# Multivariate Statistical Tools in Ecology

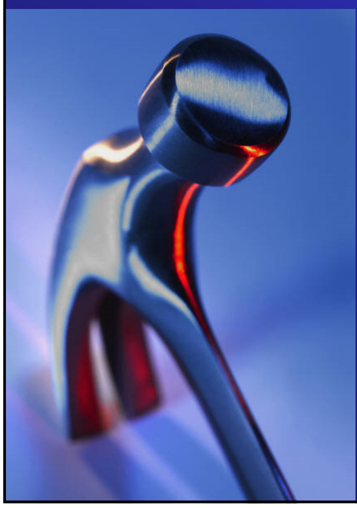
ISCED, Lubango, March 2016

**J. Paulo Sousa**
**Laboratory of Soil Ecology and Ecotoxicology**
**Centre for Functional Ecology**
**Universidade de ISCED, Lubango, Portugal**
jps@zoo.uc.pt
http://cfe.uc.pt/paulosousa
http://www.facebook.com/labsolos

Soil Ecology & Ecotoxicology Lab



## Why to use multivariate techniques ?



*"If the only tool you know is a hammer you will tend to see all your problems as nails!"*

Multivariate Statistical Tools in Ecology

ISCED, Lubango, March 2016

# Introductory notes to Multivariate Analysis Tools



## Why to use multivariate techniques ?

- **Several attributes describe each subject or each sample**
- **Examples:**
  - Effects of a chemical on soil fauna communities
  - Plant, animal or microbial communities under different treatments along with the measurement of several environmental variables
  - Monitoring data with the evaluation of several variables along time

## Why to use multivariate techniques ?

- **Data matrix (part)**
  - ➢ **Sparse data** (many zeros)
  - ➢ **Most species are infrequent** (present in a few locations)
  - ➢ The **number of factors** influencing species composition is **potentially very large**
  - ➢ The **number of important factors is typically few**
  - ➢ There is **much noise** (replicate samples will vary substantially from each other)

| Sample | Dip | Het | Hom | Lep | Col | Thy | Ort | Pso |
|---|---|---|---|---|---|---|---|---|
| C1 | 0 | 0 | 0 | 0 | 16 | 0 | 0 | 0 |
| C2 | 0 | 0 | 0 | 0 | 48 | 0 | 0 | 0 |
| C3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C4 | 0 | 16 | 16 | 0 | 0 | 0 | 0 | 0 |
| C5 | 0 | 0 | 0 | 0 | 64 | 32 | 0 | 0 |
| PDA1 | 16 | 32 | 0 | 0 | 112 | 0 | 0 | 80 |
| PDA2 | 0 | 0 | 0 | 0 | 96 | 0 | 0 | 16 |
| PDA3 | 0 | 0 | 16 | 0 | 96 | 0 | 0 | 16 |
| PDA4 | 0 | 16 | 0 | 0 | 208 | 0 | 0 | 32 |
| PDA5 | 0 | 144 | 0 | 0 | 192 | 0 | 0 | 48 |
| PDB1 | 0 | 16 | 0 | 0 | 0 | 0 | 16 | 0 |
| PDB2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 |
| PDB3 | 0 | 0 | 16 | 0 | 80 | 0 | 0 | 0 |
| PDB4 | 0 | 16 | 0 | 0 | 32 | 0 | 0 | 0 |
| PDB5 | 0 | 0 | 0 | 16 | 80 | 16 | 0 | 32 |
| PDC1 | 0 | 32 | 16 | 0 | 32 | 0 | 0 | 16 |
| PDC2 | 0 | 0 | 0 | 0 | 16 | 0 | 0 | 0 |
| PDC3 | 0 | 0 | 0 | 0 | 64 | 0 | 0 | 0 |
| PDC4 | 0 | 48 | 0 | 0 | 16 | 0 | 16 | 0 |
| PDC5 | 0 | 0 | 0 | 0 | 96 | 48 | 0 | 0 |
| M1 | 0 | 0 | 32 | 0 | 0 | 16 | 0 | 0 |
| M2 | 0 | 0 | 16 | 0 | 48 | 0 | 0 | 16 |
| M3 | 0 | 0 | 0 | 0 | 128 | 16 | 0 | 16 |



## Why to use multivariate techniques ?

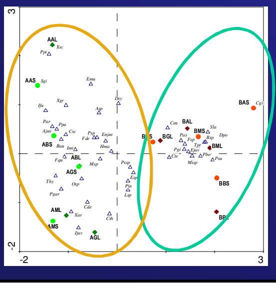
- **The different measurements (variables) can separated into:**
  - **'response variables'**,e.g., number of individuals of different species, microbial parameters, physiological variables (biomarkers), ecotoxicological endpoints, presence-absence of a band (in a DGGE gel) – measure the effect
  - **'explanatory variables'** ,e.g., concentration of chemicals, soil/water chemical and physical variables – they are related to the cause

# Why to use multivariate techniques ?

**Detect and represent the underlying structure of the data**

**(samples vs. response variables)**

***"See the forest out of the trees"***

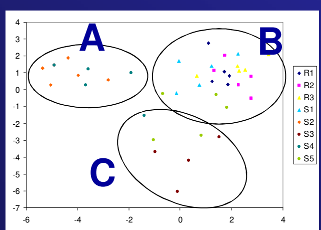




# Why to use multivariate techniques ?

**Discriminate groups**

**(based on the response variables)**

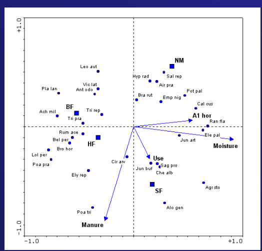



Is group A significantly different from group B or C ?

## Why to use multivariate techniques ?

Relate that structure with the explanatory variables

(response variables vs. explanatory variables)






## Why to use multivariate techniques ?

Have the advantage to analyse all variables simultaneously

## (Some) Available methods

- **Similarity analysis** (similarity or distance indices)
  - Use to evaluate the similarity between samples; these measures can be used afterwards to classify samples into clusters and construct dendrograms or used together with inferential statistics to evaluate or discriminate groups. Ex: ANOSIM.
  - Qualitative or quantitative indices (e.g.,Bray-Curtis index)



## (Some) Ordination methods available

- Reduce the complexicity of the data and represent it into a system of new variables or dimentions – the axes
- Used to represent and interpret the underlying structure of the data
- Examples:
  - Principal Component Analysis (PCA)
  - Correspondence Analysis (CA)

## (Some) Ordination methods available

Samples and species (= response variables) are projected onto a system of axes formed by linear combinations of the original variables where:
- Axis 1 explains a certain amount of variation of the data set
- Axis 2 explains a smaller amount of variation, etc

These new variables (exes) cannot be correlated with each other, otherwise the analysis does not work



## (Some) Ordination methods available

– **Discriminating groups** (samples and response variables)

– Different ways to reach the same end
- Discriminant analysis (DA) – samples are plotted on axes "derived" from the best discriminating variables
- Non-Metric Multidimentional Scalling + ANOSIM (NMDS & ANOSIM) – samples are plotted is a system based on their similarity
- PERMANOVA

## (Some) Ordination methods available

- **Relationship between two data sets** (response variables and explanatory variables)
- Two ways to reach the same end
  - Indirect analysis (e.g., PCA, CA + passive explanatory variables)
  - Direct analyais (RDA, CCA): canonical or constrained analysis



## Indirect vs. Direct Analysis

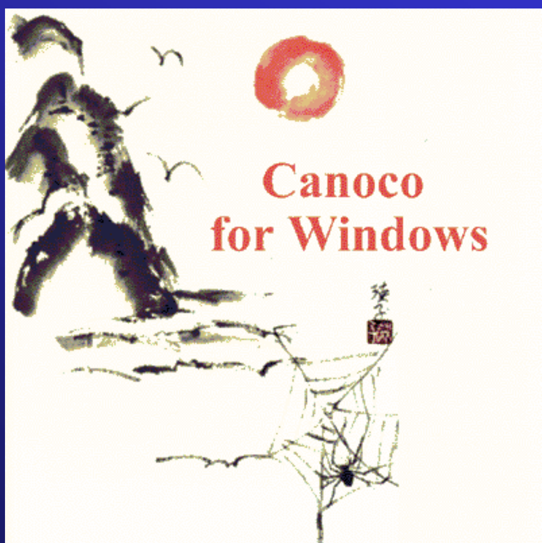
| Indirect | Direct |
|---|---|
| Analysis of total variation | Analysis of total variation |
| ⇩ | ⇩ |
| Samples ⬄ Response variables (e.g. species) | Analysis of the variation explained by the explanatory variables |
| ⇩ | ⇩ |
| Interpretation with explanatory variables using *a posteriori* regression | Samples ⬄ Response variables ⬄ Explanatory variables |

Multivariate Statistical Tools in Ecology
ISCED, Lubango, March 2016

## Ordination Tools I:
## Representing the underlying structure of a dataset
## Part 1



## CANOCO (version 4.5)
### CANOnical Community Ordination



# CANOCO response models

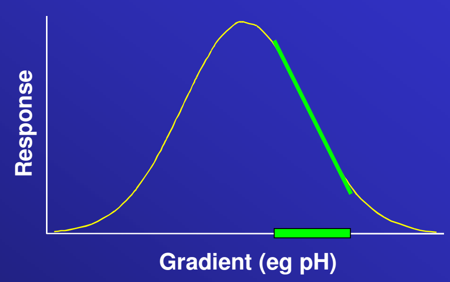
- **Linear**

**PCA, RDA**

Each species assumes a linear response in relation to the axis (gradient); species coordinate in the axis is the slope of that line.

