Tutorial: Methods for assessing functional responses to environmental gradients

Kleyer et al.

June 22, 2009

This tutorial presents the different methods used in the paper. A short description of each method is given. We describe the procedures to carry out the analyses using the R language and interpret the results.

Contents

1 Preliminary steps

1.1 Reading the data

The data corresponding to the three tables **R** (sites-by-environmental variables), **L** (sites-by-species) and **Q** (species-by-traits) are available in three text files (TAB-separated). The function `read.table` is used to read the data.

```
> traits <- read.table(file="data/Species_traits.txt", sep="\t")
> spe<- read.table(file="data/Site_species.txt", sep="\t")
> env<-read.table("data/Site_env.txt", sep="\t")
```

1.2 Sourcing the files with R code (functions)

```
> source("scripts/Inference_modelset.r")
> source("scripts/Inference_compute.r")
> source("scripts/corratio.R")
> source("scripts/calinski.R")
> source("scripts/VarScoreOMI.r")
> source("scripts/doublerda.R")
```

1.3 Loading R packages

Several packages must be installed and loaded to perform the analyses presented in this tutorial:

```
> library(ade4)
> library(MASS)
> library(vegan)
> library(ecodist)
> library(maptools)
> library(rpart)
> library(splines)
> library(gam)
> library(pgirmess)
> library(utils)
> library(combinat)
> library(mvpart)
> library(cluster)
> library(fpc)
> library(clusterSim)
> library(lmtest)
> library(Hmisc)
> library(gplots)
```

Other packages are also required by the different scripts associated to this tutorial.

2 Species responses to environmental gradients

As a baseline, we analysed the responses of species to environmental gradients using canonical correspondence analysis (ter Braak, 1986). This analysis was performed using CANOCO for Windows 4.5 in the article and reproduced using the R language in this tutorial. We used the `pcaiv` function of the `ade4` package but other functions are also available in other packages (e.g., `cca` function in package `vegan`).

```
> coa1 <- dudi.coa(spe, scannf = F)
> cca1 <- pcaiv(coa1, env, scannf = F)
```

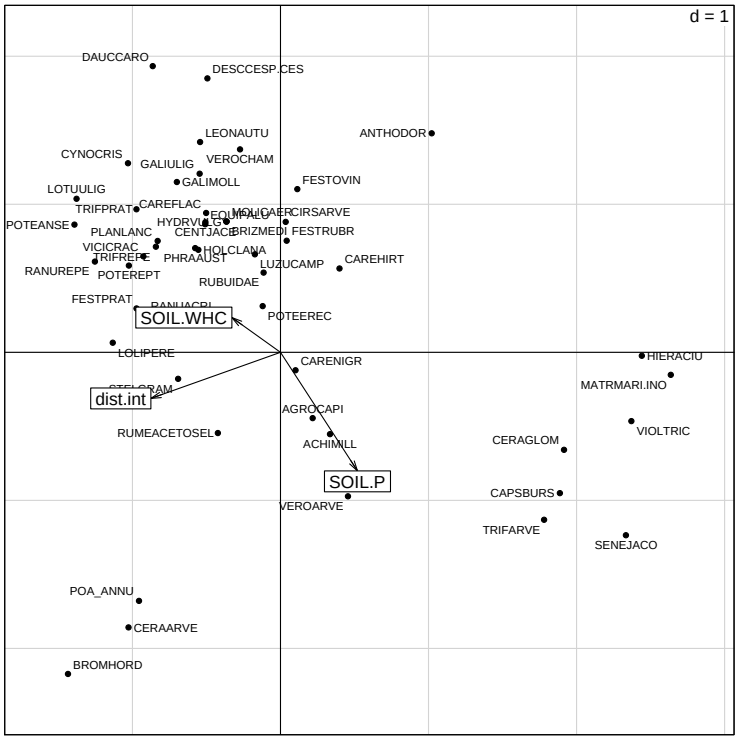
The percentage of variation in species composition explained by the three environmental variables:

```
> 100 * sum(cca1$eig) / sum(coa1$eig)
```

```
[1] 14.89
```

The species responses can be interpreted on the biplot (this corresponds to Figure 2 of the article).

```
> s.label(cca1$c1, clabel = 0)
> par(mar = c(0.1, 0.1, 0.1, 0.1))
> pointLabel(cca1$c1,row.names(cca1$c1), cex=0.7)
> s.arrow(cca1$cor[-1,], add.plot=TRUE)
```



3 A-"CWM-RDA": redundancy analysis of community weighted mean trait responses to environmental gradients

3.1 Description of the method

A sites-by-traits matrix for analysis of trait-environment relationships was created from the original **L** matrix (species-by-sites) and **Q** matrix (species-by-traits). This matrix was weighted by abundance so that each entry in the matrix was the weighted mean of the trait values of all species present in that site (for continuous traits) or the weighted proportion of species with a categorical trait (e.g. polycarpic). Initially the sites-by-traits matrix was subjected to redundancy analysis (RDA, Rao, 1964) constrained by the sites-by-environment matrix, with no transformation of the response variables (i.e. the trait weighted means, 'species' data in the CANOCO program terminology), response variables centred and standardised (no standardisation by samples)and forward selection of environmental variables. In addition, the relationship between the weighted trait data with individual environmental parameters was assessed by repeating the RDA using the other environmental variables as covariables. This analysis was performed using CANOCO for Windows 4.5 in the article and reproduced using the R language in this tutorial.

3.2 Results

The table of weighted means is constructed by matrix multiplication:

```
> cwm.tab <- prop.table(as.matrix(spe),1)%*%as.matrix(scale(traits))
```

The redundancy analyis is then perfomed using the `pcaiv` function of the `ade4` package. Other functions are also available in other packages (e.g., `rda` function in package `vegan`). Contrary to the paper, No forward selection is performed in the tutorial.

```
> pca.cwm <- dudi.pca(cwm.tab,scannf=FALSE)
> rda.cwm <- pcaiv(pca.cwm,env, scannf=FALSE)
```

The percentage of variation in community traits explained by the three environmental variables:

```
> 100 * sum(rda.cwm$eig) / sum(pca.cwm$eig)
```

```
[1] 32.08
```

The percentages of explained variation associated to each axis:

```
> pca.cwm <- dudi.pca(cwm.tab,scannf=FALSE)
> 100 * rda.cwm$eig / sum(rda.cwm$eig)
```

```
[1] 62.515 35.434  2.052
```

The relationships between the trait and the environmental variables (this corresponds to Figure 3a of the paper)

```
> s.arrow(rda.cwm$c1, xlim=c(-1,1), boxes = FALSE)
> s.label(rda.cwm$cor[-1,], add.plot=T, clab=1.5)
```


d = 0.5
dist.int
seed.mass.log
Polycarpic
SOIL.WHC
SLA
height
Cnratio
Onset.flower
SOIL.P

4 B-"CLUS-MOD": modelling functional groups on environmental gradients

4.1 Description of the method

CLUS-MOD firstly builds species groups from their traits (component 3), then searches for the trait combination with the best response to the environmental variables (component 1 and 2).

- **Step 1:** Component 3 (grouping; see Table 2 in the main text)

  In order to group species according to their traits, we applied Ward's hierarchical clustering (Everitt et al., 2001). Clustering was repeatedly conducted both based on single traits and based on all possible combinations of single traits (i.e. combinations of two to six traits; 63 clusterings in total). The cophenetic correlation (Legendre and Legendre, 1998) was used to assess how closely the clustering results correspond to the original resemblance matrix. It represents the correlation between the phenetic distances (pair-wise distances across the dendrogram) with the pair-wise distances in the distance matrix (Sneath and Sokal, 1973). Maximizing this correlation, will ensure that branch lengths in the dendrogram best match the biological differences measured among the organisms (Petchey and Gaston, 2006). For each of the 63 combinations, the optimal number of groups (clusters) was determined via Calinski and Harabasz's index (Gordon, 1999). Group stability we assessed by bootstrapping (Hennig, 2007). To this end, many (500) bootstrap samples (drawing with replacement) of the data are clustered, and the species of the resulting groups are compared to those of the original data by calculating the Jaccard index. The higher the average Jaccard index of the bootstrap replications, the more stable the group.

- **Step 2:** Component 1 (responses of clusters to environmental variables) and B (identification of responsive traits through iteration of A)

  In the second step, we modelled group responses to the environmental variables. Since the frequency data are strictly bounded to values between 0 and 100, we used logistic regression (Agresti, 2002). For each group, univariate models were estimated to determine the shape of the relationship (monotonic or unimodal) to the environmental factor. Based on all significant variables, multiple models were built. For multiple models, all possible combinations of parameters were tested. For the data used in this work (three environmental variables) this led to a maximum of seven models if all three parameters were significant (three univariate models, three models with two variables, one model with three variables). For models with more than one parameter, LR-tests were performed to test if each variable significantly improved the model. All significant univariate and multiple models were then subjected to model averaging, leading to one averaged model for each group (Burnham and Anderson, 2002; Strauss and Biedermann, 2006). Model averaging avoids the often spurious choice of a single best model and the pitfalls of stepwise variable selection (Whittingham et al., 2006; Mac Nally, 2000).

- **Step 3:** Component 2 (identification of responsive traits through iteration of A)

  Each clustering could be rated for the quality of the clustering and the responsiveness of its groups to the environment. This process does not necessarily imply one best solution, but typically a limited number of good clusterings that lead to (often very similar) groups exhibiting strong relationships to the environment. We used the following criteria to rate the clustering of each trait combination: (i) the cophenetic correlation coefficient; (ii) the mean jaccard index indicating group (=cluster) stability; (iii) the mean R^2 of the group models, indicating the responsiveness to the environment across all groups; (iv) the minimum R^2 to ensure that each group had a minimum goodness of fit. For these criteria, the following thresholds were used: (i) cophenetic correlation coefficient > 0.7, (ii) mean jaccard index of each group > 0.7, (iii) mean $R^2 \geq 0.3$, (iv) minimum $r^2 \geq 2$. The number of groups per trait combination was taken as an additional criterion. Clusterings with just two groups can sometimes lead to models with high goodness of fit. In this case, within each group, the response is mainly driven by the species with high abundance, the majority of species has only marginal influence. In the case of nominal traits, these groups are also highly stable. However, more groups can allow species with lower abundance to be more influential and thus to yield a more subtle picture of the trait – environment relationships. From the trait combinations meeting all of the above criteria, we selected the one yielding the highest number of groups.

4.2 Results

The analysis can be perfomed by sourcing the five scripts given in the folder `scripts`. Several output files are created to perform the complete analysis. At the start of the analysis, some parameters must be edited:

- `max.traits` is the maximum numbers of traits considered at one time. If set to a value exceeding number of traits present in the data, it is automatically reduced to that value.
- `min.cophenetic.corr` is the minimum cophenetic correlation coefficient. A reasonable value is 0.7 - 0.8. Set to 0 if all combinations are to be retained in the output.
- `max.no.groups` is the maximum number of groups within one clustering. A reasonable setting depends on the number of species present.
- `noisecut.value` represents the minimum number of species within a group
- `min.prop` is the minimum proportion of species that have to be in stable clusters. Set to 0 if all combinations are to be retained in the output.
- `no.boot` indicates the bootstrap replications for bootstrapping cluster stability. Should be at least 200, for serious analyses. Even with 500 replications, results are not completely replicable (deviations of 0.03 in `mean.jac` still possible).

- `mean.jac` is the minimum mean jaccard index for bootstrapping. Groups below this value are not considered to be stable and will not be considered for modelling (step 2). Reasonable setting: 0.65-0.75. In case of very homogeneous trait value distributions, lower values might be necessary. Set to 0 if all combinations are to be retained in the output.

- `name.output.table` is the name of the .txt file where outputs are written. More output is produced to file "cluster.output.txt". Do not delete this file, it is required for the following steps.

