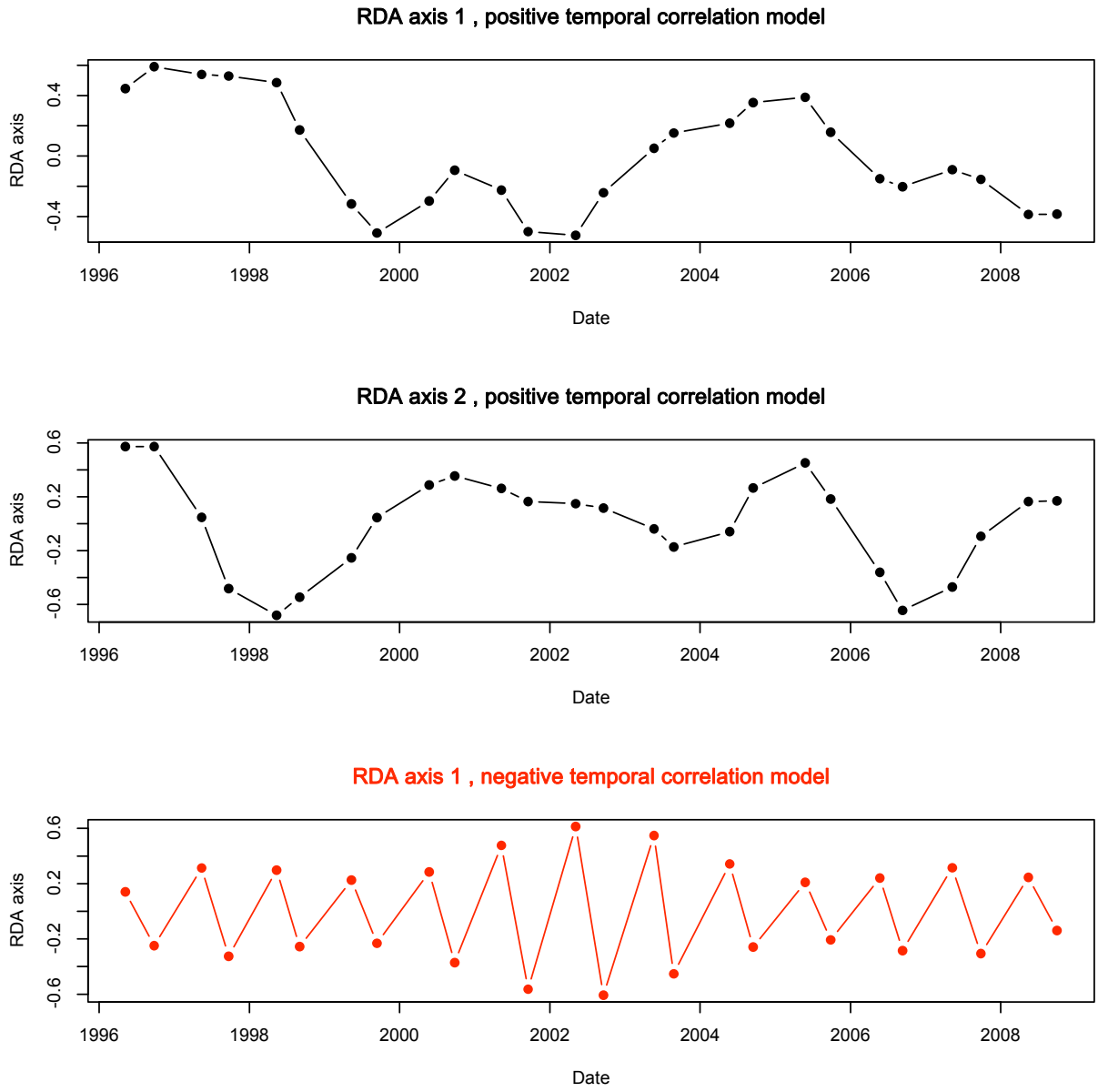

Figure S3.5. Temporal variation of values along the significant ($p < 0.05$) canonical axes of dbMEM models of the Hellinger-transformed site 40 macrofaunal data constructed using forward-selected positive (top and middle panels) and negative (bottom panel) dbMEM. Along the abscissa, year labels are shown on January 1st of the year. dbMEM were selected at the 0.05 level.