Even in the presence of an unimodal response, if the gradient is small, a linear response is considered

Interpretation: biplot rule






# CANOCO response models

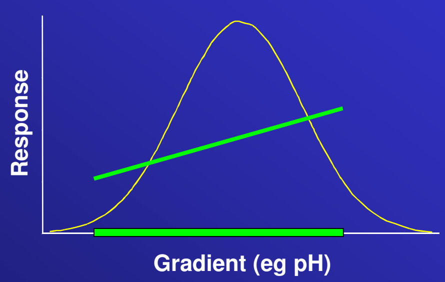
- **Unimodal**

**CA,CCA,DCA**

Each species assumes an unimodal response in relation to the axis (gradient); species coordinate in the axis is the centre of the curve.

This model assumes an environmental optimum for each species

Interpretation: centroid rule



# CANOCO response models

| | Linear response | Unimodal response |
|---|---|---|
| Simple ordination | Principal Component Analysis (PCA) | Correspondence Analysis (CA) |
| Constrained ordination | Redundancy Analysis (RDA) | Canonical Correspondence Analysis (CCA) |

*Which model to choose ?*



# CANOCO response models

## Which model to choose ?

**PCA**

- Linear (grad. < 3 SD)
- Absolute data (analysis of absolute differences)

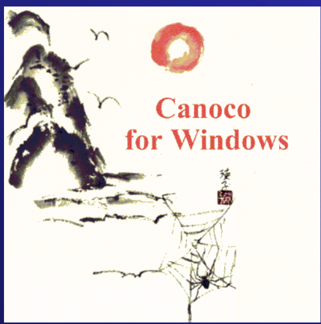
**CA**

- Unimodal (grad. > 4 SD)
- Relative data (analysis of compositional differences)

- Unimodal methods cannot be used when response variables are in different units. Also no empty samples allowed
- Variability explained by axes (questionable since way to calculate total variation is different between methods)

# CANOCO response models

## Which model to choose ?




**Steps in CANOCO !**




# CANOCO response models

## Soil fauna in different planting systems in Brazil (Adriana Aquino – EMBRAPA, RJ)

- Data on soil macrofauna (File: DataAdrEx1.xls),
- 5 areas with different treatments: woods (M), direct plantation A, B e C (PDA, PDB, PDC), conventional plantation (C);
- 23 samples collected in all areas; 24 taxonomic groups;
- Aim: to verify the association between fauna and treatments

# CANOCO for Windows (v. 4.5)

- CanoImp ➲ creating input files
- CANOCO ➲ analysis
- CanoDraw ➲ creating & editing diagrams + aditional options

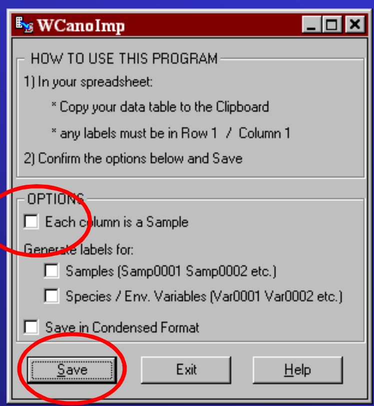
© JPSousa

# CANOCO for Windows - DCA

CanoImp

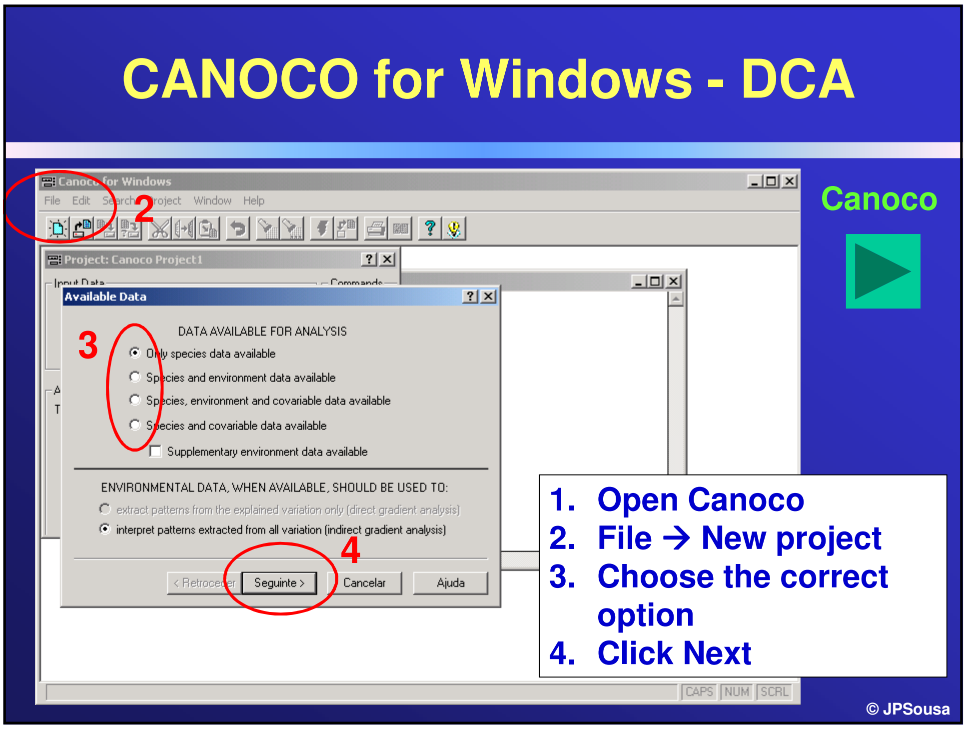
- Example Adriana (data input)

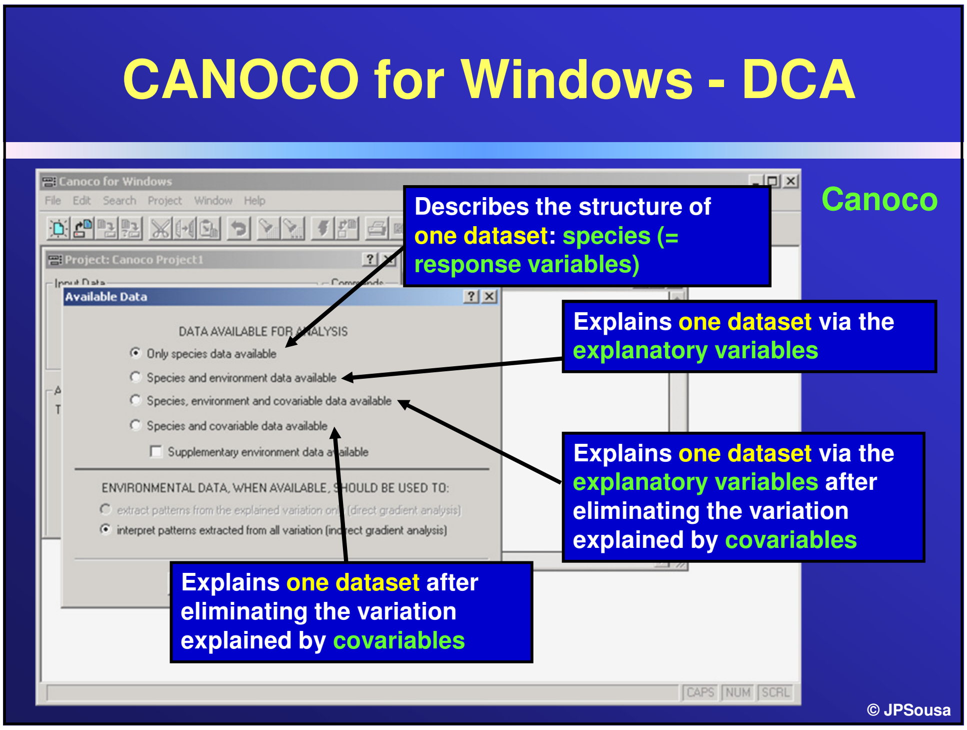
1. Copy Excel matrix (species matrix)
2. Open CanoImp
3. Click "Save"
4. Name file (do not forget the extension .dta)