- `traits.to.consider` refers to the trait to consider in the clustering procedure. There are 3 options on how to set this. To consider all traits in the table, use `traits.to.consider<-"all"`. To consider only traits in the respective columns, use `traits.to.consider<-c(2,3,4)`. To select traits by their names, use `traits.to.consider<-c("Polycarpic","Cnratio","SLA","height")`

```
> max.traits<-6
> min.cophenetic.corr<-0.7
> max.no.groups<-10
> noisecut.value<-5
> min.prop<-0.9
> no.boot<-200
> mean.jac<-0.7
> name.output.table<-"result.cluster.boot.txt"
> traits.to.consider<-"all"
```

Correlations between traits can be investigated. For instance, SLA is negatively correlated with canopy height and onset of flowering. Clusterings based on these traits may be similar in species composition.

```
> cor(traits)
```

```
              Polycarpic  Cnratio seed.mass.log      SLA   height
Polycarpic       1.00000  0.17289     0.0824449 -0.28643  0.26878
Cnratio          0.17289  1.00000     0.0844576 -0.30082  0.30188
seed.mass.log    0.08244  0.08446     1.0000000  0.04623 -0.01168
SLA             -0.28643 -0.30082     0.0462325  1.00000 -0.50164
height           0.26878  0.30188    -0.0116774 -0.50164  1.00000
Onset.flower     0.28795  0.25905    -0.0001324 -0.51958  0.34466
              Onset.flower
Polycarpic       0.2879485
Cnratio          0.2590469
seed.mass.log   -0.0001324
SLA             -0.5195817
height           0.3446553
Onset.flower     1.0000000
```

Start the clustering procedures with bootstrapping. Print the trait combinations that result in stable clusters with a cophenetic correlation coefficient higher than `min.cophenetic.corr` and a Jaccard Index higher than `mean.jac`. The output shown here is truncated by the head function.

```
> source("scripts/script1_cluster_analysis_report.r")
```

```
> clus.boot<-read.table("result.cluster.boot.txt",sep="\t",header=TRUE)
> head(clus.boot[,c(1:6,13:14)])
```

```
  no.combi no.traits                          involved.traits
1       12         4 Polycarpic, Cnratio, SLA, Onset.flower
2       14         4 Polycarpic, seed.mass.log, SLA, height
3       17         4  Polycarpic, SLA, height, Onset.flower
4       24         3               Polycarpic, Cnratio, SLA
5       25         3            Polycarpic, Cnratio, height
6       26         3      Polycarpic, Cnratio, Onset.flower
```

```
  cophenetic.corr no.clusters clust.stable spec.per.clust
1            0.74           3            3        21,18,6
2            0.78           4            4       19,15,8,7
3            0.80           3            3        28,10,12
4            0.82           3            3         27,15,8
5            0.74           4            4       22,12,8,8
6            0.83           3            3         25,14,6
  mean.jacc.clust.stable
1          0.83,0.8,0.76
2    0.71,0.78,0.95,0.75
3          0.89,0.87,0.79
4          0.84,0.75,0.97
5    0.84,0.77,0.98,0.76
6          0.92,0.91,0.94
```

In the second step, group responses to environmental variables are modelled. Output from Script 1 (i.e. component 3) is required (read in from file "cluster.output.txt"). There are some parameters to be set:

- `max.var` denotes the maximum number of variables used in multiple models. Reasonable values depend on sample size.
- `r2.min` is the minimum r^2 for variables in univariate models. Significant variables below this threshold will not be considered for multiple models. Set to 0 if all significant variables are to be retained.
- `min.cum.cov`: Group has to reach this cumulate frequency or coverage [percent, 0-100] at least in 1 plot so that models will be estimated
- `min.prev` is the minimum prevalence of a group (=proportion of plots where group occurs [0,1]) so that models will be estimated

```
> max.var<-3
> r2.min<-0
> min.cum.cov<-0.5
> min.prev<-0.1
```

Thereafter, script 2 - Group models must be invoked. It calculates models for functional groups (group frequencies or coverages depending on environmental parameters).

```
> source("scripts/script2_group_models_report.r")
```

Script 3 writes the results of the clustering and modelling scripts into a readable format. Two output files are written in .txt format, they can be read easily in Excel. The files created by Scripts 1 and 2 are required, so do not delete them.

Names of output files: "outputfile.1" contains output for each individual functional group. "outputfile.2" contains summarized output for each trait combination.

```
> name.outputfile.1<-"modelling.output.groupwise.txt"
> name.outputfile.2<-"modelling.output.clusterwise.txt"
> source("scripts/script3_output_tables_report.r")
> mod.groupwise<-read.table(name.outputfile.1,sep="\t",header=TRUE)
> mod.clusterwise<-read.table(name.outputfile.2,sep="\t",header=TRUE)
> head(mod.clusterwise[order(-mod.clusterwise$no.clusters,-mod.clusterwise$r2.av),c(1:9,17)])
```

```
   no.combi no.traits                        involved.traits
5        25         3            Polycarpic, Cnratio, height
20       63         1                           Onset.flower
2        14         4 Polycarpic, seed.mass.log, SLA, height
4        24         3              Polycarpic, Cnratio, SLA
15       44         2             Polycarpic, seed.mass.log
```

```
6        26         3      Polycarpic, Cnratio, Onset.flower
   cophenetic.corr r2.av r2.min                 r2.all no.clusters
5             0.74  0.33   0.23 0.23, 0.27, 0.49, 0.32           4
20            0.88  0.30   0.12  0.3, 0.28, 0.12, 0.49           4
2             0.78  0.28   0.16 0.16, 0.19, 0.49, 0.29           4
4             0.82  0.39   0.23       0.23, 0.44, 0.49           3
15            0.80  0.39   0.14       0.14, 0.49, 0.55           3
6             0.83  0.38   0.22       0.22, 0.43, 0.49           3
   clust.stable mean.jacc.clust.stable
5             4    0.84,0.77,0.98,0.76
20            4         0.99,1,0.97,1
2             4    0.71,0.78,0.95,0.75
4             3         0.84,0.75,0.97
15            3         0.84,0.97,0.83
6             3         0.92,0.91,0.94
```

```
> head(mod.groupwise[,1:3])
```

```
  no.combi                              trait.combi group.no
1       12 Polycarpic, Cnratio, SLA, Onset.flower        1
2       12 Polycarpic, Cnratio, SLA, Onset.flower        2
3       12 Polycarpic, Cnratio, SLA, Onset.flower        3
4       14 Polycarpic, seed.mass.log, SLA, height        1
5       14 Polycarpic, seed.mass.log, SLA, height        2
6       14 Polycarpic, seed.mass.log, SLA, height        3
```

The file "modelling.output.clusterwise.txt" lists all trait combinations with stable clusters, their cophonetic correlation coefficient and Rsquare (r^2). The output shown here is truncated by the head function and shows only selected columns. The table can be sorted according to `r2.av` which is the average r^2 of all clusters of a certain trait combination.

The second file "modelling.output.groupwise.txt" gives information regarding the trait ranges of each cluster, the regression parameters and the weights of the parameters. Again, the output is truncated by the head function and shows only selected columns. Six out of 63 trait combinations passed all thresholds: 1. (Polycarpic, CNratio, canopy height); 2. (Polycarpic, CNratio, onset); 3. (Polycarpic, CNratio, SLA); 4. (Polycarpic, SLA, height); 5. (Polycarpic, SLA, height, onset); 6. (Polycarpic, CNratio). The first combination was considered to match the selection criteria in the best way.

The selected trait combinations should be identified by their combination number (`no.combi`) and retained for plotting the results.

For each combination, boxplots are produced for the distribution of trait values within each functional group. Group models are plotted with respect to each environmental parameter (while all other environmental paramters are held constant at their median values). If a group does not respond to a parameter, it is not plotted. Note that model averaging can lead to coefficients close to 0, resulting in very shallow slopes and thus almost straight lines for some parameters. Here we show the response curves for trait combination no. 25 "Polycarpic, Cnratio, canopy height". This was the only combination yielding 4 groups, with a cophenetic correlation coefficient > 0.7, a mean jaccard index of each group > 0.7, a mean $R^2 \geq 0.3$, and minimum $r^2 \geq 2$. The third group of this combination, small monocarps, increased with soil P while showing no response to grazing and a negative response to soil water content (see also Fig. 5, main text). The other groups comprised only polycarpic species. The first group consisted of small polycarpic perennials with low CN-ratio. This group showed a positive response to grazing (dist.int) and a unimodal response to soil water content (soil.WHC). The second group, comprising small polycarps with high CN-ratio, decreased with soil P and grazing. The fourth group, comprising

large polycarpic species, decreased with soil P, increased with soil water and showed a unimodal response to grazing.

To plot results, the following parameters have to be set or may be set:

- `combis.to.plot` is used to identify the combinations of traits to be plotted. This could be a vector of mode numeric which contains the combination numbers. To plot all combinations, use

  `combis.to.plot<-as.vector(output.table.2[,"no.combi"])`.

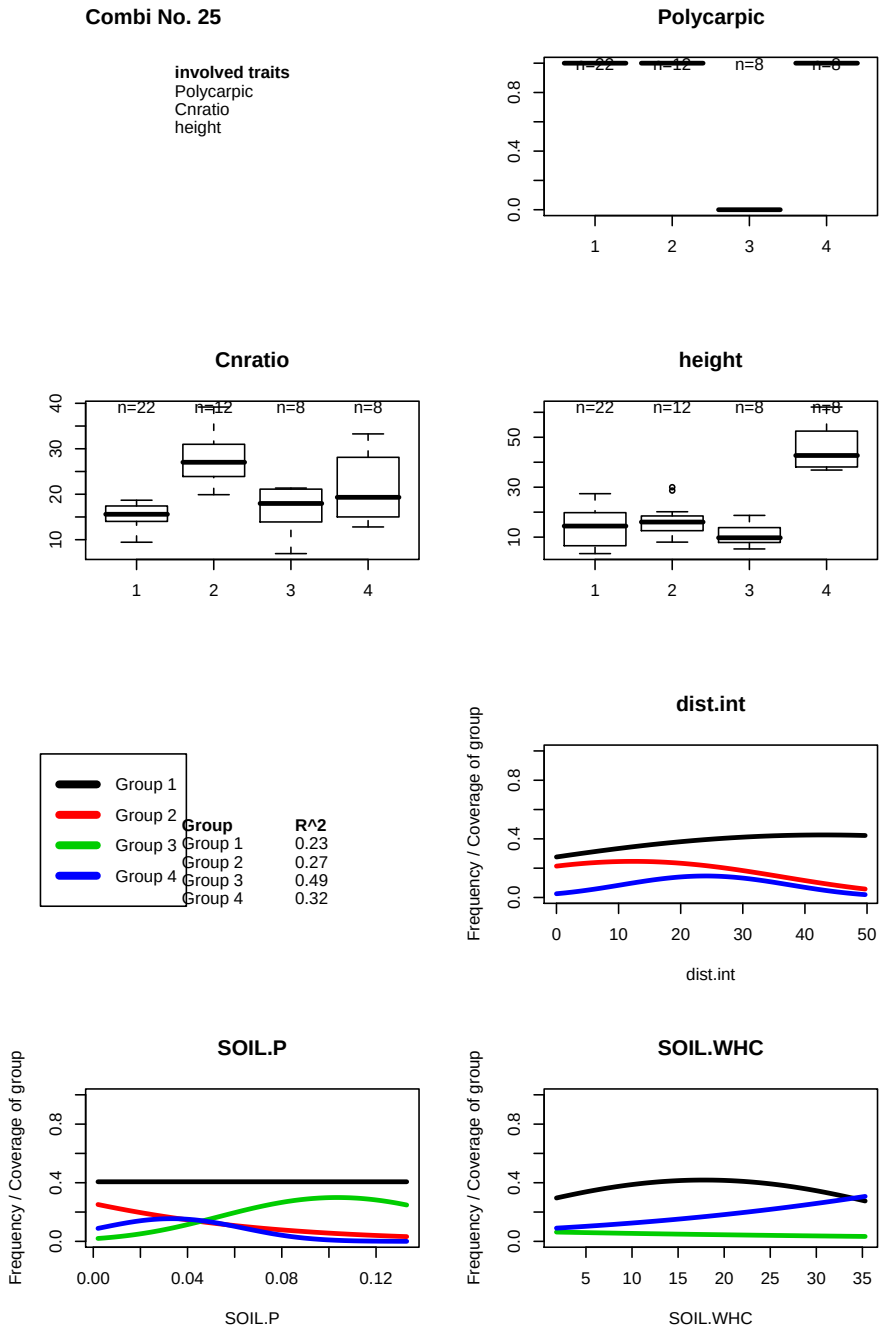
- `name.pdf` is the name of output plot file.

- `min.weight` is the minimum weight [0,100] of a variable in a group model so that it is plotted. Small weights mean small coefficients and shallow slopes, thus not much to see but a straight line.

- `mycol` denotes colours for output.

- `plot.points` is either 0 or 1. If the original data are also to be plotted, set`plot.points<-1`. Plots get easily overloaded, use with caution!

```
> combis.to.plot<-c(25)
> min.weight<-10
> mycol<-palette()
> plot.points<-0
> name.table<-"spec.groups.txt"
> source("scripts/script4_group_plots_report.r")
```



The following section produces a table where, for selected combinations, group assignements for each species and combination can be compared. A .txt file is produced that can be read. Species marked `NA` in a certain combination were either not in a cluster at all or not in a stable cluster. Note that for combinations that lead to almost identical groups, comparable group do not necessarily have the same number (even though they often do)! The

combinations of traits to be compared are identified by `combis.to.compare<-c(13,26,43,25,32)`. The figures in brackets denote the combination number. To compare all combinations with stable clusters, use `combis.to.compare<-as.vector(output.table.2[,"no.combi"])`. `name.table` is the name of the table with the species and their assignment to the functional groups.

```
> name.table<-"spec.groups.txt"                #name of output file
> #combis.to.compare<-as.vector(output.table.2[,"no.combi"])
> combis.to.compare<-c(25,24,44,26,12,17)
> source("scripts/script5_group_assignements_report.r")
> spec.in.groups<-read.table("spec.groups.txt",sep="\t",header=TRUE)
> head(spec.in.groups)
```

```
  combis.to.compare X25 X24 X44 X26 X12 X17
1      no.of.groups   4   3   3   3   3   3
2          ACHIMILL   1   1   1   1   1   1
3          AGROCAPI   1   1   1   1   1   1
4          ANTHODOR   1   1   1   1   1   1
5          BRIZMEDI   2   2   1   2   2   1
6          BROMHORD   3   3   2  NA  NA   2
```

Comparing the six combinations revealed that the trait ”polycarpic” was always involved and yielded a coarse separation into annual and perennial species. However, when ”onset of flowering” came into play, a few early flowering polycarpic species were assigned to the monocarpic group. C:N ratio separated the perennials into two groups with either low or high C:N ratios. However, this only applied to low growing perennials. Large plants were indifferent in terms of C:N ratio. Thus, the combination of life cycle, C:N ratio and canopy height led to more distinct groups. Even though SLA was negatively correlated to C:N ratio, it did not discriminate among the low growing perennials as clearly as C:N ratio. Functional groups resulting from the trait seed mass were either instable and / or showed weak relationships to the environment.