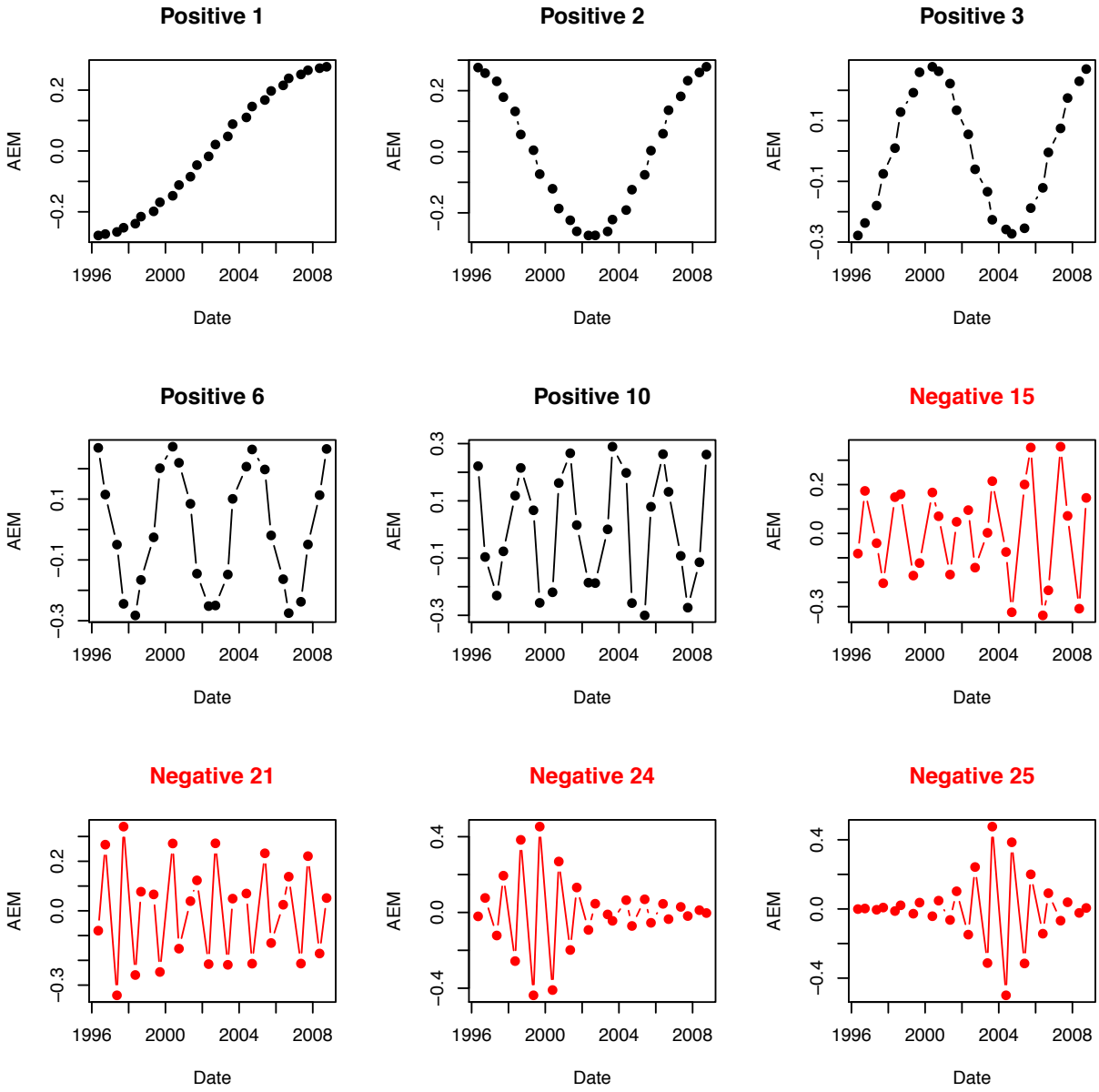

Figure S3.6. The 9 AEM selected by forward selection ($p < 0.08$) from the pool of all available AEM for the site 40 data. AEM 1, 2, 3, 6 and 10, which model positive temporal correlation, are plotted in black, while AEM 15, 21, 24 and 25, which correspond to negative temporal correlation, are plotted in red. Along the abscissa, year labels are shown on January 1st of the year.