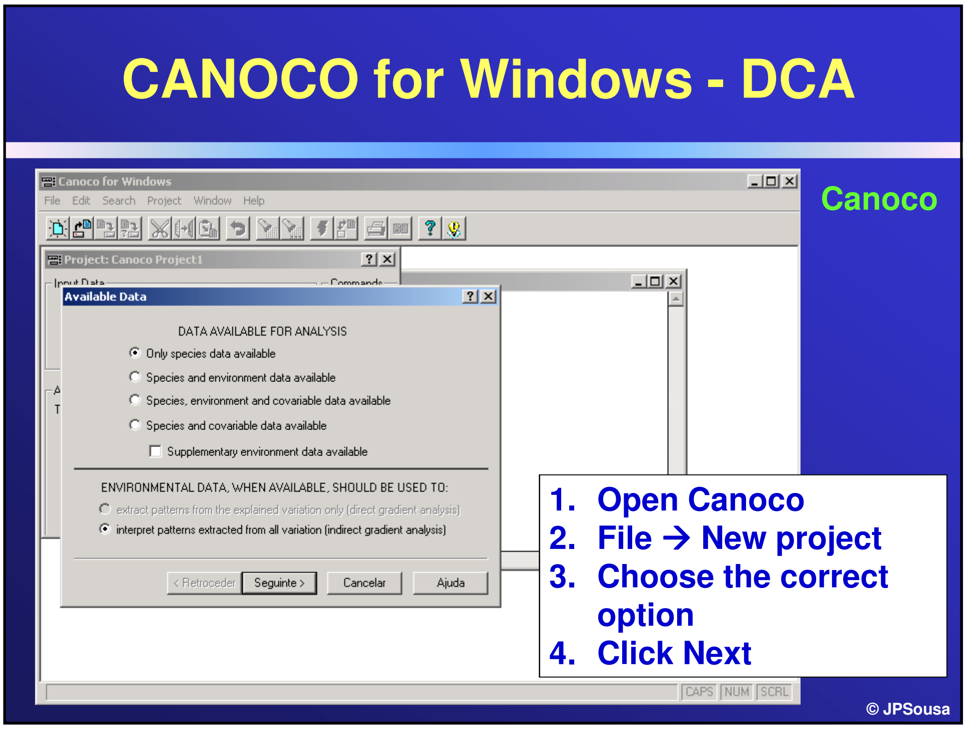


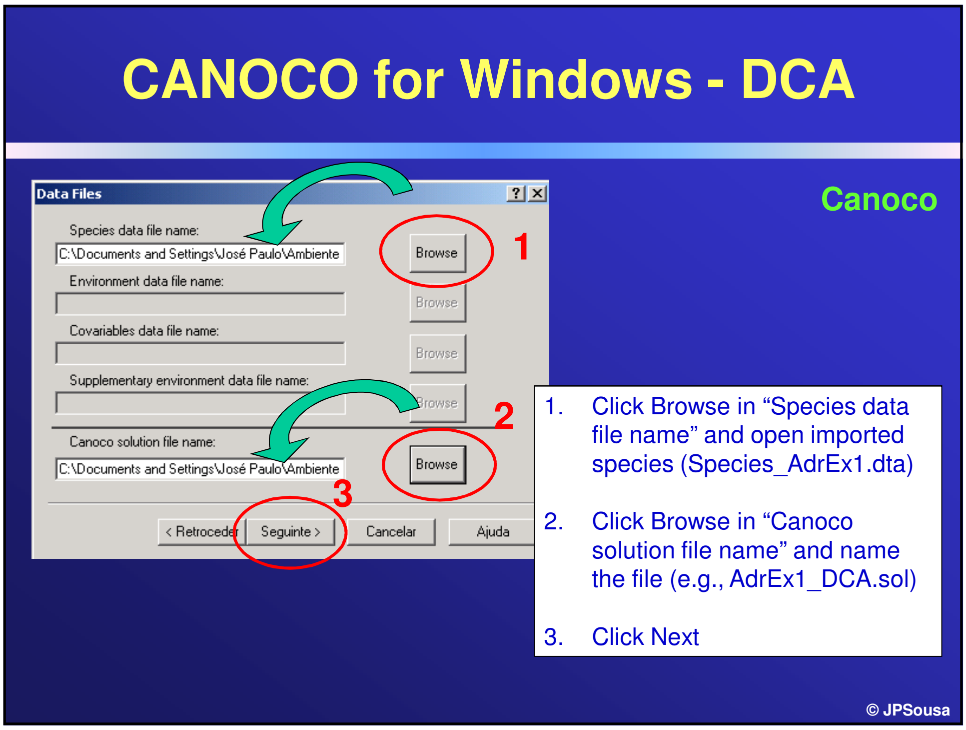

CanoImp Excel

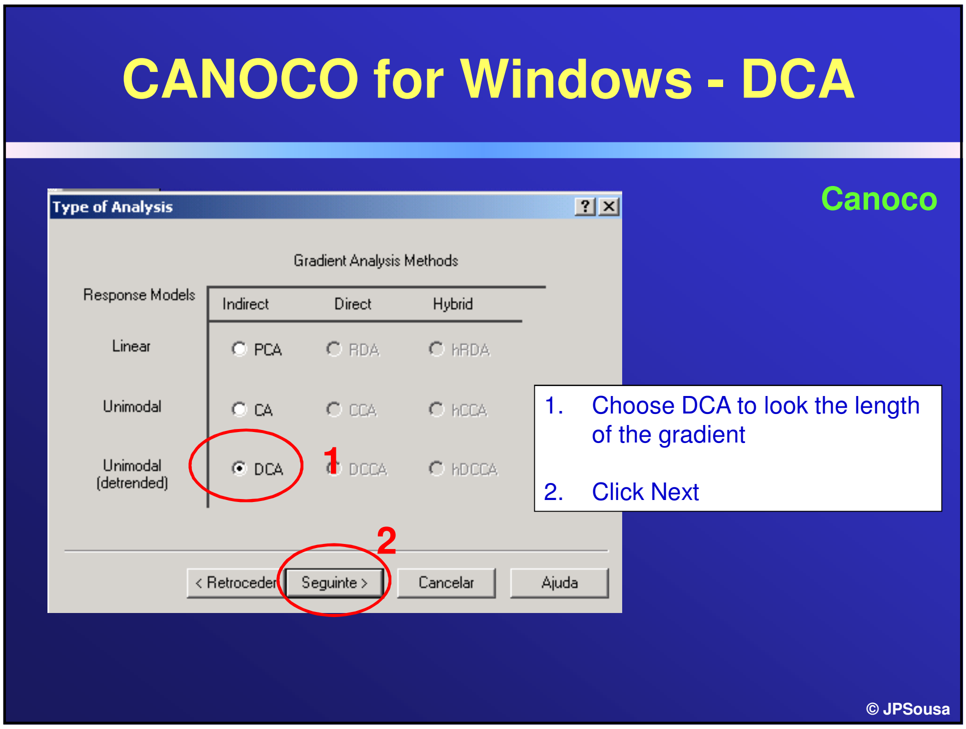
© JPSousa

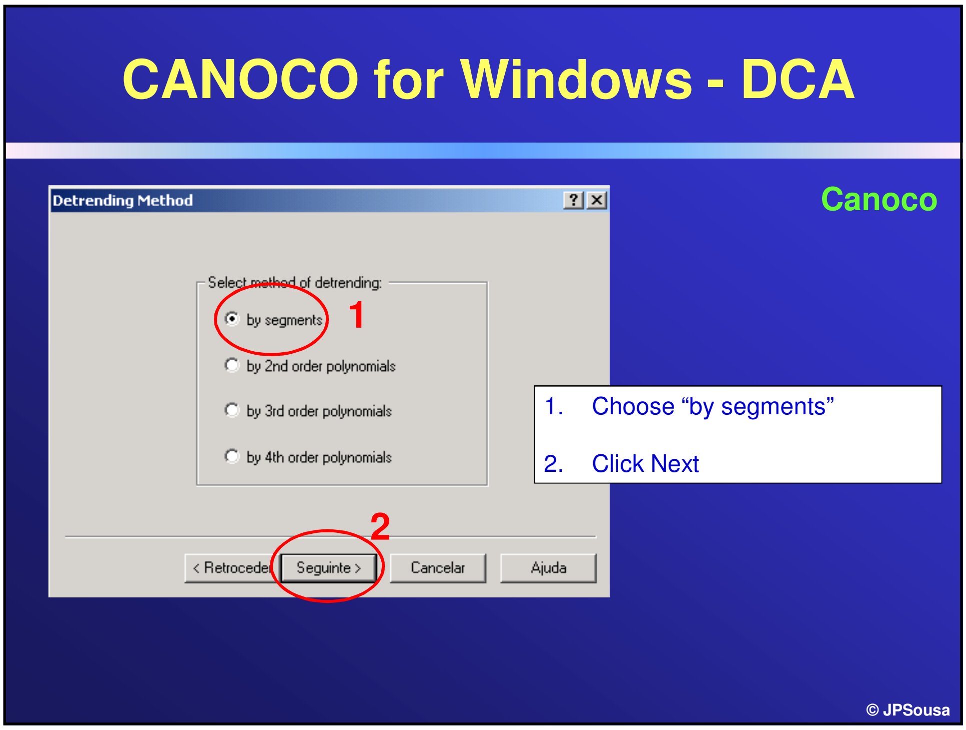

CANOCO for Windows - DCA
Canoco
Canoco for Windows
Available Data
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
2
3
4
1. Open Canoco
2. File → New project
3. Choose the correct option
4. Click Next
© JPSousa