5 C-"RDA-sRTA" and D-"RDA-mRTA": redundancy analysis and regression tree

5.1 Description of the method

RDA-RTA is a two step procedure. In the first step (component A), the response of each individual species to each individual environmental gradient (component A) is calculated using the redundancy analysis (RDA). Then, the response is predicted by the species traits, using the regression tree method (component B). Component C, the grouping of species based on responsive traits is a direct outcome of component B. RDA-single RTA (RDA-sRTA) predicts the response to each environmental gradient separately, whereas RDA-multi RTA (RDA-mRTA) uses multivariate regression tree and predicts the response to both the gradients simultaneously.

- **Step 1:** Component A (species responses to environmental variables)

  The first step, determination of the species response to individual gradients was used for RDA-sRTA only for the environmental variables, which were selected as significant in a forward selection procedure (i.e. grazing intensity and soil phosphorus). Consequently, two separate RDAs were calculated, for grazing intensity and for soil phosphorus as explanatory (environmental) variable; the other variable was used as a covariable. We have used the RDA on the correlation matrix (i.e. the option center and standardize by species), no standardization by samples was applied. The species scores correspond to the species correlation with the environmental axis, and can be considered a species response. In the CANOCO implementation, the polarity of axes is arbitrary – consequently, we have always changed the polarity so that the positive values signify positive response to the gradient (positive correlation with the environmental variable). As grazing intensity and soil phosphorus were nearly uncorrelated (see Appendix S1), the effect of using a covariable (i.e. the distinction between marginal and partial effect of the variable) is negligible. In cases with correlated environmental variables, the distinction between marginal and partial effects might be much more pronounced and can change the ecological interpretation considerably.

- **Step 2:** Component B (identification of responsive trait combinations)

  Regression trees were used to select the traits predicting individual responses. The regression tree is a non-parametric regression that produces a binary tree built through binary recursive partitioning. In our case, the traits were predictors and the species response (i.e. RDA species scores) the predicted value. The trait that best distinguishes species' responses splits the species into two groups; then, within each subset, another trait splits the species further. We have used the pruned regression tree obtained by reducing a fully grown regression tree, with the extent of the reduction based on the cross-validation procedure. In the univariate regression tree (for RDA-sRTA), response to each environmental gradient was predicted separately (procedure `rpart` of the package `rpart`). For the RDA-mRTA, the multivariate regression tree, predicting the species responses to the two gradients simultaneously (procedure `mvpart` of the

package `mvpart`). Component C (forming groups) is a direct outcome of the Component B. Nevertheless, it should be noted that the main goal of the regression tree analysis is the prediction – and the tree (and consequently groups) are formed just as a mean to achieve the main goal.

This analysis was performed using CANOCO for Windows 4.5 in the article and reproduced using the R language in this tutorial. The forward selection is not presented in the tutorial.

5.2 Results for C-"RDA-sRTA"

Partial RDAs were perfomed using the `rda` function of the `vegan` package:

```
> rda.phosp <- rda(X=spe, Y=env$SOIL.P, Z=env$dist.int, scale = TRUE)
> rda.dist <- rda(X=spe, Y=env$dist.int, Z=env$SOIL.P, scale = TRUE)
> rda.both <- rda(X=spe, Y=env[,c("dist.int","SOIL.P")], scale = TRUE)
```

Each gradient explained roughly the same amount of variation in species composition:

```
> 100 * rda.phosp$CCA$tot.chi/rda.phosp$tot.chi
```

```
[1] 5.499
```

```
> 100 * rda.dist$CCA$tot.chi/rda.dist$tot.chi
```

```
[1] 5.576
```

Whereas the values of the environmental parameters are uncorrelated, the species response to them is not independent – species responding positively to P tend to respond negatively to disturbance and vice versa – the correlation between the species responses to P and Disturbance is -0.4.

```
> cor(env$SOIL.P,env$dist.int)
```

```
[1] -0.03306
```

```
> cor(rda.phosp$CCA$v,rda.dist$CCA$v)
```

```
        RDA1
RDA1 -0.3956
```

The correlation between RDA axis and phosphorous (contrary to the paper, axes are not reversed in the tutorial):

```
> rda.phosp$CCA$biplot
```

```
        RDA1
[1,] -0.9995
```

Then, regression trees are constructed with traits as explanatory variables and RDA species scores as response variables. Note that both packages `rpart` and `mvpart` have a `rpart` function. Thus, we specify by the `::` operator that we use the function of `rpart` package. We add the constraint that each leaf of the tree must contain at least 3 species:

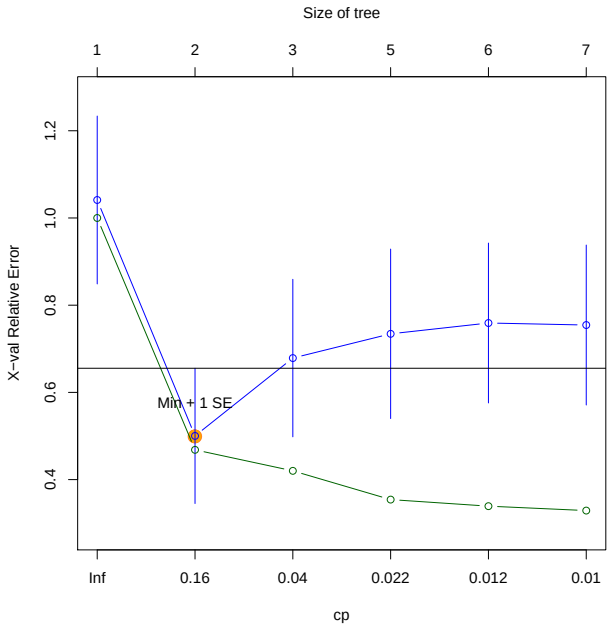
```
> df.phosp <- cbind(RDA1=rda.phosp$CCA$v[,1],traits)
> df.dist <- cbind(RDA1=rda.dist$CCA$v[,1],traits)
> rta.dist<-rpart::rpart(RDA1~.,data=df.dist, xval = 100, minbucket = 3)
> rta.phosp<-rpart::rpart(RDA1~.,data=df.phosp, xval = 100, minbucket = 3)
```

With respect to phosphorous, the regression tree selected the monocarpic/polycarpic as the first and best predictor. Seed weight and onset of flowering were then suggested to improve the fit within polycarpic plants:

```
> rta.phosp


n= 50

node), split, n, deviance, yval
      * denotes terminal node

  1) root 50 0.97400  0.02278
    2) Polycarpic< 0.5 8 0.03345 -0.21040 *
    3) Polycarpic>=0.5 42 0.42270  0.06720
      6) seed.mass.log< -0.7935 5 0.09702 -0.02374 *
      7) seed.mass.log>=-0.7935 37 0.27870  0.07949
       14) Onset.flower< 162.5 7 0.04459  0.02671 *
       15) Onset.flower>=162.5 30 0.21010  0.09180
         30) height>=29.32 6 0.11060  0.01863 *
         31) height< 29.32 24 0.05930  0.11010
           62) seed.mass.log< -0.1525 16 0.02913  0.09262
            124) height>=10.48 9 0.00823  0.07076 *
            125) height< 10.48 7 0.01107  0.12070 *
           63) seed.mass.log>=-0.1525 8 0.01551  0.14510 *
```

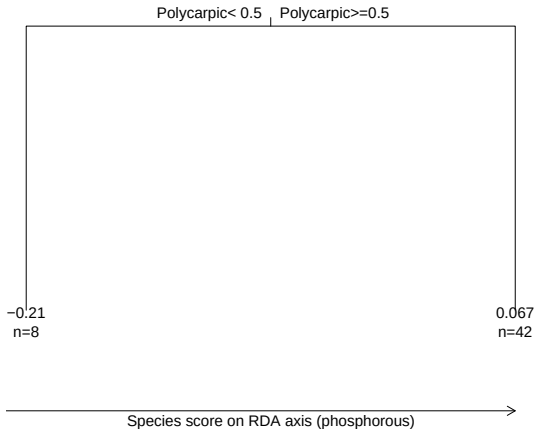
To prune the trees, we used cross validation:

```
> plotcp(rta.phosp)
```



This suggest that the best tree has a size of 2 (i.e. Polycarpic vs monocarpic). The new pruned tree is then:

```
> idx.min <- which.min(rta.phosp$cptable[,4])
> rta.phosp.pruned <- prune(rta.phosp, cp = 1e-6+rta.dist$cptable[,1][idx.min])
> plot(rta.phosp.pruned, ylim = c(0.28,1.1))
> text(rta.phosp.pruned, use.n=T)
> arrows(0.9, 0.28, 2, 0.28, length = 0.1)
> text(1.5, 0.26, "Species score on RDA axis (phosphorous)", cex = 1)
```



This analysis yielded two functional groups:

```
> table(rta.phosp.pruned$where)
```

```
 2  3
 8 42
```

Group 1 comprised monocarpic species which occurred at high P level. Group 2 was polycarpic. The inclusion of seed weight was not sufficiently supported by crossvalidation. Whereas the Mono/Polycarpy explains itself 53.17 % of variability, the tree with seed mass explains 57.99 %.

```
> 100 * (1-rta.phosp$cptable[2,3])
```

```
[1] 53.17
```

```
> 100 * (1-rta.phosp$cptable[3,3])
```

```
[1] 57.99
```

Concerning grazing, the correlation with RDA axis (contrary to the paper, axes are not reversed in the tutorial):

```
> rda.dist$CCA$biplot
```

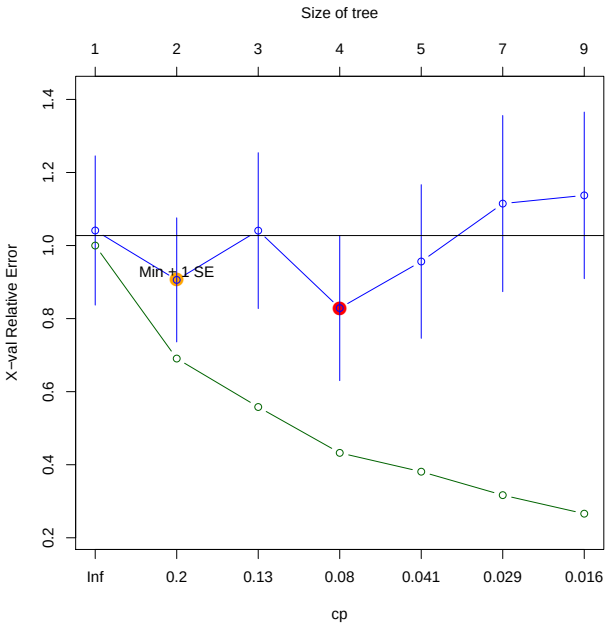
```
        RDA1
[1,] -0.9995
```

The regression tree selected C:N ratio as the best splitting rule (the high C>N ratio plants respond negatively to disturbance). Then, the best predictors differ. In low C:N plants, the seed mass is important (surprisingly, the heavier seeds predict more positive response to disturbance). For the high C:N plants, a second split based on Polycarpy is also selected:

```
> rta.dist


n= 50

node), split, n, deviance, yval
      * denotes terminal node

 1) root 50 0.980800 -0.019570
   2) Cnratio< 18.19 28 0.391400 -0.088620
     4) seed.mass.log>=-0.843 22 0.234900 -0.123300
       8) Onset.flower< 168.5 15 0.115100 -0.156000
        16) seed.mass.log< -0.479 4 0.037770 -0.229300 *
        17) seed.mass.log>=-0.479 11 0.048020 -0.129300
          34) seed.mass.log>=0.084 3 0.003994 -0.219700 *
          35) seed.mass.log< 0.084 8 0.010340 -0.095440 *
       9) Onset.flower>=168.5 7 0.069340 -0.053110 *
     5) seed.mass.log< -0.843 6 0.033370  0.038360 *
   3) Cnratio>=18.19 22 0.286000  0.068310
     6) Polycarpic>=0.5 18 0.144900  0.032050
      12) height< 16.04 6 0.021510 -0.020480 *
      13) height>=16.04 12 0.098550  0.058320
        26) height>=34.67 4 0.004003 -0.006159 *
        27) height< 34.67 8 0.069610  0.090560 *
     7) Polycarpic< 0.5 4 0.010980  0.231500 *
```

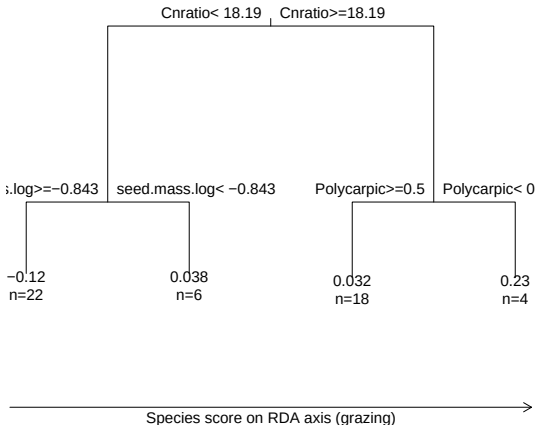
To prune the trees, we used cross validation:

```
> plotcp(rta.dist)
> cptab <- rta.dist$cptable
```