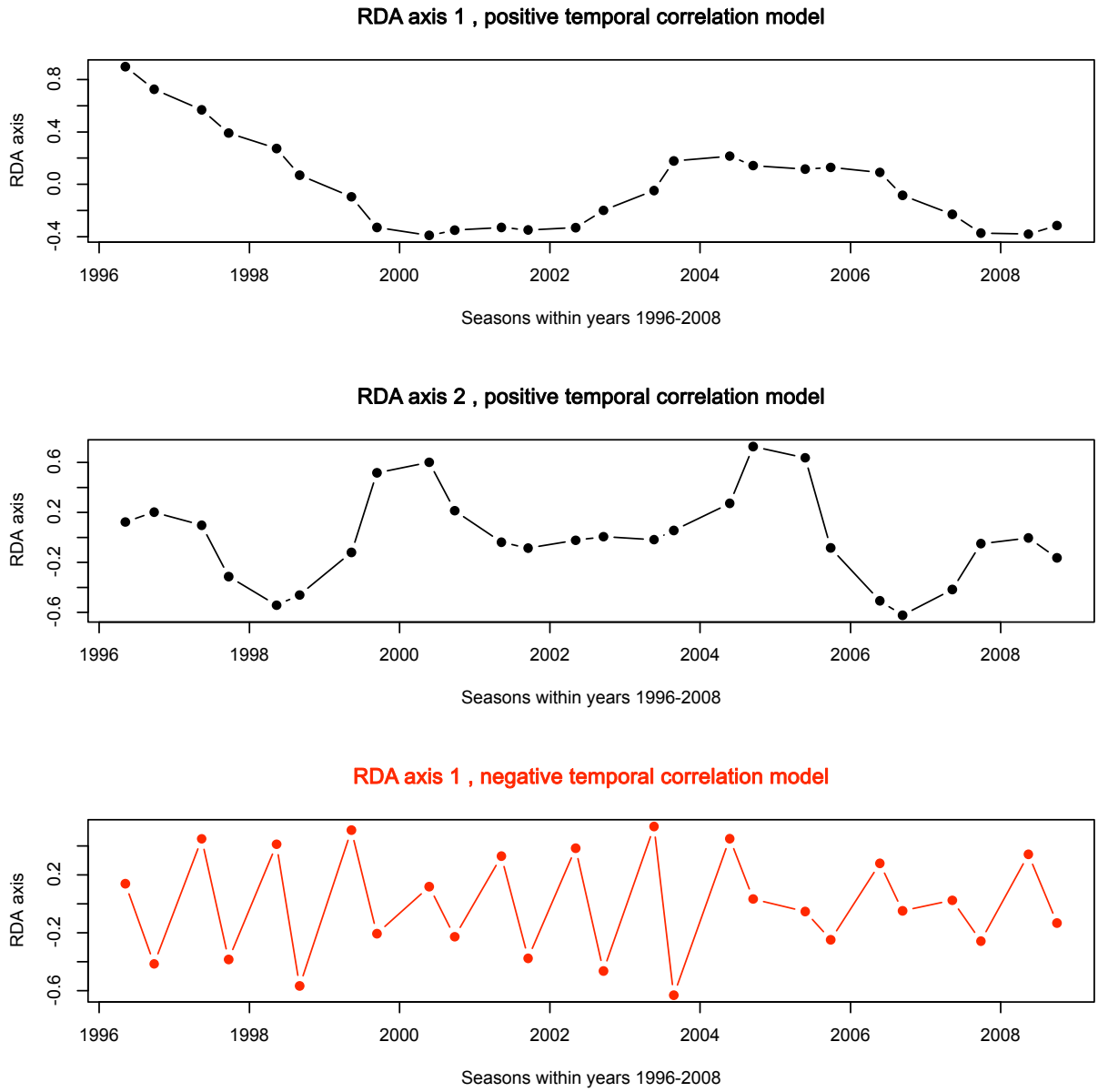

Figure S3.7. Temporal variation of values along the significant ($p < 0.05$) axes of canonical AEM models of the Hellinger-transformed site 40 macrofaunal data constructed using forward-selected positive (top and middle panels) and negative (bottom panel) AEM. Along the abscissa, year labels are shown on January 1st of the year. AEM were selected at the 0.08 level.

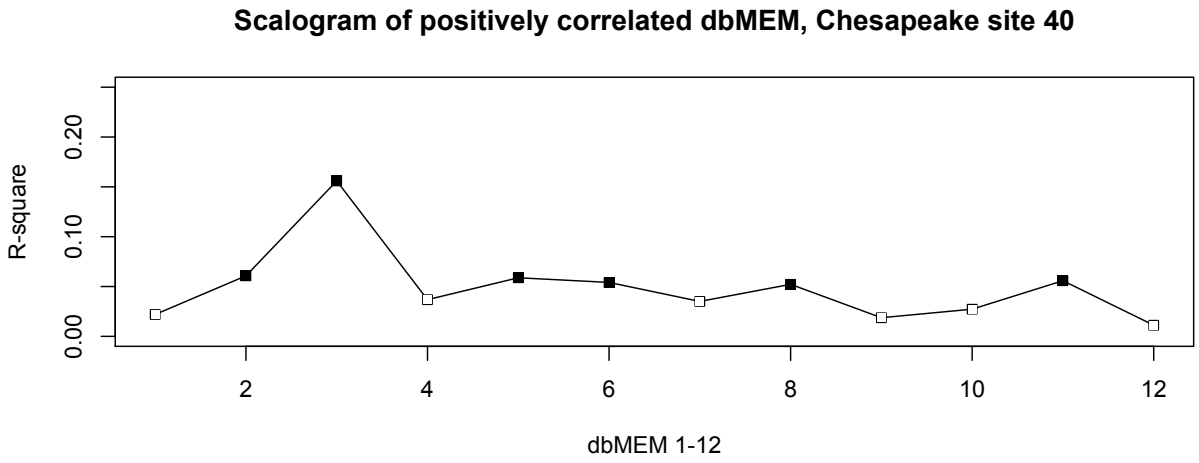

Figure S3.8. Scalogram of the 12 dbMEM modelling positive temporal correlation of the site 40 data. Values on the ordinate are semipartial R^2 computed separately for each dbMEM. The statistics of the dbMEM that were selected by forward selection (i.e. the six that significantly contributed to model the faunal response data; see Figure S3.3) are represented by black squares.

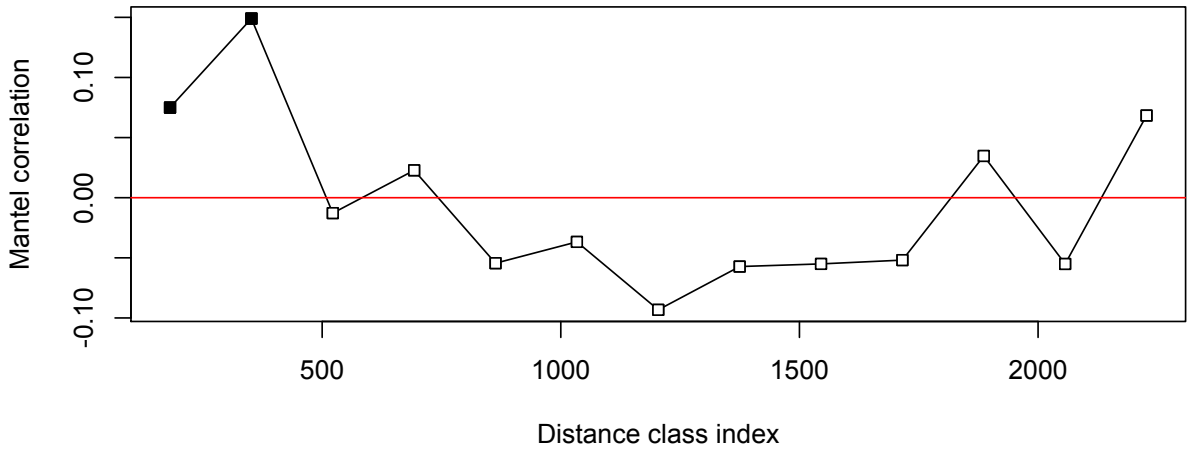

Figure S3.9. Multivariate Mantel correlogram of the site 40 Hellinger-transformed macrofaunal data. Maximum and significant positive correlation is found between observations in the second distance class, which corresponds to pairs of observations one year apart. Observations made in subsequent seasons are also, but somewhat less, positively correlated. Larger distance classes exhibit no significant temporal correlation.