CANOCO for Windows - DCA
Canoco
Describes the structure of one dataset: species (= response variables)
Explains one dataset via the explanatory variables
Explains one dataset via the explanatory variables after eliminating the variation explained by covariables
Explains one dataset after eliminating the variation explained by covariables
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
© JPSousa


CANOCO for Windows - DCA
Canoco
Canoco for Windows
File Edit Search Project Window Help
Project: Canoco Project1
Available Data
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Open Canoco
2. File → New project
3. Choose the correct option
4. Click Next
CAPS NUM SCRL
© JPSousa


CANOCO for Windows - DCA
Canoco
Data Files
Species data file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
1
Environment data file name:
Browse
Covariables data file name:
Browse
Supplementary environment data file name:
Browse
2
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
3
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Click Browse in "Species data file name" and open imported species (Species_AdrEx1.dta)
2. Click Browse in "Canoco solution file name" and name the file (e.g., AdrEx1_DCA.sol)
3. Click Next
© JPSousa


CANOCO for Windows - DCA
Canoco
Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
1
2
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Choose DCA to look the length of the gradient
2. Click Next
© JPSousa


CANOCO for Windows - DCA
Canoco
Detrending Method
Select method of detrending:
by segments
1
by 2nd order polynomials
by 3rd order polynomials
by 4th order polynomials
2
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Choose “by segments”
2. Click Next
© JPSousa

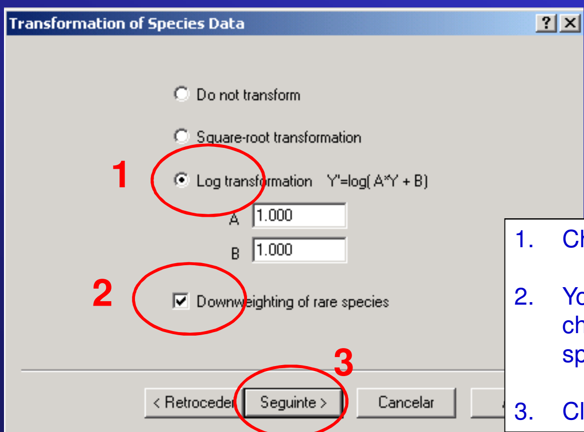
# CANOCO for Windows

## Methods of “detrending”

- The length of the gradient can only be obtained via “Detrending by segments”;
- In case a “normal” DCA or DCCA is wanted, choose “detrending by polynomials” – the second axis becomes not correlated with the first axis, preventing the Gutman (arch) effect and the compression of the ends of the gradient.



# CANOCO for Windows - DCA




Canoco

1. Choose “log transformation”
2. You have the possibility to choose “downweighting rare species” (in unimodal models)
3. Click Next

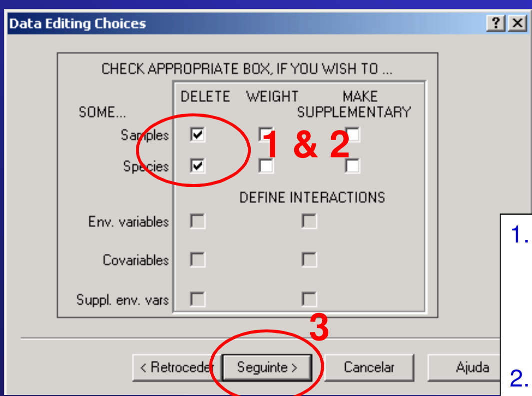
# CANOCO for Windows

## Transformation

- Advisable when the distribution of species abundances is highly skewed. Prevents some few species to dominate the analysis.
- In unimodal response models rare species can influence the results. They should be downweighted.