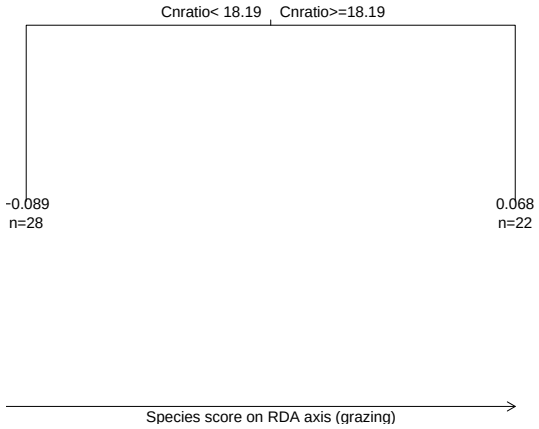
According to the method used (1-SE rule or minimization of `xerror`), the selected size of tree is different. For the minimization of `xerror`, we obtain:

```
> idx.min <- which.min(cptab[,4])
> rta.dist.pruned <- prune(rta.dist, cp =  1e-6+cptab[,1][idx.min])
> plot(rta.dist.pruned, ylim = c(0.33,1.1))
> text(rta.dist.pruned, use.n=T)
> arrows(0.9, 0.33, 4.1, 0.33, length = 0.1)
> text(2.5, 0.31, "Species score on RDA axis (grazing)", cex = 1)
```



For the 1-SE rule, we obtain:

```
> idx.min2 <- min((1:nrow(cptab))[cptab[,4]<min(cptab[,4]+cptab[,5])])
> rta.dist.pruned2 <- prune(rta.dist, cp =  1e-6+cptab[,1][idx.min2])
> plot(rta.dist.pruned2, ylim = c(0.33,1.1))
> text(rta.dist.pruned2, use.n=T)
> arrows(0.9, 0.33, 2, 0.33, length = 0.1)
> text(1.5, 0.31, "Species score on RDA axis (grazing)", cex = 1)
```



Whereas tree with just one predictor (C:N) explains 30.93 % of variability, the tree selected by the minimization of the `xerror` explains 56.75 %:

```
> 100 * (1-cptab[idx.min,3])
```

```
[1] 56.75
```

```
> 100 * (1-cptab[idx.min2,3])
```

```
[1] 30.93
```

5.3 Results for D-"RDA-mRTA"

```
> rda.both <- rda(X=spe, Y=env[,c("dist.int","SOIL.P")], scale = TRUE)
```

The amount of variation in species composition explained by both gradients together:

```
> 100 * rda.both$CCA$tot.chi/rda.both$tot.chi
```

```
[1] 11.24
```

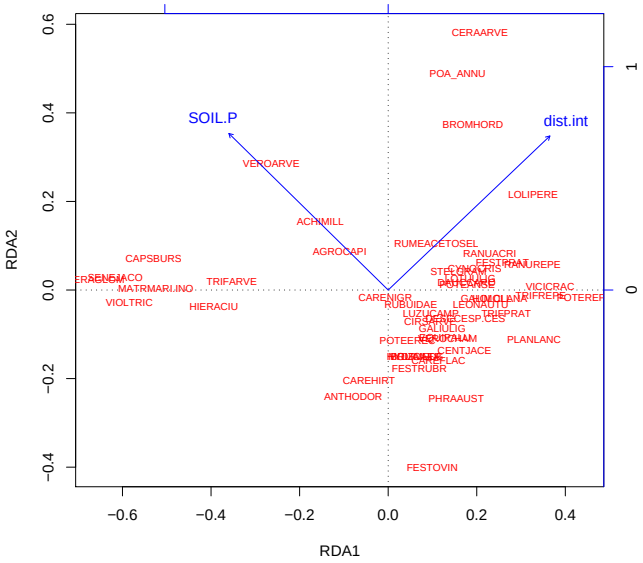
The explained variation is decomposed onto two axes:

```
> 100 * rda.both$CCA$eig/rda.both$CCA$tot.chi
```

```
 RDA1  RDA2
71.81 28.19
```

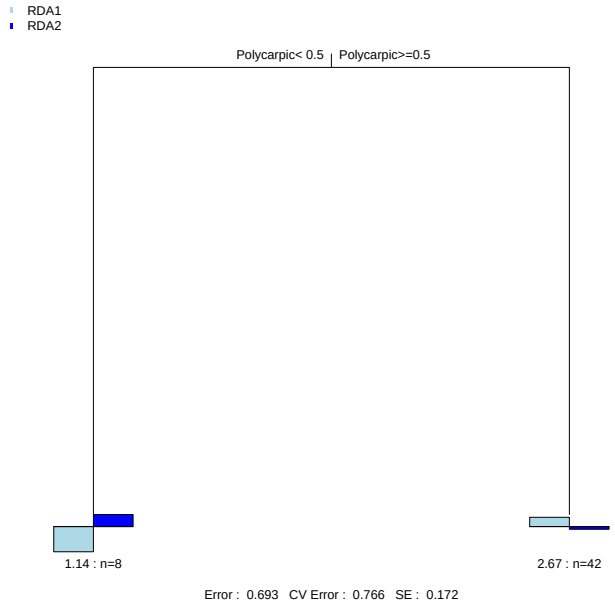
We can then represent the species and the environment variables on a plot:

```
> plot(rda.both, display=c("bp","sp"))
```



When applying the multivariate regression trees (with cross-validation and pruned using the 1-SE rule) to predict responses to both grazing and phosphorous together, only one of the predictors was able to improve the prediction of the two species responses: Monocarpic/Polycarpic.

```
> df.both <- cbind(RDA1=rda.both$CCA$v.eig[,1],RDA2=rda.both$CCA$v.eig[,2],traits)
> rta.both<-mvpart(data.matrix(df.both[,1:2])~Polycarpic+Cnratio+seed.mass.log
                  +SLA+height+Onset.flower,data=df.both, xval = 100, xv="1se")
```



The monocarpic plants respond positively to phosphorus, and negatively to disturbance.

6 E-"OMI-GAM": outlying mean index and generalised additive model

6.1 Description of the method

The Outlying Mean Index (OMI) is a multivariate method to separate species niches and to measure the distance between the mean habitat conditions used by each species and the mean habitat conditions of the study area (Dolédec et al., 2000; Thuiller et al., 2004). OMI-GAM starts by determining the species response to the environmental gradients (component 1) and models the contribution of each trait to the response (component 2). Finally, it clusters the species according to the responsive species - trait models (component 3).

- **Step 1:** Component 1 (species responses to environmental variables)

  The Outlying Mean Index makes no assumption about the shape of species response curves to the environment (e.g. unimodal or linear) and, unlike CCA and RDA, gives equal weight to species-rich and species-poor sites. The result of this analysis describes the mean position of the species in the environmental space (along each environmental axis), which represents a measure of the distance between the mean habitat conditions used by the species and the mean habitat conditions of the study area. It measures the propensity of the species to select a specialized environment.

- **Step 2:** Component 2 (identification of responsive trait combinations)

  They are various techniques to analysis the relationship between species' niche position and selected functional traits (e.g. regression-type, classification-type). Here we used inference-based generalized additive models. Stepwise regression-backward, forward or both–is an obvious method for examining the relative importance of each functional trait to explain species niche position on the selected axes. However, using usual stepwise regression to find the optimal combination of explanatory variables to model a response is often considered to be a high-variance operation because small perturbations of the response data can sometimes lead to vastly different subsets of the variables (Johnson and Omland, 2004). To avoid this problem, and to measure the actual power of each functional trait we used multimodal inference based on all-subsets selection of generalised additive models (Burnham and Anderson, 2002; Thuiller et al., 2007). In the case of six functional traits, there are $2^6 = 64$ possible models in an all-subsets selection. We thus estimated a small-sample (second order) bias adjustment of AIC (AICc) for each submodel. To estimate the weight of evidence of each functional trait (wpi) to explain species niche position on each OMI axis, we simply summed the model AICs weights (wi) over all models in which predictor appeared. To derive predicted species niche position, we averaged the predictions from each submodel weighted by the model AICc weight (see also method 2). This procedure was carried out for the two selected OMI axes.

- **Step 3:** Component 3 (grouping of species based on responsive traits)

  Outputs of inference-based GAM were used to define functional groups. Euclidean distances between species were computed on the predictions

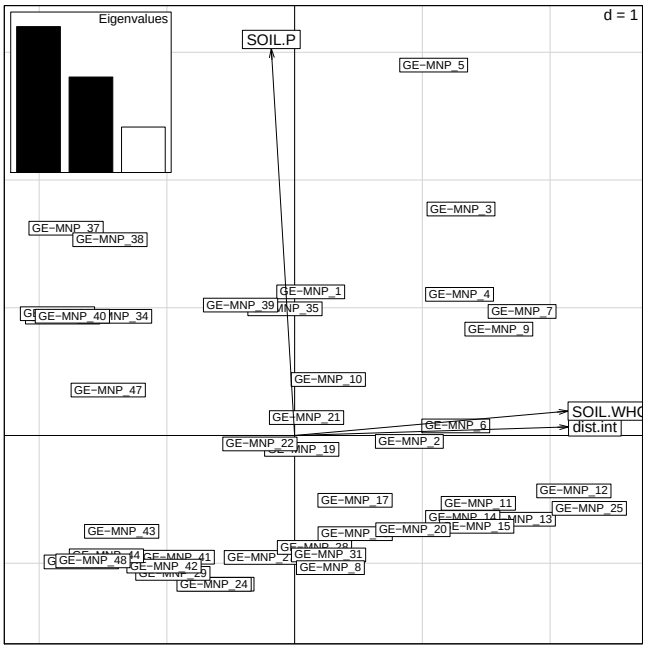
from inference-based GAM over the selected axes of OMI analysis and Ward's hiercharchical clustering was then performed. Clusters were extracted from the dendrogram and the optimal number of functional groups was determined with the Calinsky-Harabasz stopping criterion. Correlation ratios were computed to measure the degree of correlation between species traits and response groups.

6.2 Results

6.2.1 Analysis of environment

Prior to OMI analysis, environmental data must be analysed. Here, we use a PCA on correlation matrix:

```
> pca.env<-dudi.pca(env, scannf=FALSE)
> scatter(pca.env)
```



The total variation is decomposed onto othogonal axes. The percentage of variation associated to each axis:

```
> 100 * pca.env$eig/sum(pca.env$eig)
```

```
[1] 50.86 33.28 15.85
```

The correlations between environmental variables and PCA axes:

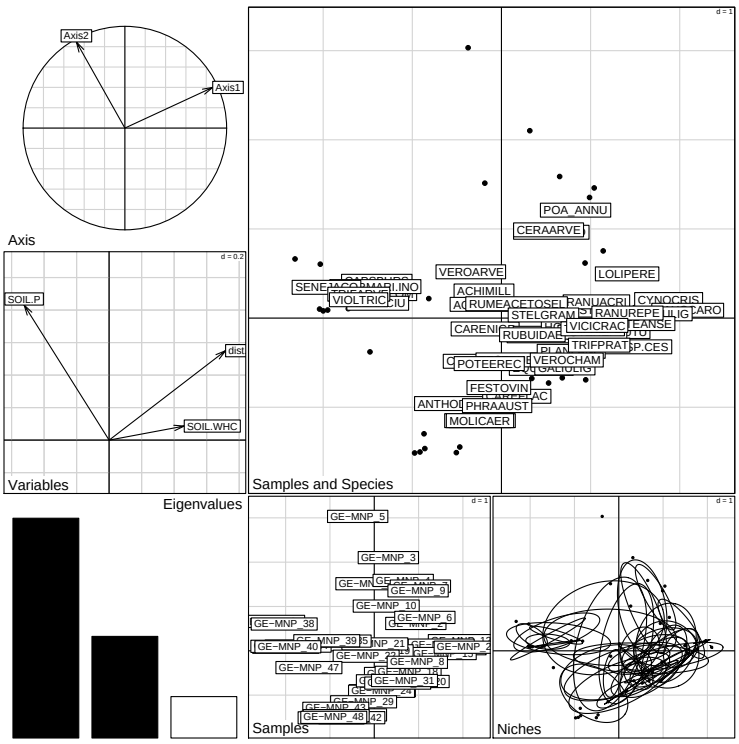
```
> pca.env$co
```

```
           Comp1   Comp2
dist.int  0.87264 0.02217
SOIL.P   -0.07431 0.99704
SOIL.WHC  0.87113 0.06285
```

6.2.2 Species niche description

Then, we analyse the distribution of species on environmental gradients using the OMI analysis. The method is implemented in the function `niche` of the `ade4` package.

```
> omi1<-niche(pca.env, spe, scannf=FALSE)
> plot(omi1)
```