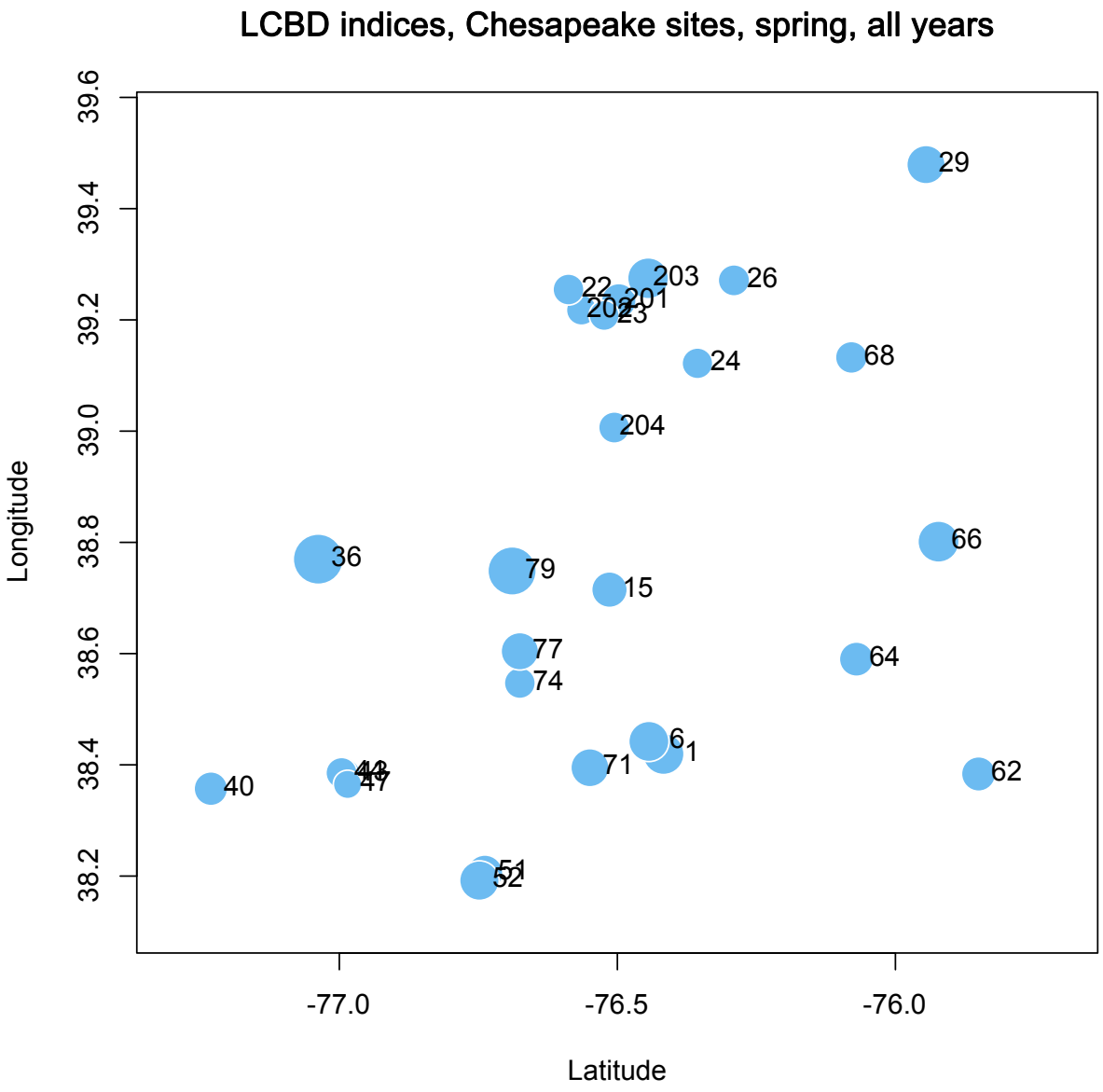

Figure S3.10. Map of LCBD values at the 27 sites, summed over the 13 years, for the spring surveys. The circle surface areas are proportional to the LCBD values.