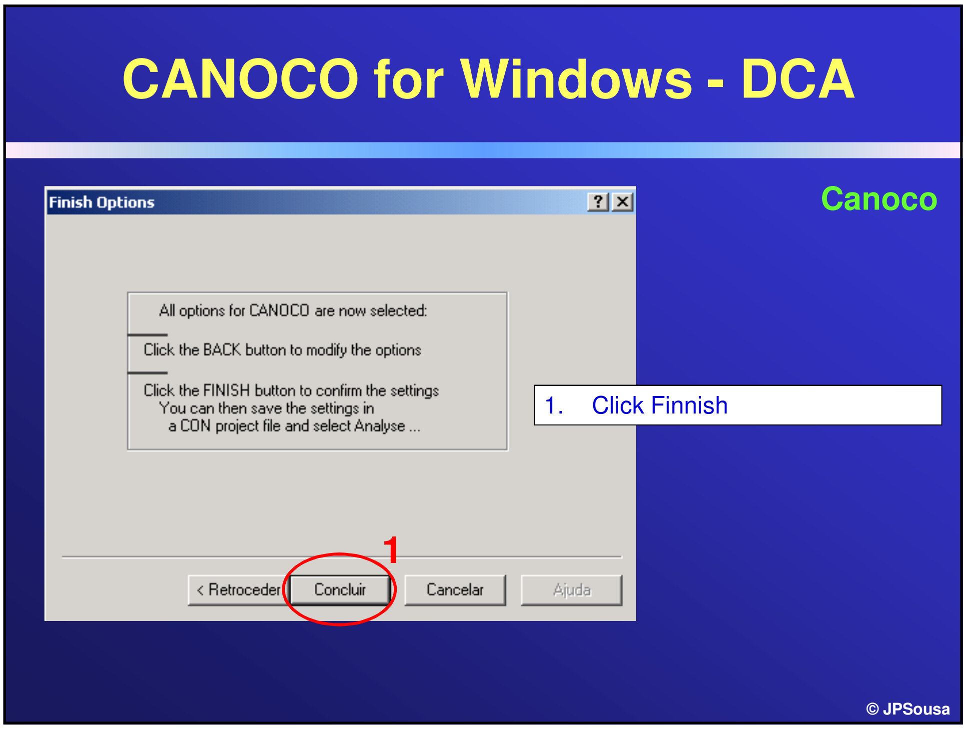


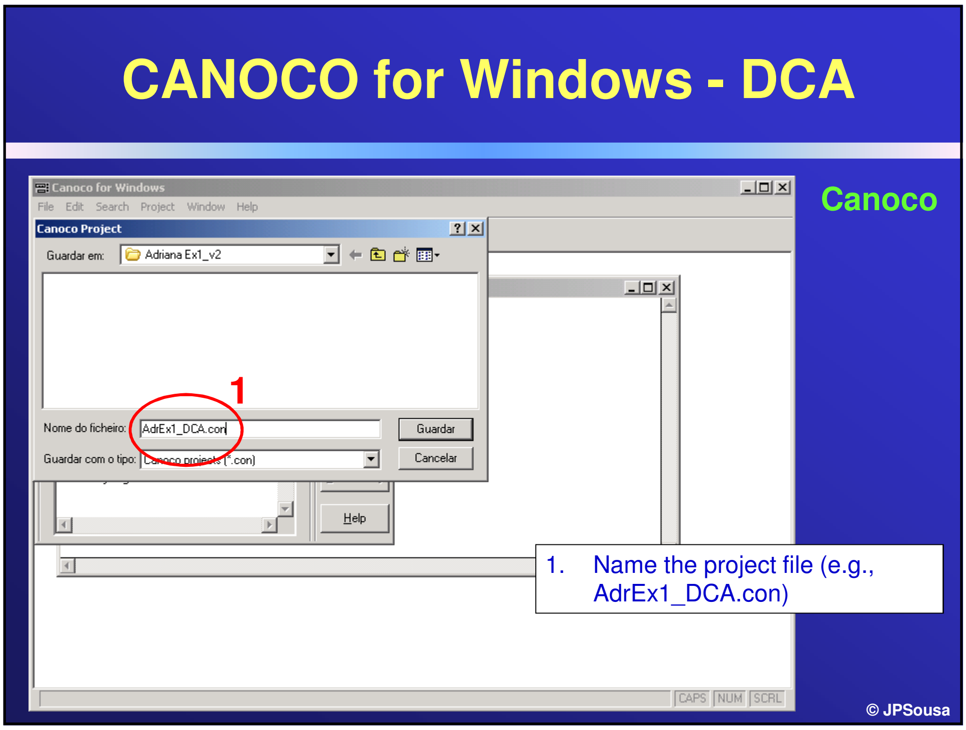
# CANOCO for Windows - DCA

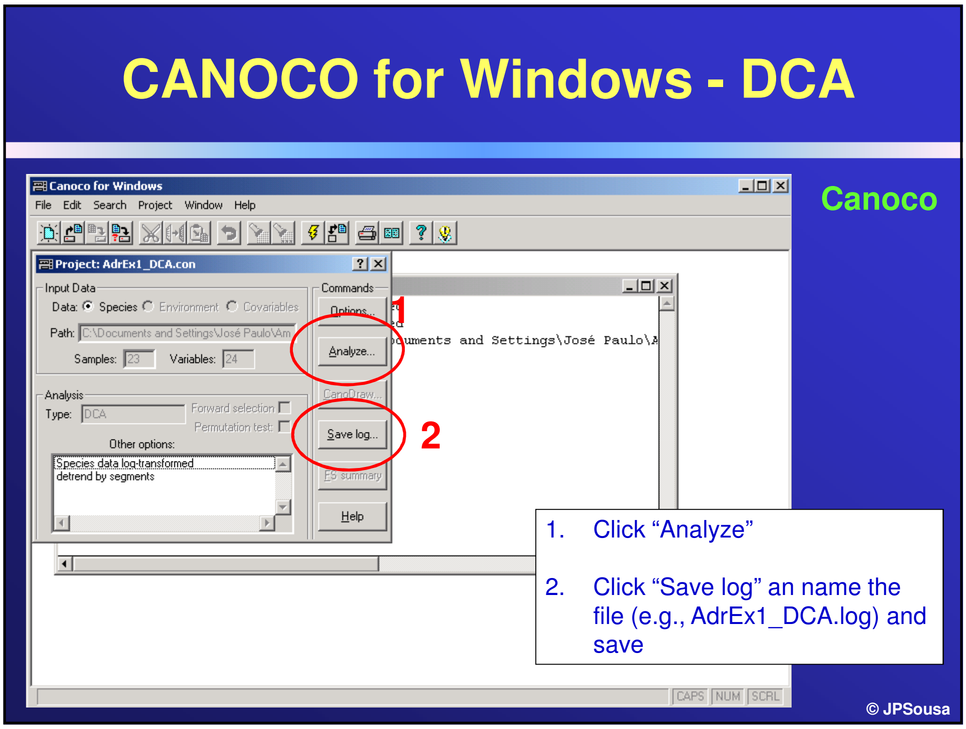


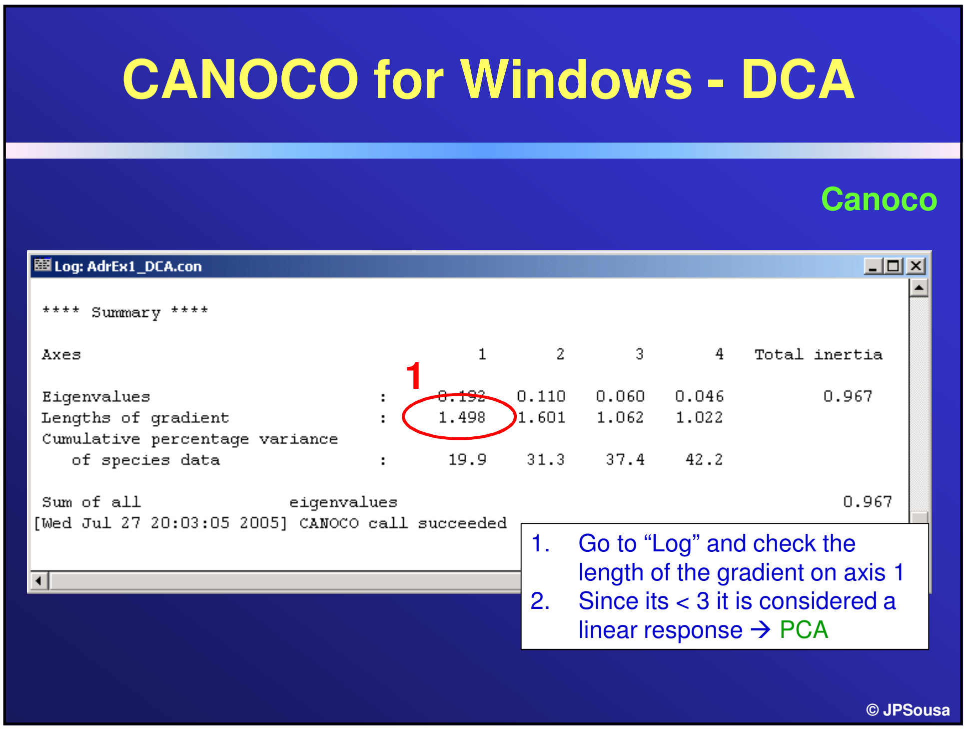

Canoco

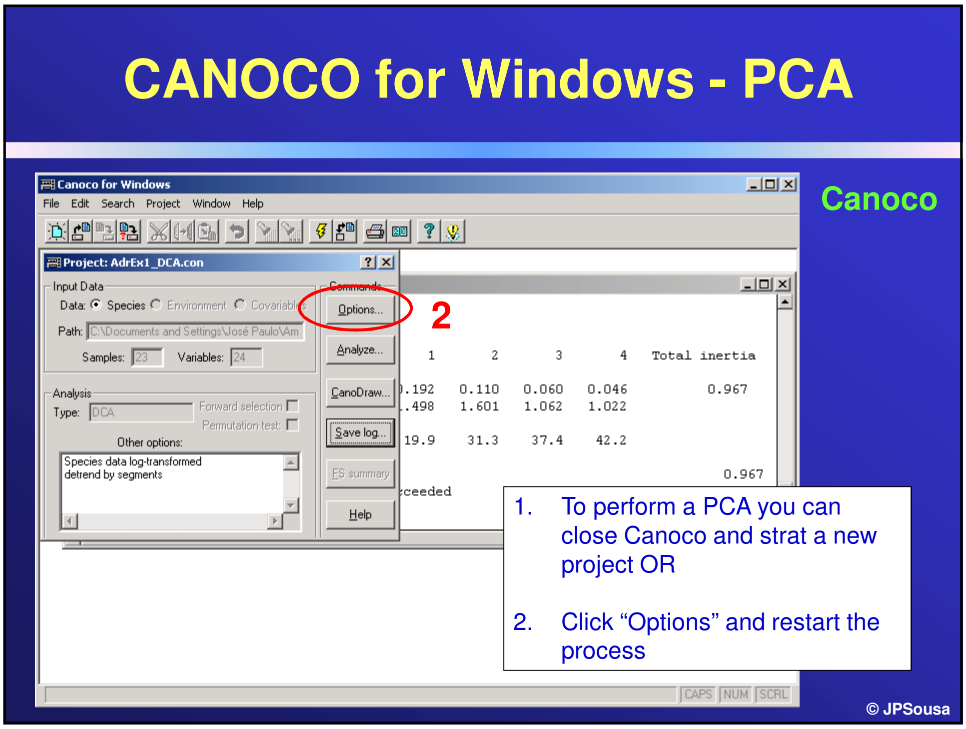
1. This option allows to eliminate "species" or "samples" from the analysis without the need to restart the process.
2. In this case all species and samples should be in, so, do not select this option
3. Click Next

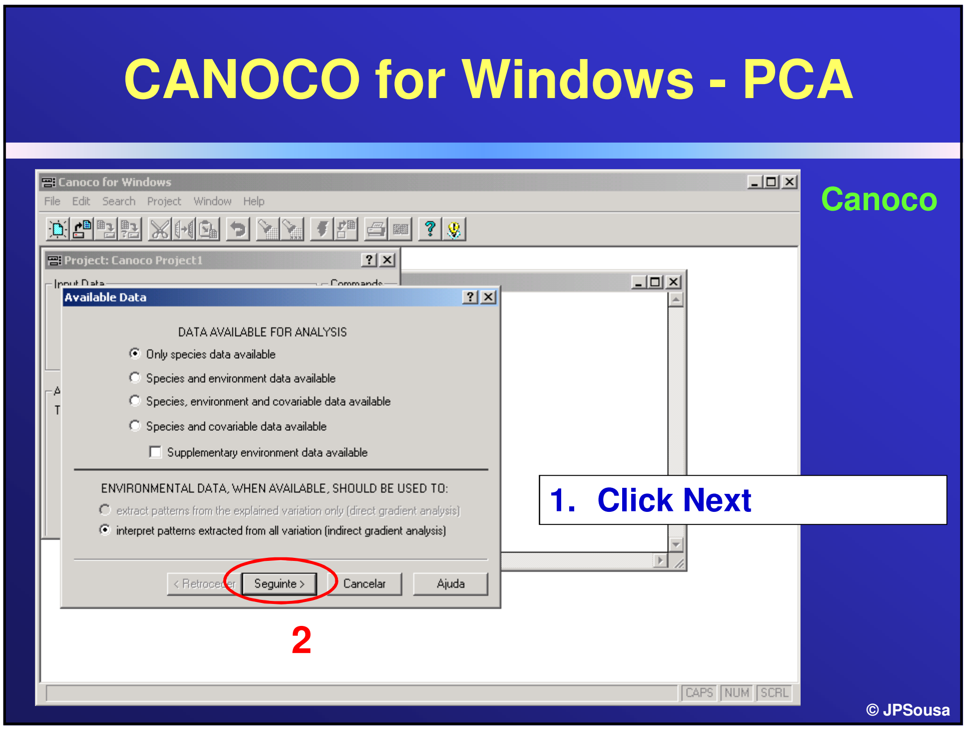

CANOCO for Windows - DCA
Canoco
Finish Options
All options for CANOCO are now selected:
Click the BACK button to modify the options
Click the FINISH button to confirm the settings
You can then save the settings in
a CON project file and select Analyse ...
1
< Retroceder
Concluir
Cancelar
Ajuda
1. Click Finnish
© JPSousa