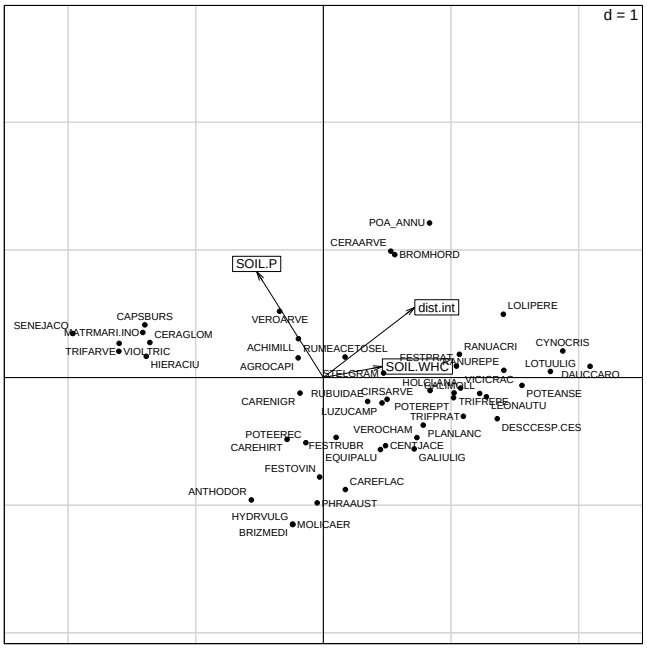
The variation of of the relationship between species and environmental gradients is along the first two axes:

```
> 100 * omi1$eig/sum(omi1$eig)
```

```
[1] 60.58 28.01 11.41
```

A biplot allows to represent species and environmental variables:

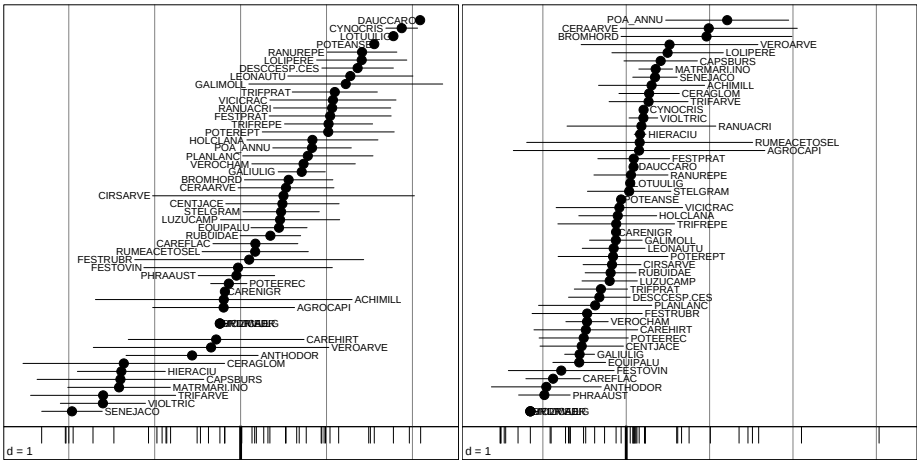
```
> s.arrow(omi1$c1, clab = 0.8, xlim=c(-2.5,2.5))
> s.label(omi1$li, xax = 1,yax = 2, clabel=0,add.plot = TRUE)
> par(mar = c(0.1, 0.1, 0.1, 0.1))
> pointLabel(omi1$li, rownames(omi1$li), cex=0.7)
```



The first axis is strongly positively linked to intensity of disturbance and to soil water holding capacity. The second axis is positively related to soil phosphorous.

Niche position and niche breadth on the first two axes of OMI analysis can be represented using the `sco.distri` function:

```
> par(mfrow=c(1,2))
> sco.distri(omi1$ls[,1],spe,clab=0.7)
> sco.distri(omi1$ls[,2],spe,clab=0.7)
```



The niche position of each species (contained in `omi1$li`) are then extracted for each axis (species scores) and used as response variable into the inference-based GAM with functional traits as explanatory variables.

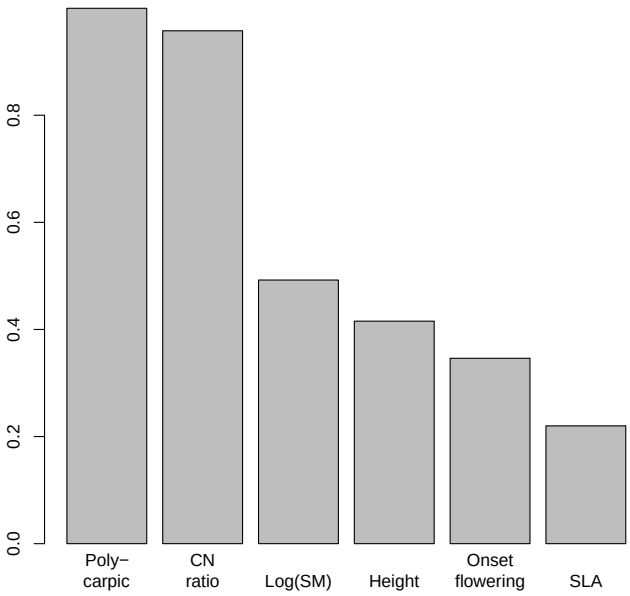
6.2.3 Inference based model - GAM

The generalised additive models will relate the mean position of species on OMI axes to species traits. The list of the 63 possible models (all possible models except the one which contains only the intercept) is created by the function `Inference_modelset`. Then, AICc and related measures corresponding to each model, are obtained by the `Inference_compute` function. Here, the 'Polycarpic' trait is coded as a factor (for a convenient GAM modelling).

```
> traits[,1]<-as.factor(traits[,1])
> modelset<-Inference_modelset(Explanatory=traits)
> inf.axis1 <- Inference_compute(Fam="gaussian",  combin=modelset[[1]], Mat=modelset[[2]],
   Response=omi1$li[,1], Explanatory=traits, Average = TRUE)
> inf.axis2 <- Inference_compute(Fam="gaussian",  combin=modelset[[1]], Mat=modelset[[2]],
   Response=omi1$li[,2], Explanatory=traits, Average = TRUE)
```

Variable importances from the inference based model (Figure 4a of the paper):

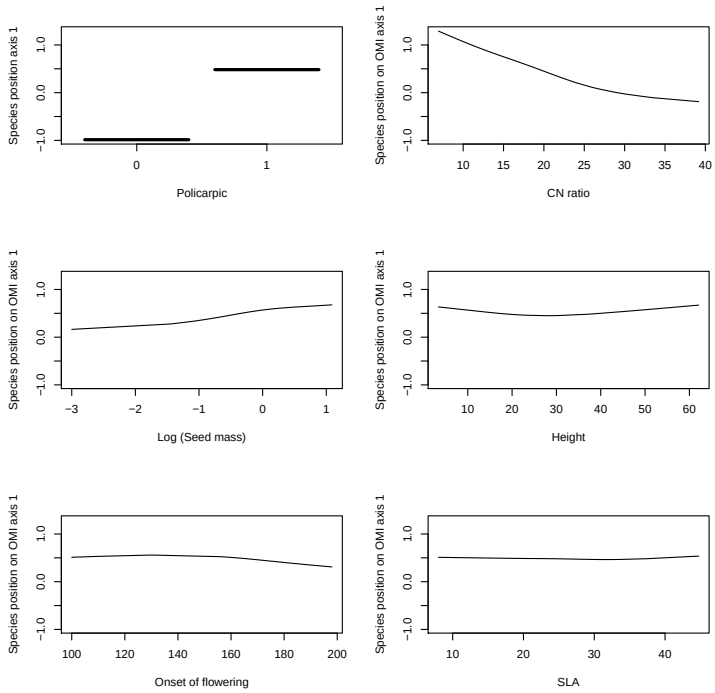
```
> dd.names <- c('Poly- carpic','CN ratio', 'Log(SM)',
  'Height','Onset flowering', 'SLA')
> dd.names.2 <- sapply(dd.names, function(x) gsub("\\s", "\\\n", x))
> barplot(inf.axis1$Var.importance[,1], names.arg=dd.names.2)
```



Along the OMI axis 1 (intensity of disturbance – soil water holding capacity), the GAM inference-based approach together with the permutation test expressed C:N-ratio and flowering mode (polycarpic vs. monocarpic) as relatively important.

Response curves for the OMI axis 1 are then plotted for each trait (Figure 4b in the paper).

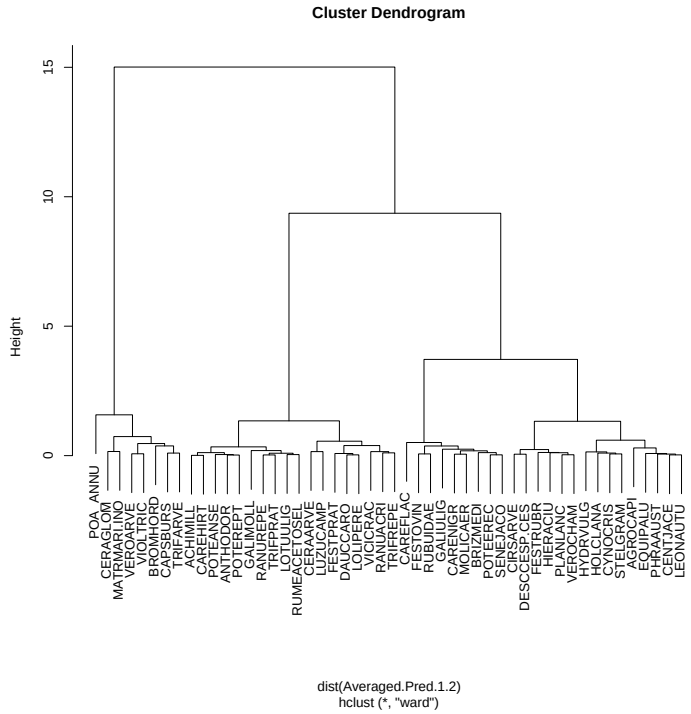
```
> Limits <- apply(inf.axis1$Plot.response[, seq(2, 12, by=2)], 2, range)
> lim <- c(min(Limits[1,]), max(Limits[2,]))
> par(mfrow=c(3,2))
> plot(as.factor(inf.axis1$Plot.response[,1]), inf.axis1$Plot.response[,2],
      ylim=lim, type="l", xlab="Policarpic", ylab="Species position axis 1")
> plot(inf.axis1$Plot.response[,3], inf.axis1$Plot.response[,4], ylim=lim,
      type="l", xlab="CN ratio", ylab="Species position on OMI axis 1")
> plot(inf.axis1$Plot.response[,5], inf.axis1$Plot.response[,6], ylim=lim,
      type="l", xlab="Log (Seed mass)", ylab="Species position on OMI axis 1")
> plot(inf.axis1$Plot.response[,9], inf.axis1$Plot.response[,10], ylim=lim,
      type="l", xlab="Height", ylab="Species position on OMI axis 1")
> plot(inf.axis1$Plot.response[,11], inf.axis1$Plot.response[,12], ylim=lim,
      type="l", xlab="Onset of flowering", ylab="Species position on OMI axis 1")
> plot(inf.axis1$Plot.response[,7], inf.axis1$Plot.response[,8], ylim=lim,
      type="l",  xlab="SLA", ylab="Species position on OMI axis 1")
```



In summary, species on intensely disturbed sites with high soil water holding capacity tend to be polycarpic, and to have lower CN ratio than species occurring in less intense disturbed places. Along the second axis (soil phosphorous gradient, not shown in the paper and the tutorial), flowering mode, CN ratio are again the most correlated to species position, followed by onset of flowering. Monocarpic species tend to be preferably on sites with high soil phosphorous content, with high CN ratio and early onset of flowering than species occurring on lower soil phosphorous content.