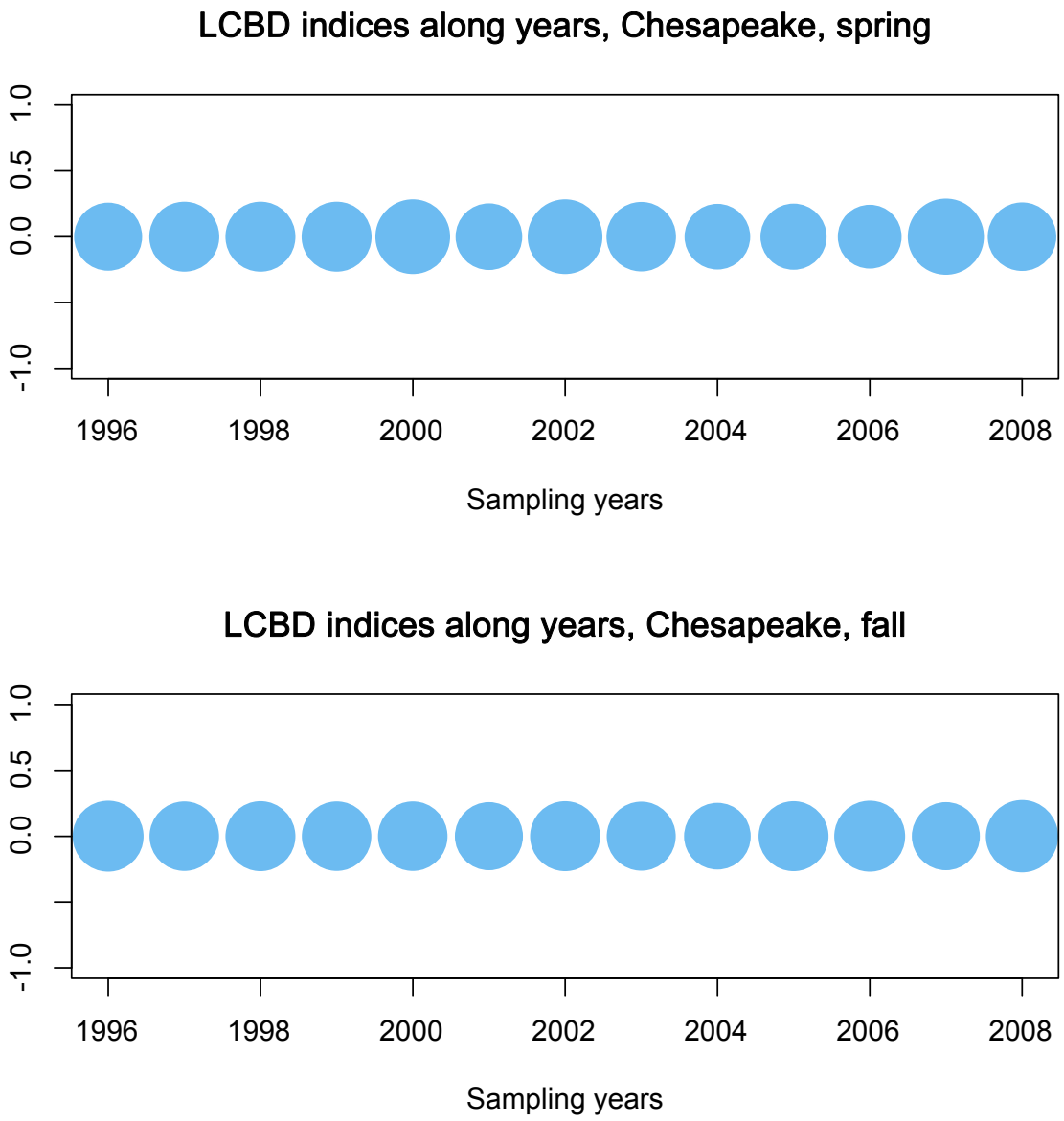

Figure S3.11. Time maps of LCBD values per year, summed over the sites, for the spring (top panel) and fall (bottom panel) surveys.

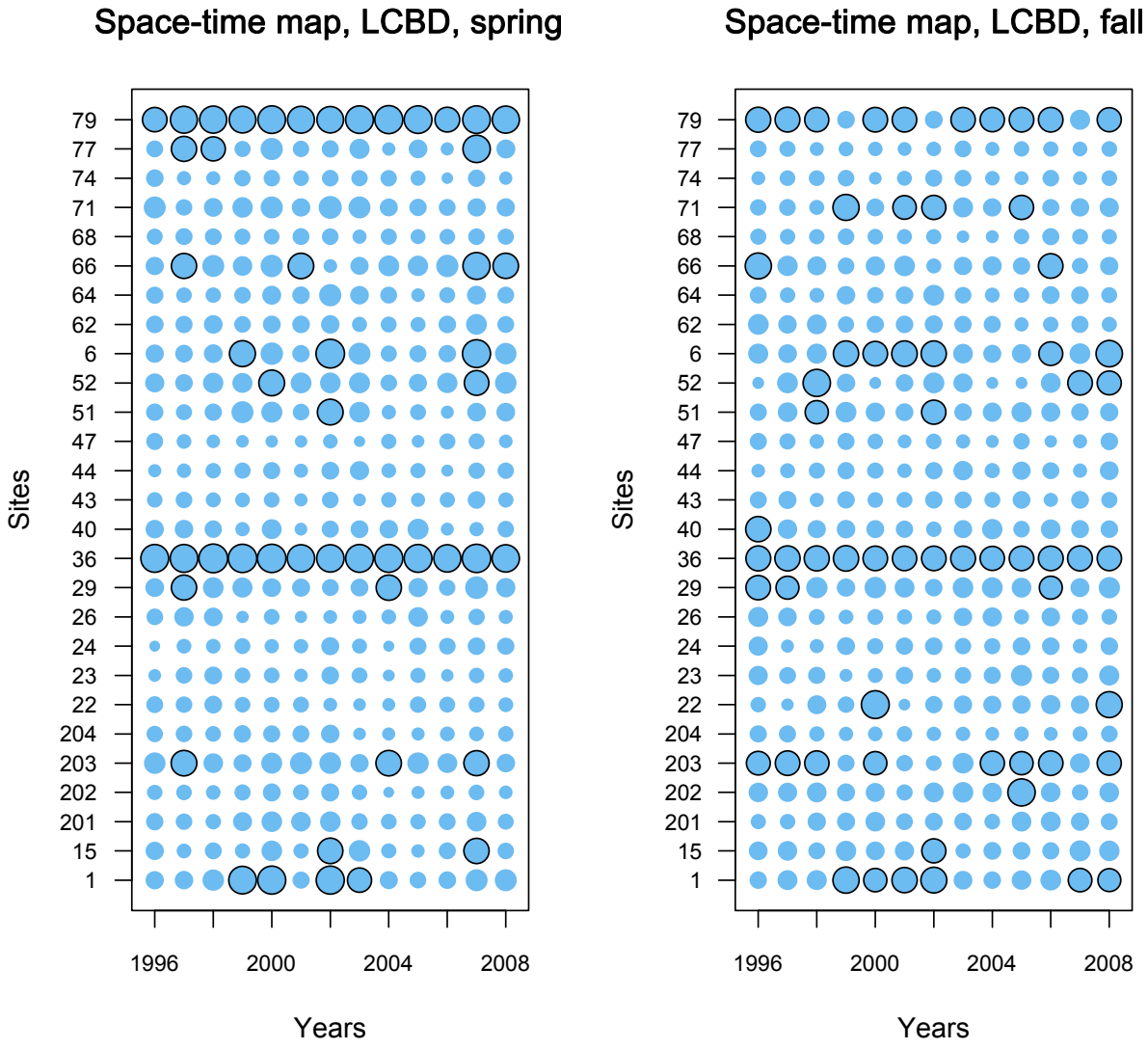

Figure S3.12. Space-time maps (27 sites, 13 years) of LCBD indices computed separately for the spring and fall data. The circle surface areas are proportional to the LCBD values. Circles with a black rim indicate significant LCBD values at the 0.05 level.