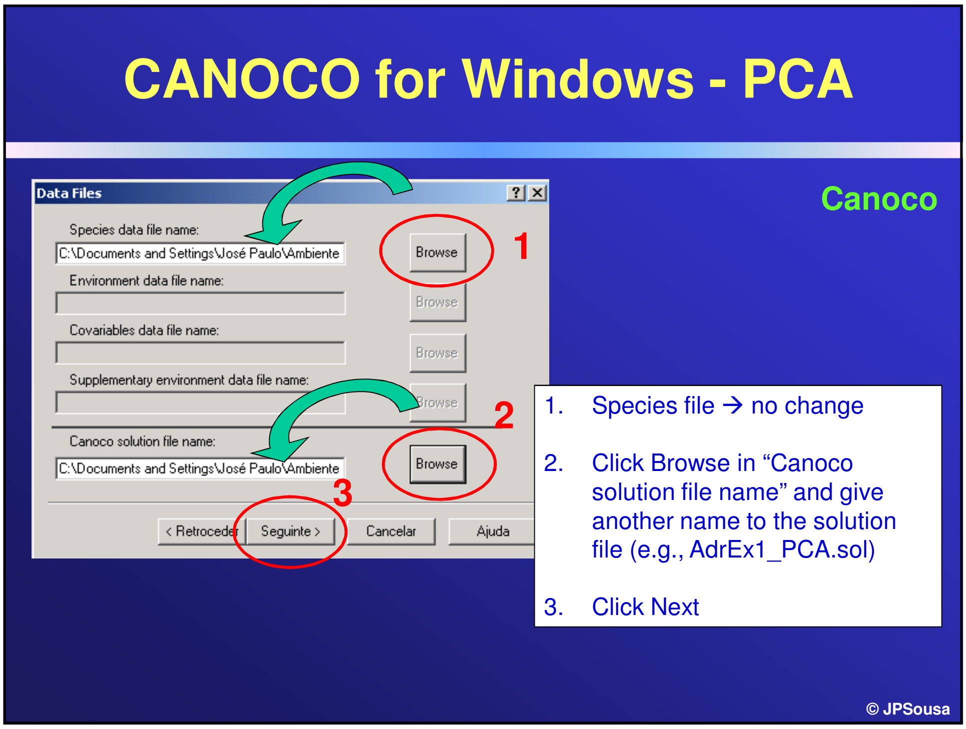

CANOCO for Windows - DCA
Canoco
Canoco for Windows
File Edit Search Project Window Help
Canoco Project
Guardar em: Adriana Ex1_v2
1
Nome do ficheiro: AdrEx1_DCA.con
Guardar
Guardar com o tipo: Canoco projects (*.con)
Cancelar
Help
1. Name the project file (e.g., AdrEx1_DCA.con)
CAPS NUM SCRL
© JPSousa

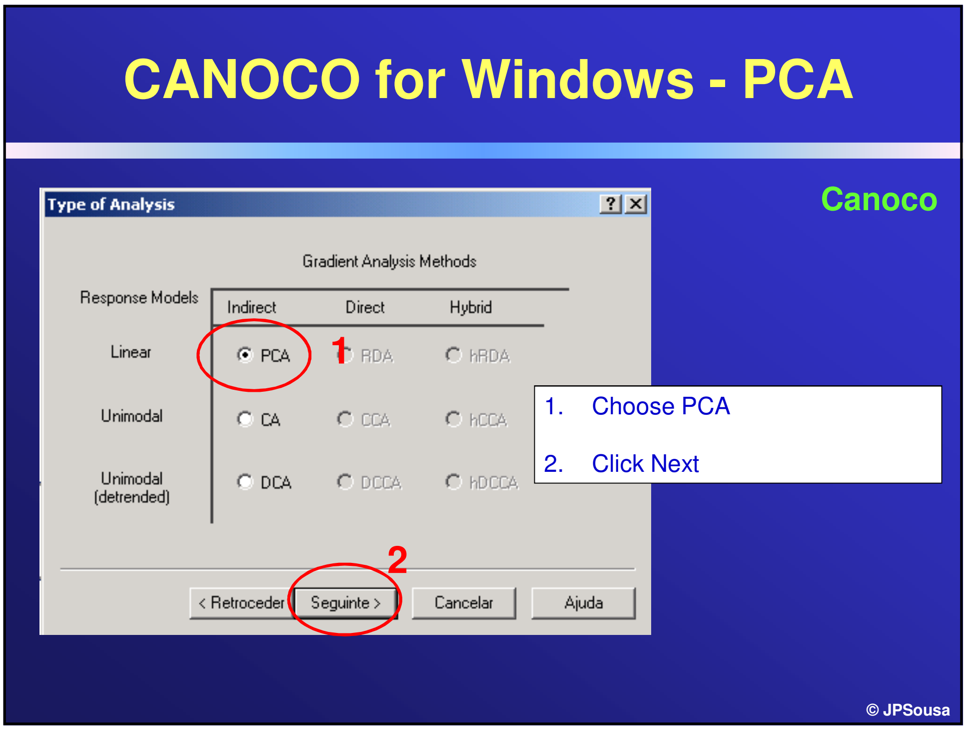

CANOCO for Windows - DCA
Canoco
Canoco for Windows
File Edit Search Project Window Help
Project: AdrEx1_DCA.con
Input Data
Data: Species Environment Covariables
Path: C:\Documents and Settings\José Paulo\Am
Samples: 23 Variables: 24
Analysis
Type: DCA
Forward selection
Permutation test:
Other options:
Species data log-transformed
detrend by segments
Commands
Options...
Analyze...
CanoDraw...
Save log...
F5 summary
Help
ocuments and Settings\José Paulo\A
CAPS NUM SCRL
1
2
1. Click "Analyze"
2. Click "Save log" an name the file (e.g., AdrEx1_DCA.log) and save
© JPSousa


CANOCO for Windows - DCA
Canoco
Log: AdrEx1_DCA.con
**** Summary ****
Axes 1 2 3 4 Total inertia
Eigenvalues : 0.192 0.110 0.060 0.046 0.967
Lengths of gradient : 1.498 1.601 1.062 1.022
Cumulative percentage variance
of species data : 19.9 31.3 37.4 42.2
Sum of all eigenvalues 0.967
[Wed Jul 27 20:03:05 2005] CANOCO call succeeded
1
1. Go to "Log" and check the length of the gradient on axis 1
2. Since its < 3 it is considered a linear response → PCA
© JPSousa

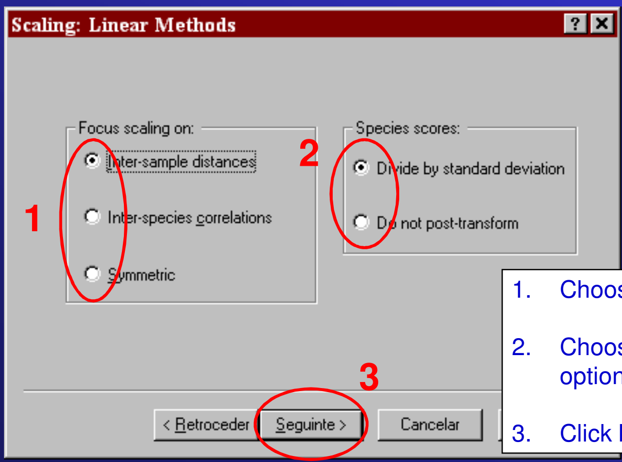

CANOCO for Windows - PCA
Canoco
Canoco for Windows
File Edit Search Project Window Help
Project: AdrEx1_DCA.con
Input Data
Data: Species Environment Covariables
Path: C:\Documents and Settings\José Paulo\Am
Samples: 23 Variables: 24
Analysis
Type: DCA
Forward selection
Permutation test
Other options:
Species data log-transformed
detrend by segments
Commands
Options...
Analyze...
CanoDraw...
Save log...
FS summary
Help
2
1 2 3 4 Total inertia
.192 0.110 0.060 0.046 0.967
.498 1.601 1.062 1.022
19.9 31.3 37.4 42.2
0.967
ceeded
1. To perform a PCA you can close Canoco and strat a new project OR
2. Click "Options" and restart the process
CAPS NUM SCRL
© JPSousa