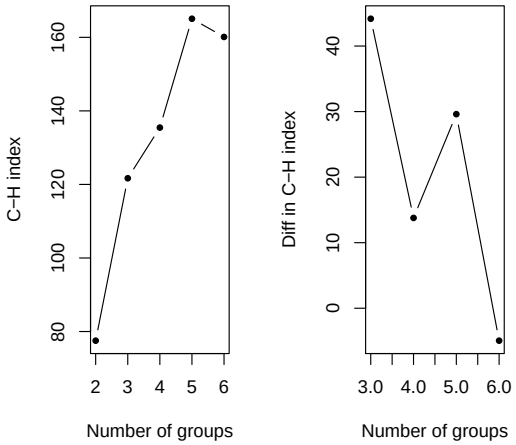
We perform the classification using these scores to obtain functional groups:

```
> Averaged.Pred.1.2<-cbind(inf.axis1$Averaged.Pred, inf.axis2$Averaged.Pred)
> hc1 <- hclust(dist(Averaged.Pred.1.2), method = "ward")
> plot(hc1)
```



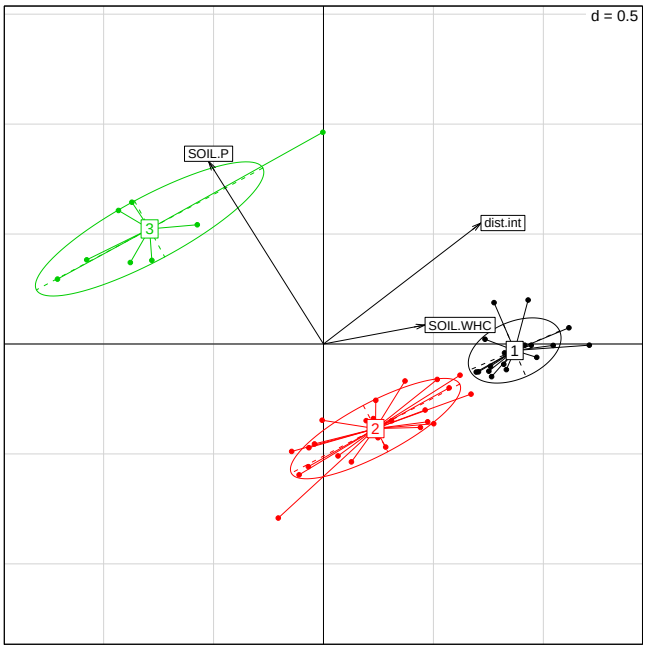
We use the Calinsky-Harabasz criteria to find the best partition (try between 2 and 6 groups).

```
> ntest <- 6
> res <- rep(0,ntest - 1)
> for (i in 2:ntest){
   fac <- cutree(hc1, k = i)
   res[i-1] <- calinski(tab=Averaged.Pred.1.2, fac = fac)[1]
 }
> par(mfrow=c(1,2))
> plot(2:ntest, res, type='b', pch=20, xlab="Number of groups", ylab = "C-H index")
> plot(3:ntest, diff(res), type='b', pch=20, xlab="Number of groups", ylab = "Diff in C-H index")
```



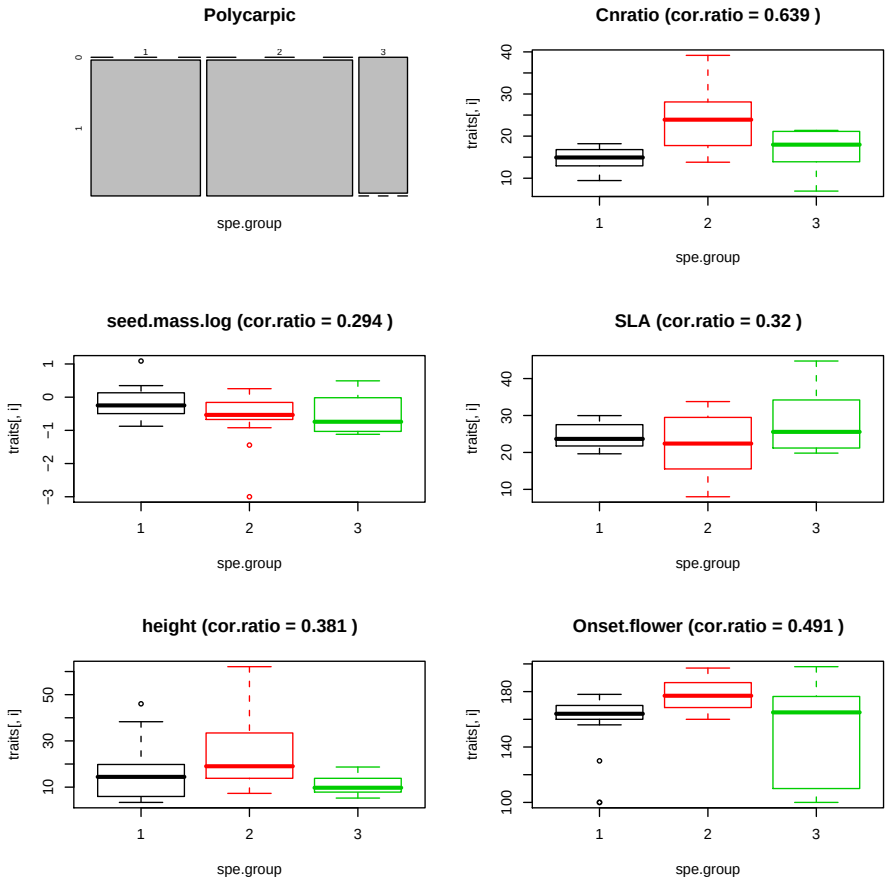
It was not possible to identify an optimal number of groups (corresponding to a maximal value of Calinski-Harabasz criterion). Differences between subsequent values of the criterion suggest a partition into 3 groups. Three functional groups are then represented (Figure 4c in the paper):

```
> nbgroup <- 3
> spe.group <- as.factor(cutree(hc1, k = nbgroup))
> spe.group <- as.factor(spe.group)
> s.class(Averaged.Pred.1.2, spe.group, col= 1:nlevels(spe.group))
> s.arrow(omi1$c1, xax=1, yax=2, csub = 1, clab = 0.8, add.plot=T)
```



We can interpret this partition in terms of traits :

```
> eta2 <- cor.ratio(traits[,-1], data.frame(spe.group), weights = rep(1, length(spe.group)))
> par(mfrow=n2mfrow(ncol(traits)))
> plot(table(spe.group,traits[,1]), main =names(traits)[1])
> for(i in 2:ncol(traits)){
   label <- paste(names(traits)[i], "(cor.ratio =", round(eta2[i-1],3), ")")
   plot(traits[,i]~spe.group, main = label, border = 1:nlevels(spe.group))
 }
```



The first group contains 18 polycarpic species with a low CN ratio, occurring in very disturbed sites, with a high soil water holding capacity and medium soil phosphorous content. The second group comprises 24 polycarpic species with a higher CN ratio than group 1 and 3, later onset of flowering that the other groups, and occurring mostly in slightly disturbed sites and soil water content, but a low soil phosphorous content. The third group contains only monocarpic species, which have intermediate CN ratio compared to the other groups and a large variance for onset of flowering. These species mostly occur in site with high soil phosphorous and low intensity of disturbance and soil water content.

The 'Polycarpic' trait is then retransformed into a binary variable:

```
> traits[,1]=as.numeric(traits[,1])
```

7 F-"RLQ": RLQ Analysis

7.1 Description of the method

RLQ analysis (Dolédec et al., 1996) is a three-table ordination method that allows the simultaneous analysis of tables **R**, **L** and **Q** in order to summarize and represent graphically the main patterns of co-variation between trait data and environmental parameters (components 1, 2). A subsequent cluster analysis based on the co-variation then produces functional groups (component 3).

- **Step 1, 2:** Component 1, 2 (species responses to environmental variables and identification of responsive trait combinations)

  RLQ analysis is an extension of the two-table method of co-inertia analysis (Dolédec and Chessel, 1994; Dray et al., 2003). It aims at finding a site score (linear combination of environmental variables) and a species score (linear combination of traits) maximizing the co-inertia criterion. This criterion is the product of the variance of the site scores by the variance of the species scores and by the squared cross-correlation between the species score and the sites score mediated by table **L**.

- **Step 3:** Component 3 (grouping of species based on responsive traits)

  Outputs of RLQ analysis were used to define functional groups. Euclidean distances between species were computed on the first two axes of RLQ analysis and Ward's hiercharchical clustering was then performed. Clusters were extracted from the dendrogram and the optimal number of functional groups was determined with the Calinsky-Harabasz stopping criterion. Correlation ratios were computed to measure the degree of correlation between species traits and response groups.

7.2 Results

Prior to the analysis, the table **L** must be analysed by correspondence analysis. Species and sites weights computed in these analysis are then used in the analyses of species traits (**Q**) and environmental variables (**R**).

```
> pca.traits <- dudi.pca(traits, row.w = coa1$cw, scannf = FALSE)
> pca.env <- dudi.pca(env, row.w = coa1$lw, scannf = FALSE)
```

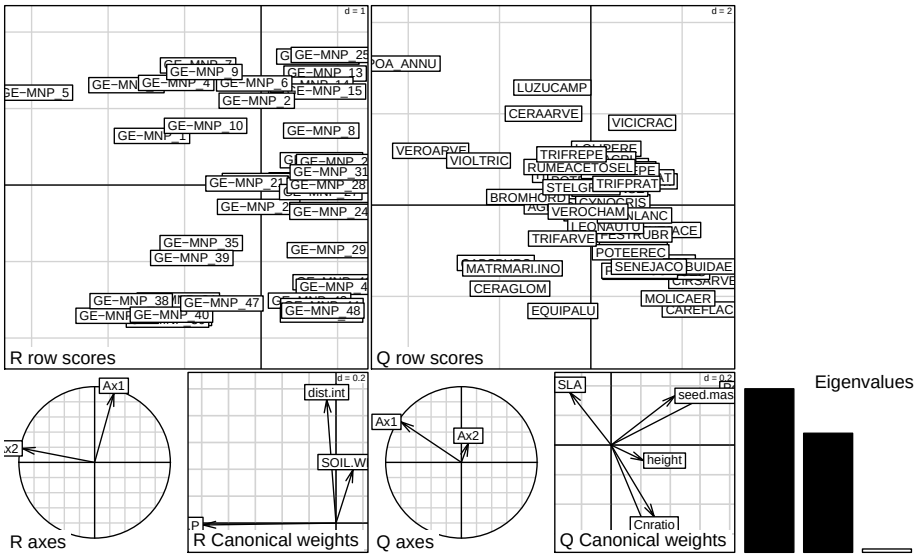
The RLQ analysis is performed using the `rlq` function of the `ade4` package:

```
> rlq1 <- rlq(pca.env, coa1, pca.traits, scannf = FALSE)
> summary(rlq1)


Eigenvalues decomposition:
     eig  covar    sdR   sdQ   corr
1 0.3474 0.5894 0.9994 1.317 0.4479
2 0.2530 0.5030 1.1860 1.136 0.3732

Inertia & coinertia R:
   inertia   max  ratio
1   0.9989 1.581 0.6316
12  2.4054 2.553 0.9423

Inertia & coinertia Q:
   inertia   max  ratio
1    1.734 2.223 0.7800
12   3.025 3.195 0.9467
```

```
Correlation L:
    corr    max  ratio
1 0.4479 0.8483 0.5280
2 0.3732 0.8160 0.4574
```

The main outputs of the analysis can be represented:

```
> plot(rlq1)
```



The co-structure between traits and environment is mainly decomposed onto the two first axes of RLQ analysis (57.18 % and 41.64 % of the co-inertia criterion for the first and second RLQ axis respectively):

```
> ## Percentage of co-Inertia for each axis
> 100*rlq1$eig/sum(rlq1$eig)
```

```
[1] 57.178 41.642  1.180
```

To interpret the results, correlations can be computed:

```
> ## weighted correlations axes / env.
> t(pca.env$tab)%*%(diag(pca.env$lw))%*%as.matrix(rlq1$mR)
```

```
            NorS1   NorS2
dist.int  0.09551  0.9600
SOIL.P   -0.99445 -0.1151
SOIL.WHC  0.17425  0.7638
```

```
> ## weighted correlations axes / traits.
> t(pca.traits$tab)%*%(diag(pca.traits$lw))%*%as.matrix(rlq1$mQ)
```

```
                NorS1   NorS2
Polycarpic     0.7171  0.1551
Cnratio        0.5583 -0.6262
seed.mass.log  0.4932  0.1148
SLA           -0.6396  0.6372
height         0.5221 -0.3725
Onset.flower   0.4782 -0.7687
```

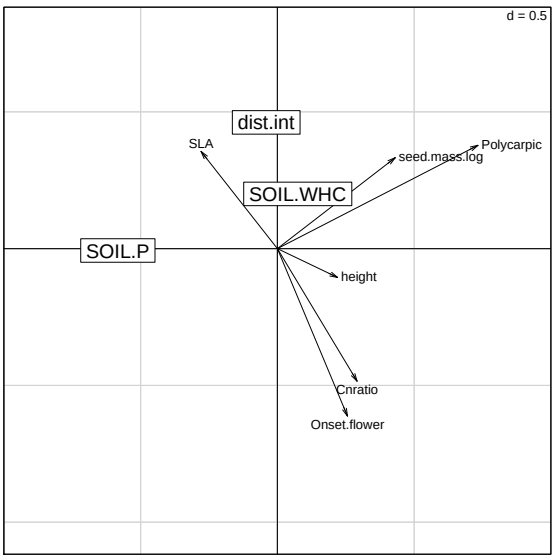
```
> ## correlations traits / env.
> rlq1$tab
```

```
         Polycarpic Cnratio seed.mass.log     SLA   height
dist.int     0.1468 -0.2183       0.13662 0.17193 -0.08406
SOIL.P      -0.4312 -0.1729      -0.25391 0.16159 -0.11785
SOIL.WHC     0.1244 -0.1153       0.09662 0.05833  0.05672
         Onset.flower
dist.int     -0.30212
SOIL.P       -0.14185
SOIL.WHC     -0.08219
```

The first axis is negatively correlated to soil phosophate. It is also negatively related to SLA and positively to all other traits. The second axis is positively correlated to disturbance frequency and soil water content. It is positively related to SLA and negatively to Onset of flowering, C:N ratio.