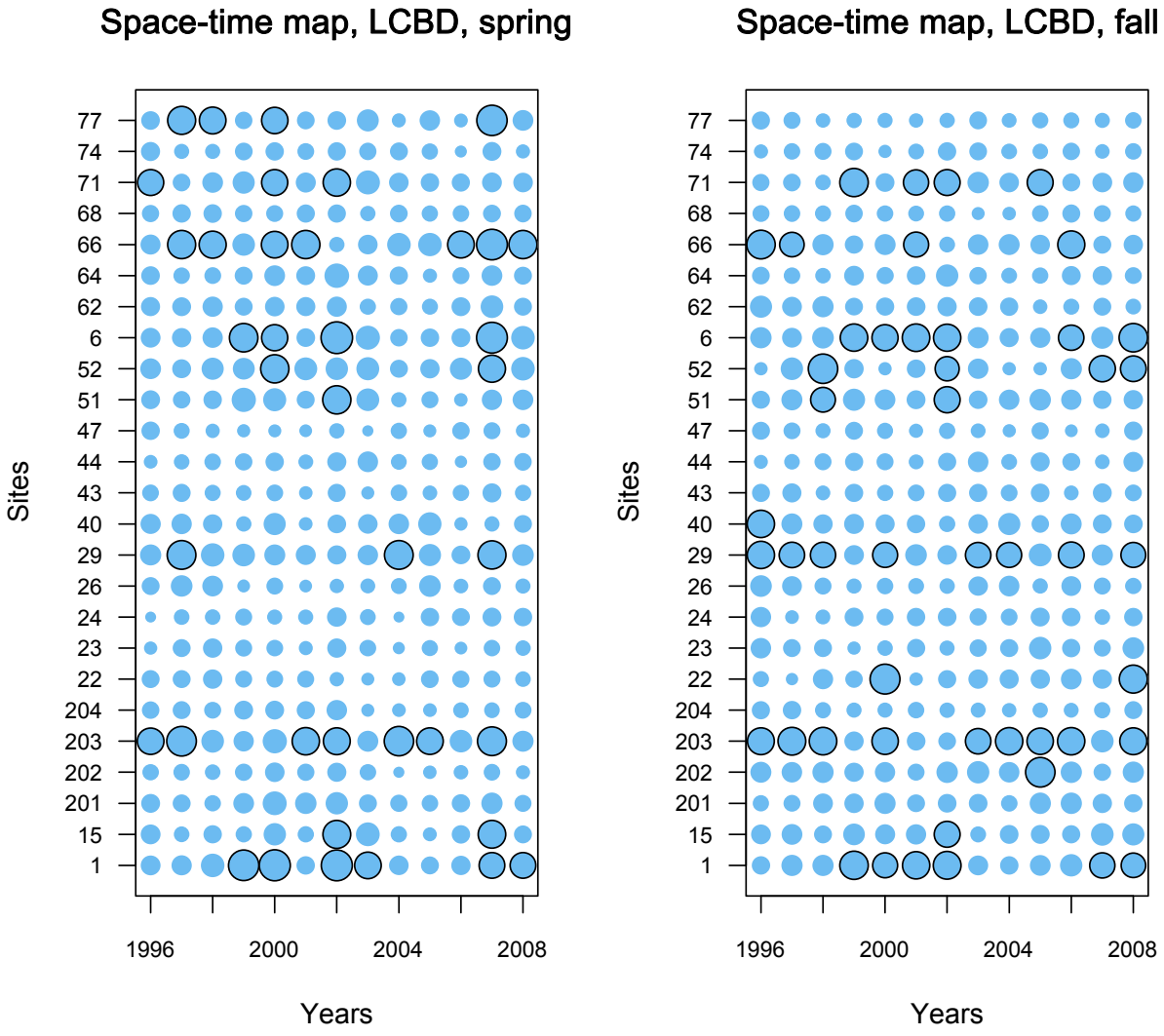

Figure S3.13. Space-time maps (25 sites, 13 years) of LCBD indices computed separately for the spring and fall data. Within each map, the surface area of the circles is proportional to the LCBD. Circles with a black rim indicate significant LCBD values at the 0.05 level.