CANOCO for Windows - PCA
Canoco
Canoco for Windows
File Edit Search Project Window Help
Project: Canoco Project1
Available Data
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
2
1. Click Next
CAPS NUM SCRL
© JPSousa


CANOCO for Windows - PCA
Canoco
Data Files
Species data file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
Environment data file name:
Browse
Covariables data file name:
Browse
Supplementary environment data file name:
Browse
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
< Retroceder
Seguinte >
Cancelar
Ajuda
1
2
3
1. Species file → no change
2. Click Browse in "Canoco solution file name" and give another name to the solution file (e.g., AdrEx1_PCA.sol)
3. Click Next
© JPSousa


CANOCO for Windows - PCA
Canoco
Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
< Retroceder
Seguinte >
Cancelar
Ajuda
1
2
1. Choose PCA
2. Click Next
© JPSousa

# CANOCO for Windows - PCA




1. Choose the "Scaling" option
2. Choose the "Species scores" option
3. Click Next



# CANOCO for Windows

## Scaling

- Ton interpret the relations between
  - **Inter-sample** (e.g. ecotoxicological data)
  - **Inter-species** (e.g. niche studies and correlation with explanatory variables)
  - **or simetric scaling**
- It is not important if the eigenvalues of the axes of importance (generally the first two) are similar

# CANOCO for Windows

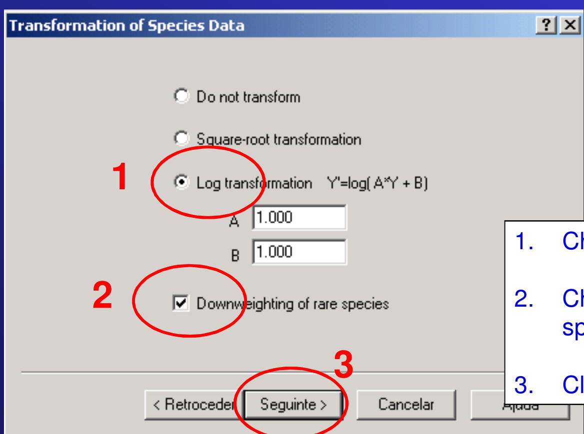
## Species scores

- **Not transformed:**
  - Species scores are proportional to species SDs, therefore species with large variance (usually the dominant species) will dominate the diagram ➲ covariance diagram (the length of the species arrow indicates the variability of that species in the ordination space)
- **Transformed:**
  - Species scores more comparable ➲ correlation diagram (the length of a species arrow measures the fit with the ordination axes)



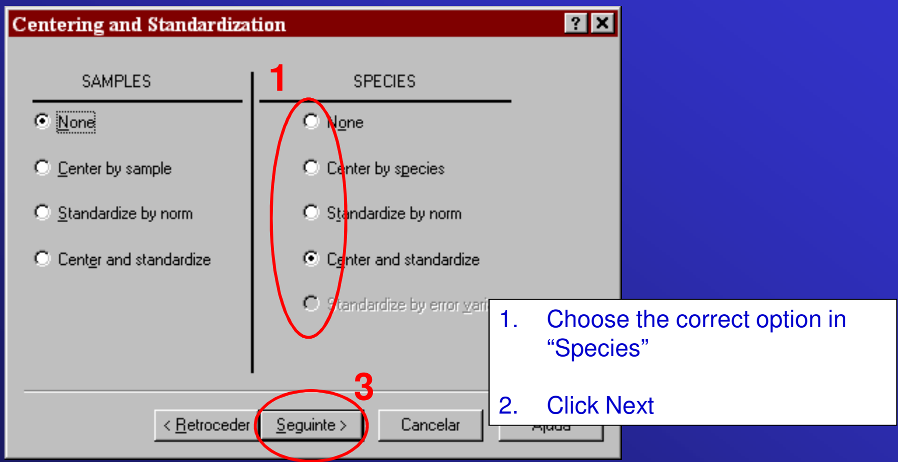
# CANOCO for Windows - PCA

Canoco




1. Choose "log transformation"
2. Choose "downweighting rare species" (in unimodal models)
3. Click Next

# CANOCO for Windows - PCA






# CANOCO for Windows

## Center

- Center by species:
  $y_{ki}^{*} = y_{ki} - y_{k+}/n$
  Mean = 0, SD does not change

- PCA based in the <u>covariance matrix</u> (Common PCA) ➲ center by species

# CANOCO for Windows

## Standardize

- Center and standardiz by species:
  $y_{ki}^{*} = (y_{ki} - y_{mean})/s_k$
  Mean = 0, SD = 1

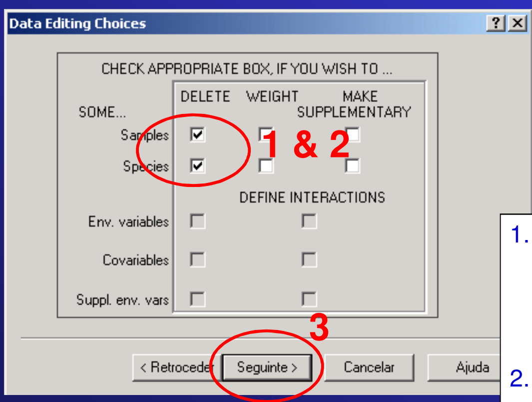
- PCA based in a <u>correlation matrix</u>

- Makes all “species” equally important
  (e.g. pH, $O_2$, nutrientes, etc)