A biplot representing traits and environmental variables (Figure 3c in the paper) can be constructed:

```
> s.arrow(rlq1$c1, xlim=c(-1,1), boxes = FALSE)
> s.label(rlq1$li, add.plot=T, clab=1.5)
```

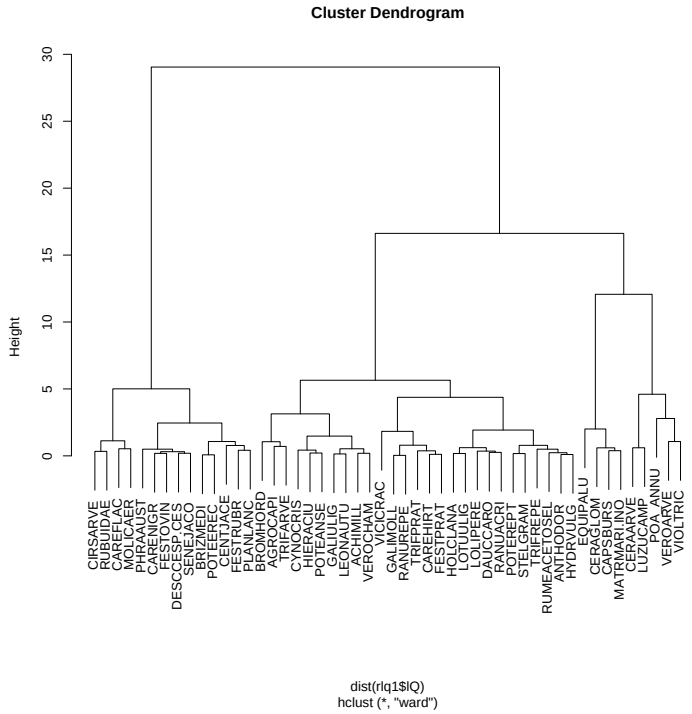


Species scores on the first two axes of RLQ analysis:

```
> s.label(rlq1$lQ, clabel = 0)
> par(mar = c(0.1, 0.1, 0.1, 0.1))
> pointLabel(rlq1$lQ,row.names(rlq1$lQ), cex=0.7)
```

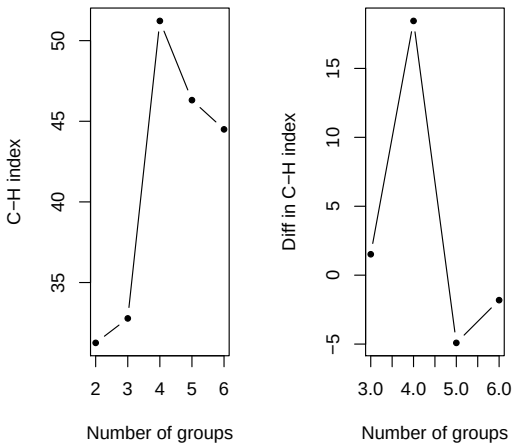
We perform the classification using these scores to obtain functional groups:

```
> hc2 <- hclust(dist(rlq1$lQ), method = "ward")
> plot(hc2)
```



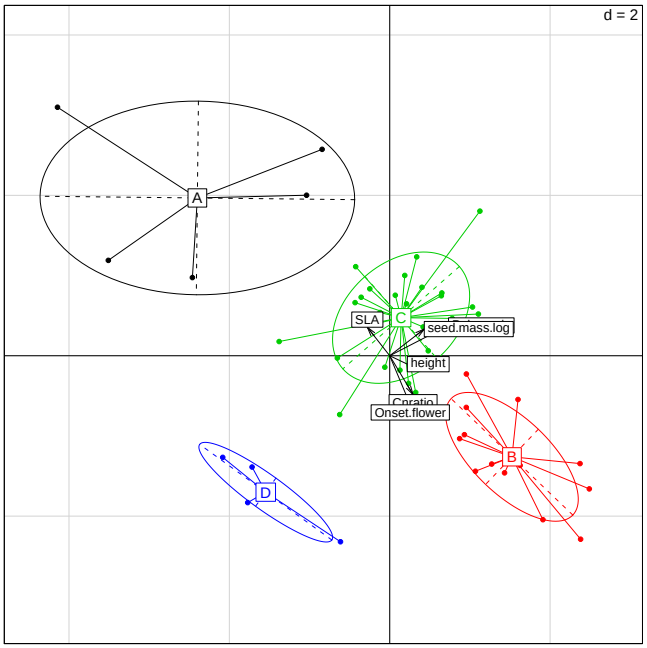
We use the Calinsky-Harabasz criteria to find the best partition (try between 2 and 6 groups) :

```
> ntest <- 6
> res <- rep(0,ntest - 1)
> for (i in 2:ntest){
   fac <- cutree(hc2, k = i)
   res[i-1] <- calinski(tab=rlq1$lQ, fac = fac)[1]
 }
> par(mfrow=c(1,2))
> plot(2:ntest, res, type='b', pch=20, xlab="Number of groups", ylab = "C-H index")
> plot(3:ntest, diff(res), type='b', pch=20, xlab="Number of groups", ylab = "Diff in C-H index")
```



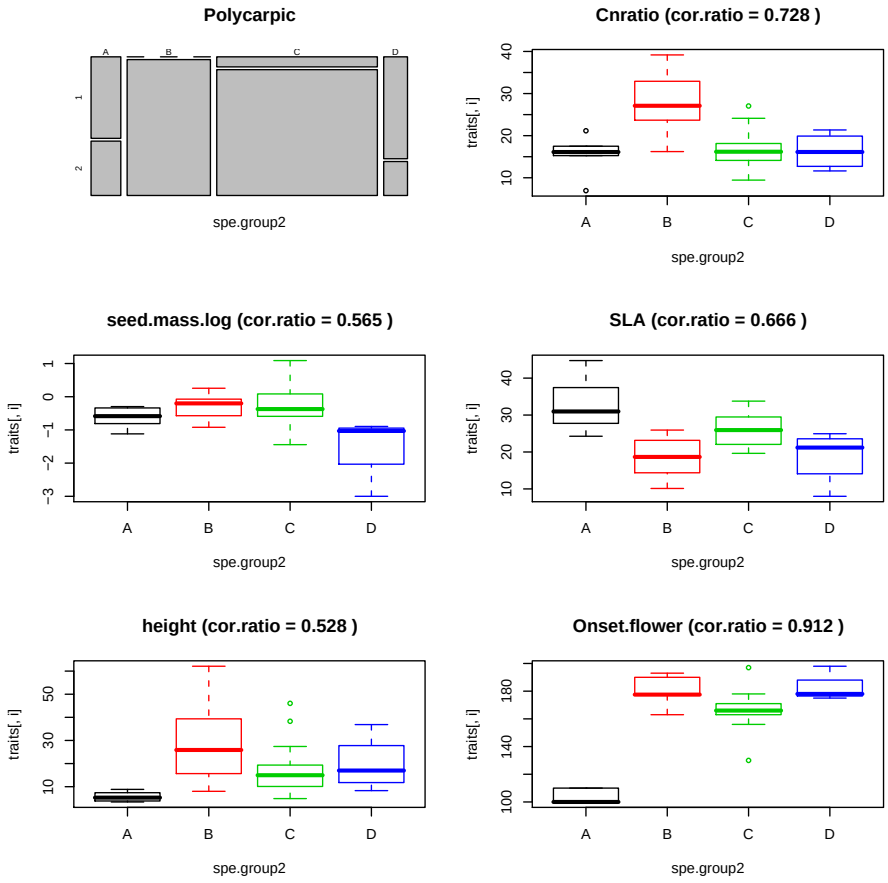
The best partition is for 4 groups. In the next figure, each point represents the modelled species position on RLQ axes 1 and 2 , and each colour the group from the cluster:

```
> spe.group2 <- as.factor(cutree(hc2, k = which.max(res) +1))
> levels(spe.group2) <- c("C","B","D","A")
> spe.group2 <- factor(spe.group2, levels=c("A","B","C","D"))
> s.class(rlq1$lQ, spe.group2, col= 1:nlevels(spe.group2))
> s.arrow(rlq1$c1, add.plot = T,clab=0.8)
```



We can interpret this partition in terms of traits :

```
> eta2 <- cor.ratio(traits[,-1], data.frame(spe.group2), weights = rep(1, length(spe.group2)))
> par(mfrow=n2mfrow(ncol(traits)))
> plot(table(spe.group2,traits[,1]), main =names(traits)[1])
> for(i in 2:ncol(traits)){
   label <- paste(names(traits)[i], "(cor.ratio =", round(eta2[i-1],3), ")")
   plot(traits[,i]~spe.group2, main = label, border = 1:nlevels(spe.group2))
 }
```



A first group (A) contains 5 species with very low value of Onset of flowering, low height and high SLA. These species occupied mainly highly disturbed environment with high phosphate. A second group (B) of 14 polycarpic species is identified. These species have very high C:N ratio, high height, low SLA and high value of Onset of flowering and they occupied sites with low phosphate and moderate disturbance intensity. A third intermediate group (C) contains 27 species which are mainly polycarpic (25 species), with quite high SLA. These species occupied mainly highly disturbed environment with low phosphate. The fourth group (D) with 4 species including 3 monocarpic corresponds to species with high value of of Onset of flowering, low seed mass and low SLA that occupied sites with high phosphate and low disturbance intensity.

The classification obtained is quite similar to the one obtained with OMIGAM:

```
> table(spe.group,spe.group2)
```

```
          spe.group2
spe.group  A  B  C  D
        1  2  0 16  0
        2  0 14  9  1
        3  3  0  2  3
```

The only difference is that RLQ has an additional group which corresponds to the splitting of a group of 24 species into two groups of 15 and 9 species.

8 G-"Double CCA": double canonical correspondence analysis

8.1 Description of the method

- **Step 1, 2:** Component 1, 2 (species responses to environmental variables and identification of responsive trait combinations)

  Double CCA is a three-table ordination method proposed by Lavorel et al. (1999, 1998). In the classical context, canonical correspondence analysis (CCA) is used to link tables **L** and **R** in order to ordinate the community data in the light of the environmental variables. It is well known that CCA implies two main steps: (1) prediction of community data by environment and (2) ordination of predicted values. Ojeda et al. (1998) performed an unusual CCA in which the ordination of **L** is constrained by the species traits table **Q**. Lavorel and coauthors proposed to combine these two CCA in one analysis nicknamed "double CCA". This approach ordinates **L** by taking the effects of **R** and **Q** simultaneously into account. Double CCA encompasses also two steps: (1) prediction of community data by both environmental variables and species traits and (2) ordination of predicted values.

- **Step 3:** Component 3 (grouping of species based on responsive traits)

  As in RLQ, functional groups were defined using Ward's hiercharchical clustering with the Calinsky-Harabasz stopping criterion using species scores for the first two axes of the double CCA. Correlation ratios were computed to measure the degree of correlation between species traits and response groups.

8.2 Results

Double CCA is based on the correspondence analysis of the species-by-sites table. In this analysis, the ordination of sites and species is constrained by both species traits and environment:

```
> dbcca1 <- dbrda(coa1,env, traits, scannf = FALSE)
```

For the double CCA, the three environmental variables and the six traits explain 6.25 % of the variation (14.89 % for the environment in standard CCA).

```
> ## percentage of explained variation by the environment
> sum(cca1$eig)/sum(coa1$eig)*100
```

```
[1] 14.89
```

```
> ## percentage of explained variation by both traits and env.
> sum(dbcca1$eig)/sum(coa1$eig)*100
```

```
[1] 6.248
```

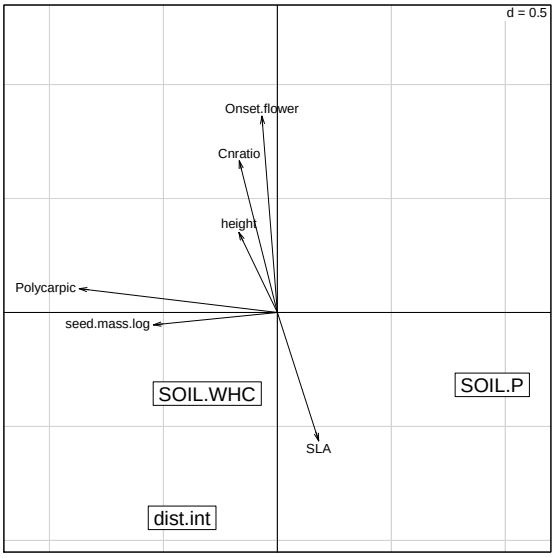
This explained variation is mainly decomposed onto the first two axes of the analysis (58.2 % and 38.34 % for the first and second axis respectively):

```
> ## Percentage of variation explained by each axis
> 100*dbcca1$eig/sum(dbcca1$eig)
```

```
[1] 58.196 38.338  3.466
```

Correlations between axes and traits and environmental variables (Figure 3c in the paper) can then be used to interpret the results:

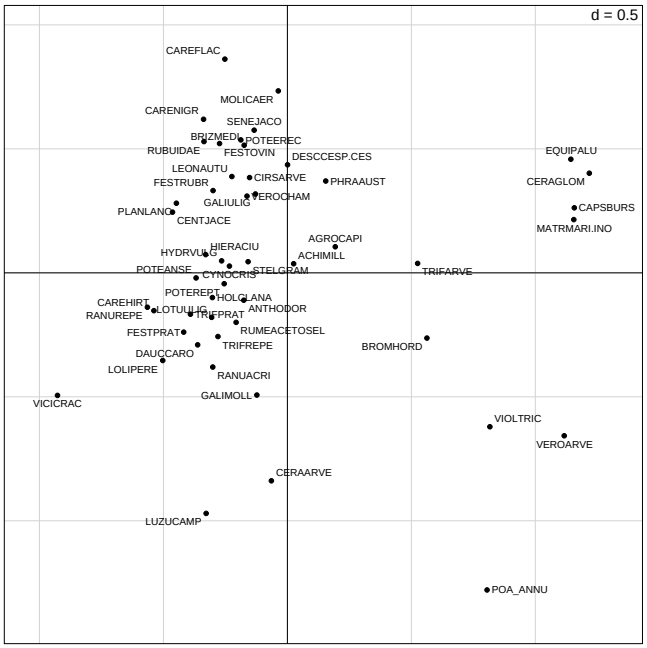
```
> s.arrow(dbcca1$corZ[-1,], xlim=c(-1.2,1.2), boxes = FALSE)
> s.label(dbcca1$corX[-1,], add.plot=T, clab=1.5)
```



The first axis is positively correlated to soil phosophate and negatively correlated to disturbance frequency. It is also negatively to polycarpic life history and seed mass. The second axis is negatively correlated to disturbance frequency and also to soil phosophate and soil water holding capacity. It is negatively related to SLA and positively to Onset of flowering, C:N ratio and height.