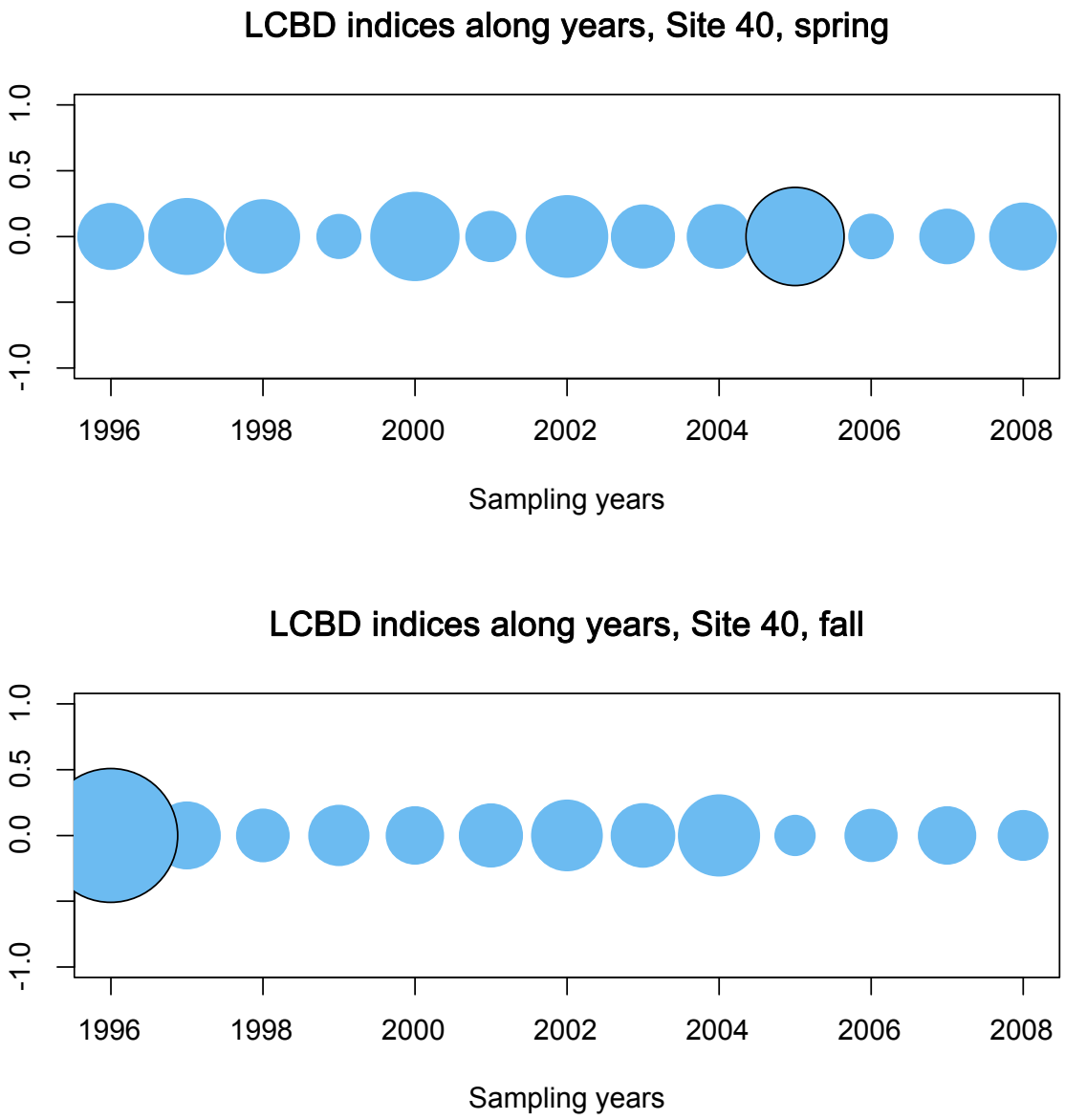

Figure S3.14. Time maps of LCBD values per year for the spring (top panel) and fall (bottom panel) surveys at site 40. Circles with a black rim indicate significant LCBD values at the 0.05 level.

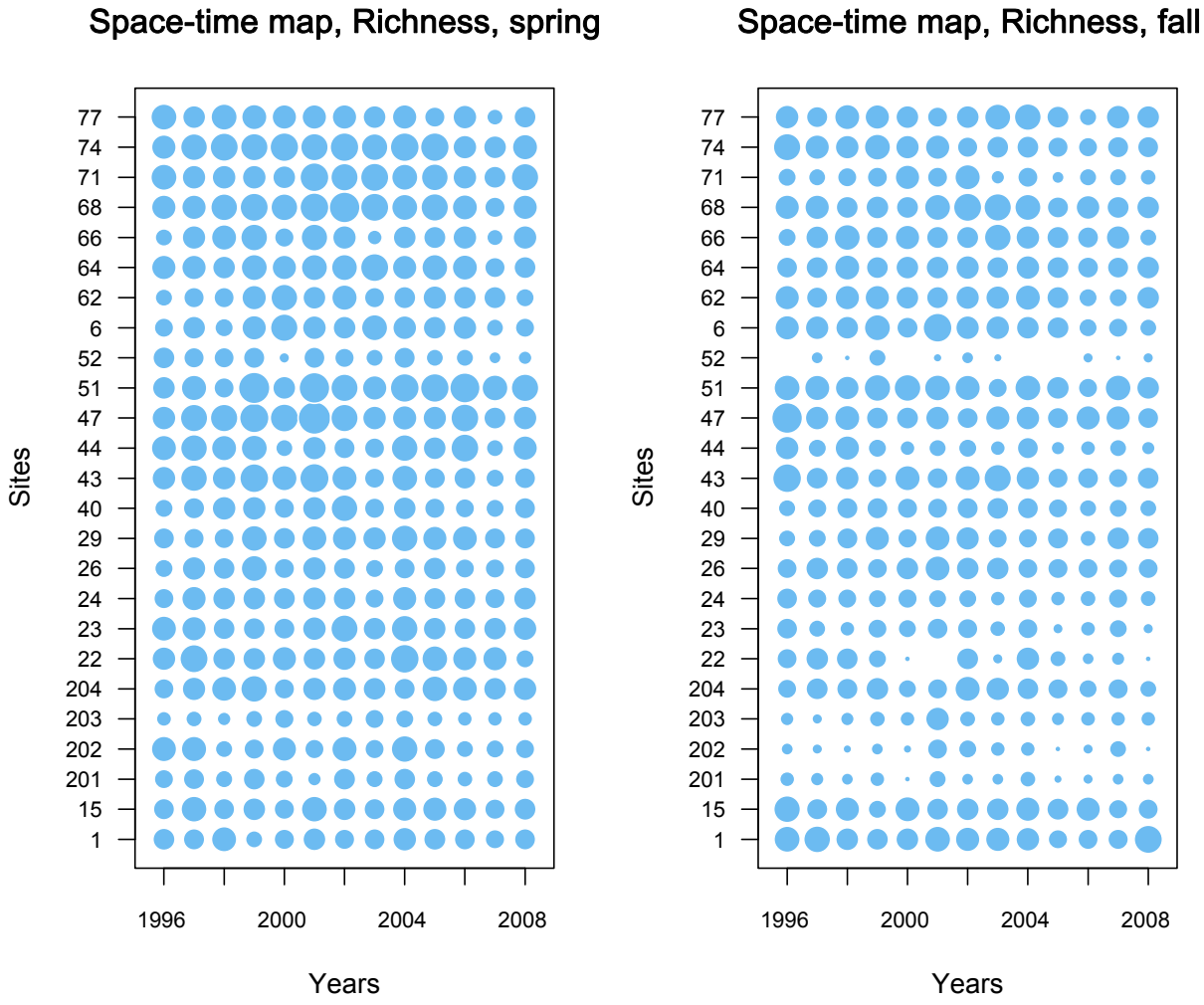

Figure S3.15. Space-time maps (25 sites, 13 years) of taxonomic richness computed separately for the spring and fall surveys.