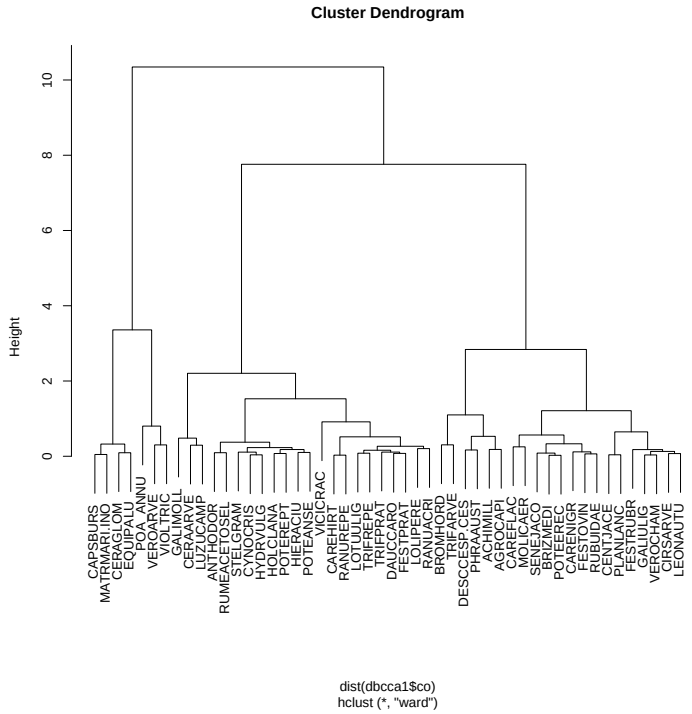
Species scores on the first two axes of double CCA:

```
> s.label(dbcca1$co, clabel = 0)
> par(mar = c(0.1, 0.1, 0.1, 0.1))
> pointLabel(dbcca1$co,row.names(dbcca1$co), cex=0.7)
```



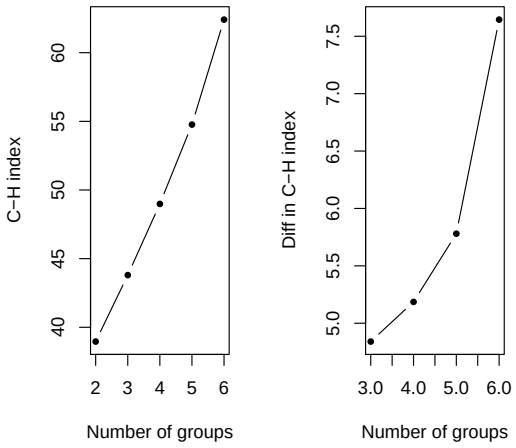
We perform the classification using these scores to obtain functional groups:

```
> hc3 <- hclust(dist(dbcca1$co), method = "ward")
> plot(hc3)
```



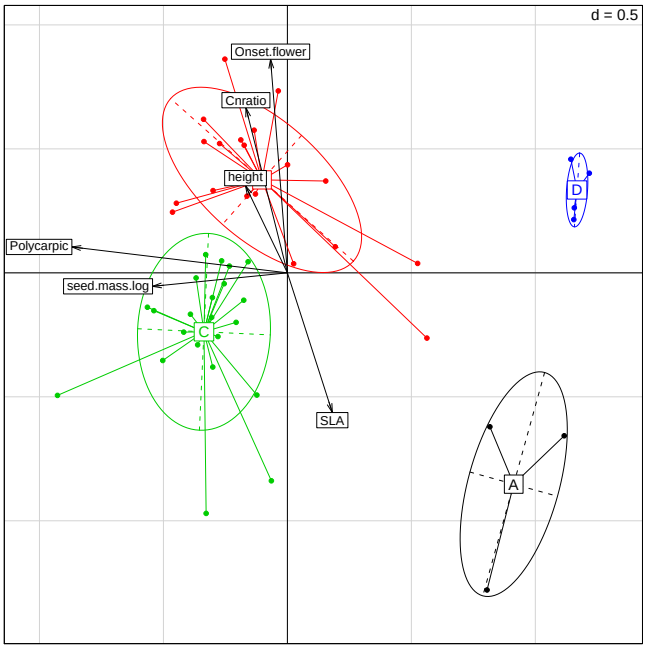
We use the Calinsky-Harabasz criteria to find the best partition (try between 2 and 6 groups) :

```
> ntest <- 6
> res <- rep(0,ntest - 1)
> for (i in 2:ntest){
   fac <- cutree(hc3, k = i)
   res[i-1] <- calinski(tab=dbcca1$co, fac = fac)[1]
 }
> par(mfrow=c(1,2))
> plot(2:ntest, res, type='b', pch=20, xlab="Number of groups", ylab = "C-H index")
> plot(3:ntest, diff(res), type='b', pch=20, xlab="Number of groups", ylab = "Diff in C-H index")
```



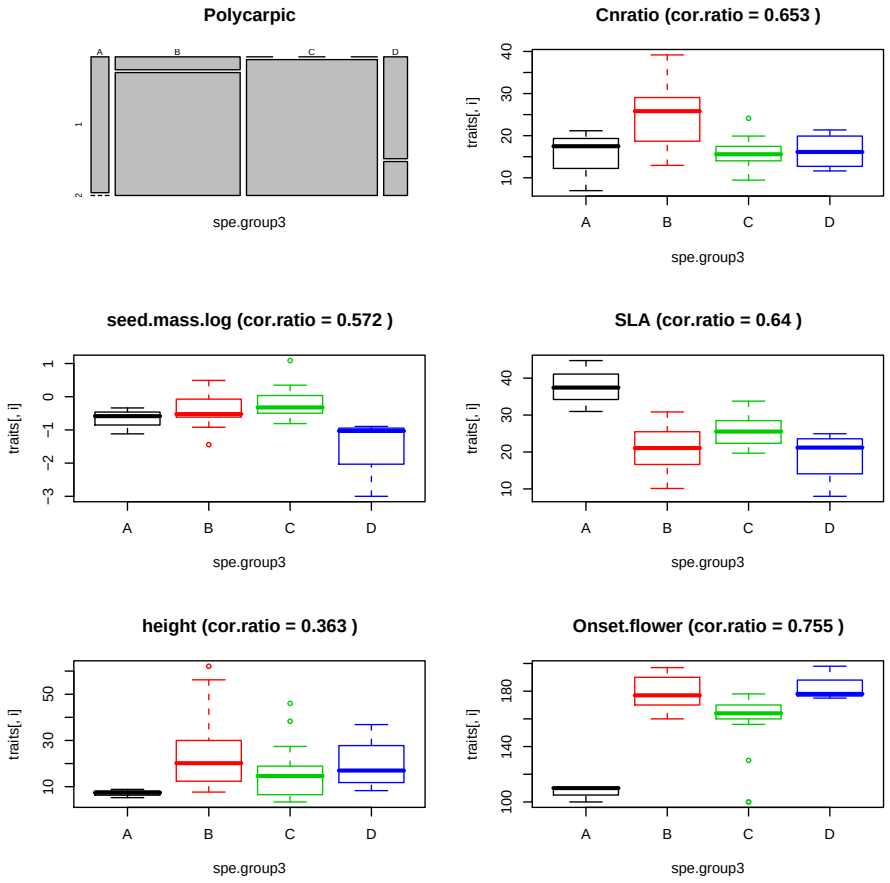
It was not possible to identify an optimal number of groups as the Calinski-Harabasz criterion increases (but never decreases) with the number of groups. Differences between subsequent values were not helpful in this case. The number of groups was then arbitrary set to 4 (i.e., the number of groups identified for the RLQ analysis).

```
> nbgroup <- ifelse((which.max(res) + 1) == ntest, nlevels(spe.group2), which.max(res) + 1)
> spe.group3 <- as.factor(cutree(hc3, k = nbgroup))
> levels(spe.group3) <- c("B","C","D","A")
> spe.group3 <- factor(spe.group3, levels=c("A","B","C","D"))
> s.class(dbcca1$co, spe.group3, col = 1:nlevels(spe.group3))
> s.arrow(dbcca1$corZ[-1,],add.plot=T,clab=0.8)
```



We can interpret this partition in terms of traits :

```
> eta2 <- cor.ratio(traits[,-1], data.frame(spe.group3), weights = rep(1, length(spe.group3)))
> par(mfrow=n2mfrow(ncol(traits)))
> plot(table(spe.group3,traits[,1]), main =names(traits)[1])
> for(i in 2:ncol(traits)){
   label <- paste(names(traits)[i], "(cor.ratio =", round(eta2[i-1],3), ")")
   plot(traits[,i]~spe.group3, main = label, border = 1:nlevels(spe.group3))
 }
```



The classification obtained is quite similar to the one obtained with RLQ analysis:

```
> table(spe.group3,spe.group2)
```

```
           spe.group2
spe.group3  A  B  C  D
         A  3  0  0  0
         B  0 14  7  0
         C  2  0 20  0
         D  0  0  0  4
```

Two species (*Cerastium arvense* and *Luzula campestris*) of group A moves to group C while 7 species (*Achillea millefolium*, *Agrostis capillaris*, *Bromus hordeaceus*, *Galium uliginosum*, *Leontodon autumnalis*, *Trifolium arvense* and *Veronica chamaedrys*) moves from group C to group B.

References

A. Agresti. *Categorical data analysis*. John Wiley and Sons, 2002.

K. Burnham and D. Anderson. *Model selection and multimodel inference. A practical information-theoretic approach*. Springer, 2002.

S. Dolédec and D. Chessel. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology*, 31:277–294, 1994.

S. Dolédec, D. Chessel, C. ter Braak, and S. Champely. Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, 3:143–166, 1996.

S. Dolédec, D. Chessel, and C. Gimaret-Carpentier. Niche separation in community analysis: a new method. *Ecology*, 81(10):2914–2927, 2000.

S. Dray, D. Chessel, and J. Thioulouse. Co-inertia analysis and the linking of ecological data tables. *Ecology*, 84:3078–3089, 2003.

B. Everitt, S. Landau, and M. Leese. *Cluster analysis*. Arnold, London, 2001.

A. Gordon. *Classification*. Chapman and Hall, Boca Raton, 1999.

C. Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, 52(1):258–271, 2007.

J. Johnson and K. Omland. Model selection in ecology and evolution. *Trends in Ecology & Evolution*, 19(2):101–108, 2004.

S. Lavorel, B. Touzard, J. Lebreton, and B. Clément. Identifying functional groups for response to disturbance in an abandoned pasture. *Acta Oecologica - International Journal of Ecology*, 19(3):227–240, 1998.

S. Lavorel, C. Rochette, and J. Lebreton. Functional groups for response to disturbance in Mediterranean old fields. *Oikos*, 84:480–498, 1999.

P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier Science, Amsterdam, 2nd edition, 1998.

R. Mac Nally. Regression and model-building in conservation biology, biogeography and ecology: the distinction between–and reconciliation of–'predictive' and 'explanatory' models. *Biodiversity and Conservation*, 9(5):655–671, 2000.

F. Ojeda, J. Arroyo, and T. Marañon. The phytogeography of European and Mediterranean heath species (Ericoideae, Ericaceae): a quantitative analysis. *Journal of Biogeography*, 25:165–178, 1998.

O. Petchey and K. Gaston. Functional diversity: back to basics and looking forward. *Ecology Letters*, 9(6):741–758, 2006.

C. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya A*, 26:329–359, 1964.

P. Sneath and R. Sokal. *Numerical taxonomy: the principles and practice of numerical classification*. Freeman, San Francisco, 1973.

B. Strauss and R. Biedermann. Urban brownfields as temporary habitats: driving forces for the diversity of phytophagous insects. *Ecography*, 29:928–940, 2006.

C. ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167–1179, 1986.

W. Thuiller, S. Lavorel, G. Midgley, S. Lavergne, and T. Rebelo. Relating plant traits and species distributions along bioclimatic gradients for 88 Leucadendron taxa. *Ecology*, 85(6):1688–1699, 2004.

W. Thuiller, J. Slingsby, S. Privett, and R. Cowling. Stochastic species turnover and stable coexistence in a species-rich, fire-prone plant community. *PLoS ONE*, 2:e938, 2007. URL `http://dx.plos.org/10.1371%2Fjournal.pone.0000938`.

M. Whittingham, P. Stephens, R. Bradbury, and R. Freckleton. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, 75(5):1182–1189, 2006.