Appendix to:

Legendre P, Gauthier O. 2014 Statistical methods for temporal and space-time analysis of community composition data. *Proc. R. Soc. B* **281**, xxx–xxx.

Appendix S4

Case study *(continued)*

This is the continuation of the presentation of the Case study of the paper (§8). It was moved to the present appendix for lack of page space.

(c) Contributions of the spatio-temporal sampling units to beta diversity

(i) Graphical analysis of LCBD variation among sites and years

LCBD analysis of the Chesapeake Bay data is detailed in section 5 of the Practicals. LCBD indices were computed for each season separately, based upon the Hellinger distance, over the 27 sites and 13 survey years (Practicals, §5.2). The indices can be summed in any which way that researchers find suitable, while retaining their interpretation. We summed them over years and plotted time maps of LCBD along the years (electronic supplementary material, figure S3.11) for each season separately. The figure shows that the indices did not vary much among the years.

Then we plotted a space-time map of the LCBD values of each site and year, for the two seasons separately (electronic supplementary material, figure S3.12). The size of the circles is proportional to the LCBD values and the circles with a black rim correspond to significant LCBD indices at the 0.05 significance level. The maps show that sites 36 and 79 had very high LCBD indices. These were freshwater sites with mean salinity of 0.2 and 0.3 PSU respectively, so the exceptional character of their faunal composition was not surprising. All the other sites in the study contained brackish water, defined as containing 0.5 to 30 g/L of salt (0.5 to 30 PSU). The map simply confirmed that LCBD indices can successfully point to sampling units with exceptional species compositions.

The analysis was repeated for the 25 brackish sites only, excluding the two freshwater sites (Practicals, §5.3). The new space-time maps (electronic supplementary material, figure S3.13) showed that the year-to-year variation at any one site was less important than the variation among sites. Two-way ANOVA without replication showed a highly significant *site* effect in both the spring and fall. The effect of factor *year* was significant for the spring data, although less so than that of *site*, and it was not significant for the fall data, indicating no consistent differences among years in the fall LCBD values. Nine sites displayed high variability in LCBD indices across years. Two were in the mid-portion of the Bay (sites 1 and 6), the others were in brackish portions of river tributaries (sites 29, 51, 52, 66, 71, 77 and 203). The spring community variation was synchronized across years among these sites (see electronic supplementary material, figure S3.13), hence the significant variation among years identified by ANOVA, but that was not the case for the fall data.

A paired *t*-test of the LCBD indices showed that the spring indices were significantly smaller than the fall values. Before the analysis, the LCBD indices had been computed using all data: 25 sites, spring and fall.

(ii) Analysis of LCBD variation across years using environmental variables

Analyses were carried out to partition the LCBD variation between the *site* factor and the available environmental variables, including the North Atlantic Oscillation (NAO) (Practicals, §5.4). For the spring data, factor *site*, NAO, as well as water quality variables salinity, conductivity and dissolved oxygen, significantly contributed to the explanation of LCBD variation. Factor *site* explained the largest portion of variation (R^2_{adj} = 0.4496) whereas the four environmental variables explained less (R^2_{adj} = 0.0901); a small fraction of variation (R^2_{adj} = 0.0198) was shared between factor *site* and the four environmental variables.

For the fall data, factor *site*, NAO, as well as salinity, conductivity and pH, significantly contributed to the explanation of LCBD variation. Factor *site* explained the largest portion of variation (R^2_{adj} = 0.4314) whereas the four environmental variables explained a smaller yet substantial fraction (R^2_{adj} = 0.2816); a large fraction of variation (R^2_{adj} = 0.2283) was explained jointly by factor *site* and the four environmental variables.

(iii) The role of species

We were interested in identifying the species that corresponded to large and small values of the LCBD indices. For each season separately, correlation coefficients were computed between the LCBD indices and the Hellinger-transformed species data for site 40. That site was chosen for its strong variability among years and because it only had 36 macrofauna taxa, which made results for this site easier to present in a Practical example session. Tables of results are shown in the Practicals, §5.5. The large positive correlations identified the species that contributed more to the observations with large LCBD indices. On the contrary, the large negative correlations indicated the species that were more abundant in the sampling units with small LCBD values; they were common species and among the least ecologically informative. These correlations could not be tested for significance because the LCBD indices were not independent from the species data, from which they were computed.

We used a threshold of $r > 0.5$ for the correlations of interest. For the spring data, species *Chaoborus punctipennis* (Insecta, Chaoboridae), *Coelotanypus spp.* (Insecta, Chironomidae), *Corbicula fluminea* (Bivalvia), *Heteromastus filiformis* (Polychaeta) and *Procladius spp.* (Insecta, Tanypodiae) were strongly and positively correlated with LCBD indices, hence these species were more abundant during exceptional years in the LCBD sense, whereas *Gammarus daiberi* (Crustacea, Gammaridae) was negatively correlated with LCBD, indicating that they were more abundant in non-exceptional years.

For the fall data, species *Cassidinidea ovalis* (Crustacea, Isopoda), *Chaoborus punctipennis* (Insecta, Chaoboridae), *Coelotanypus spp.* (Insecta, Chironomidae) and *Littoridinops tenuipes* (Gastropoda) were strongly and positively correlated with LCBD indices, hence these species were more abundant during exceptional years in the LCBD sense, whereas species *Leptocheirus plumulosus* (Crustacea, Amphipoda) and *Rangia cuneata* (Bivalvia) were negatively correlated with LCBD, indicating that they were more abundant in non-exceptional years.

(iv) Species richness

For comparison, taxonomic richness was also computed for the 25 brackish sites and plotted on space-time maps (electronic supplementary material, figure S3.15). A paired *t*-test showed that the richness values were significantly larger in the spring than in the fall. The correlation coefficients between LCBD indices and taxonomic richness, computed independently for spring and fall, were negative and significant, indicating that in this study, high LCBD values point to species-poor sampling events.

==========