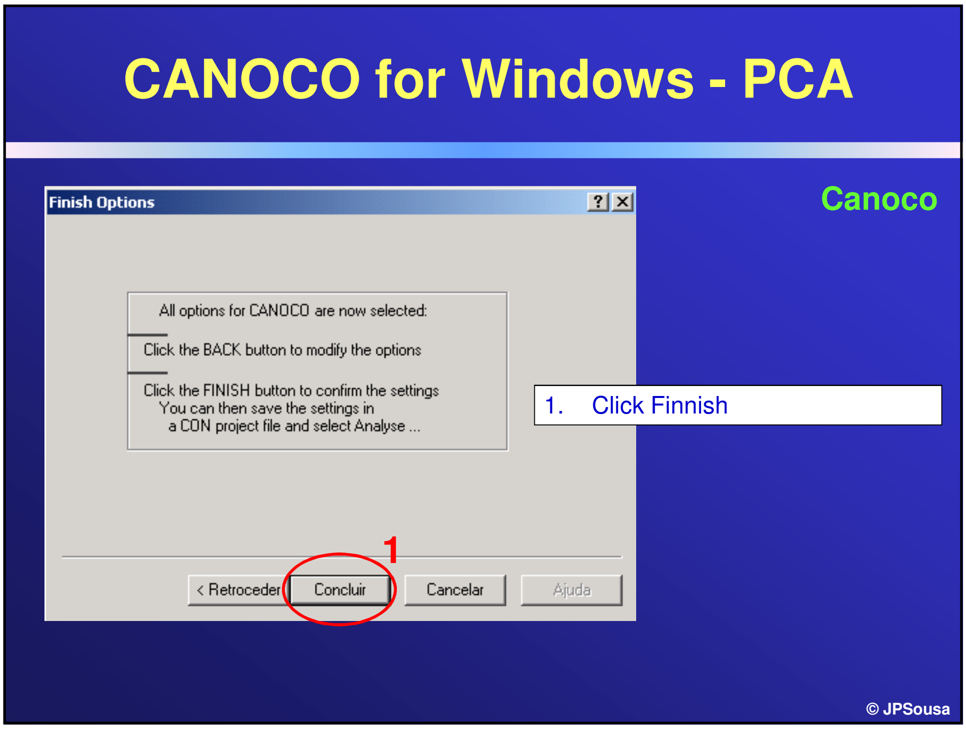


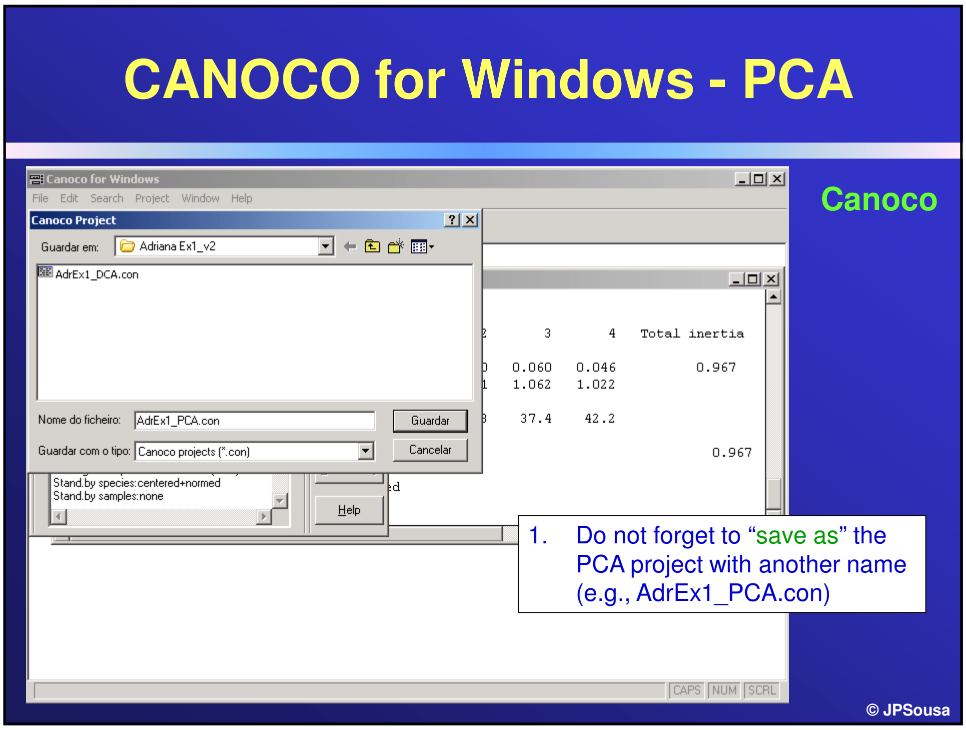
# CANOCO for Windows - PCA

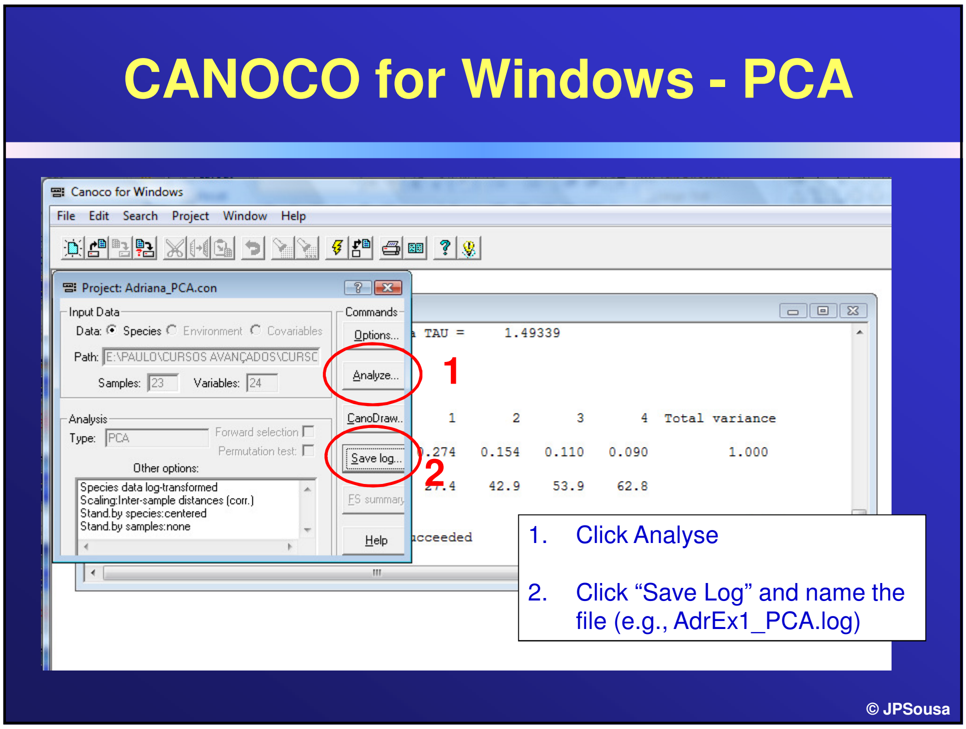
Canoco

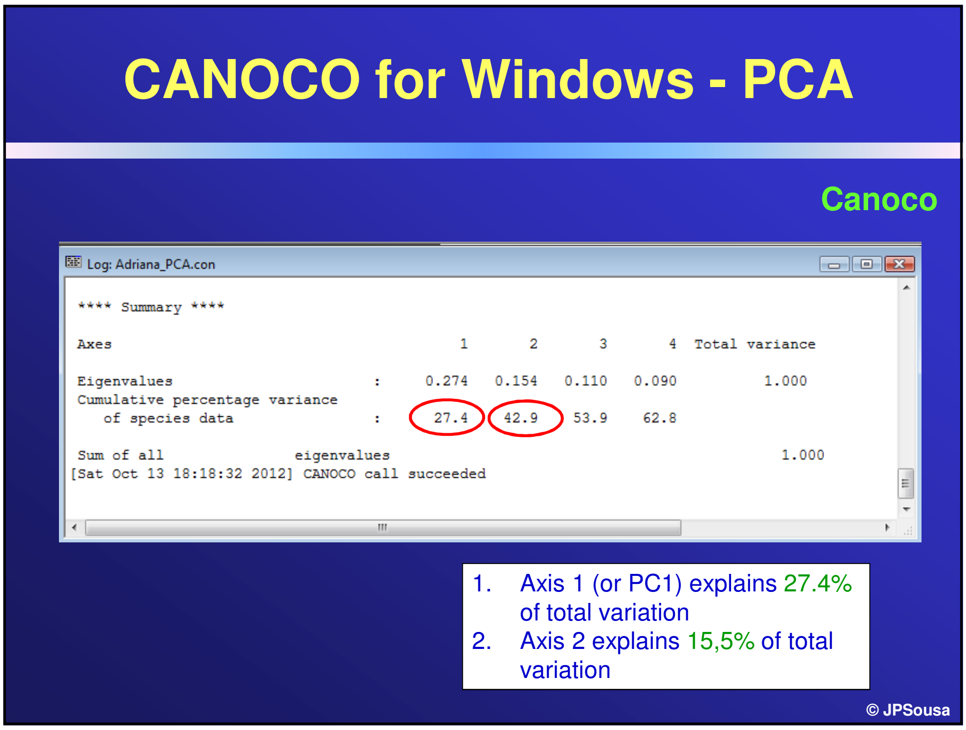


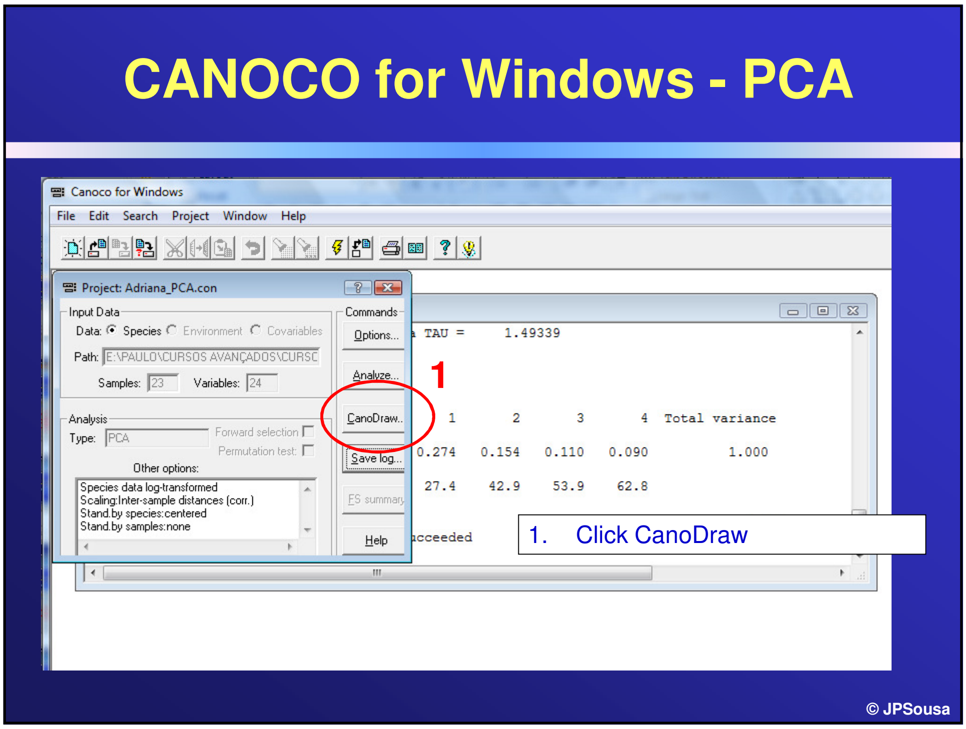

1. This option allows to eliminate “species” or “samples” from the analysis without the need to restart the process.
2. In this case all species and samples should be in, so, do not select this option
3. Click Next

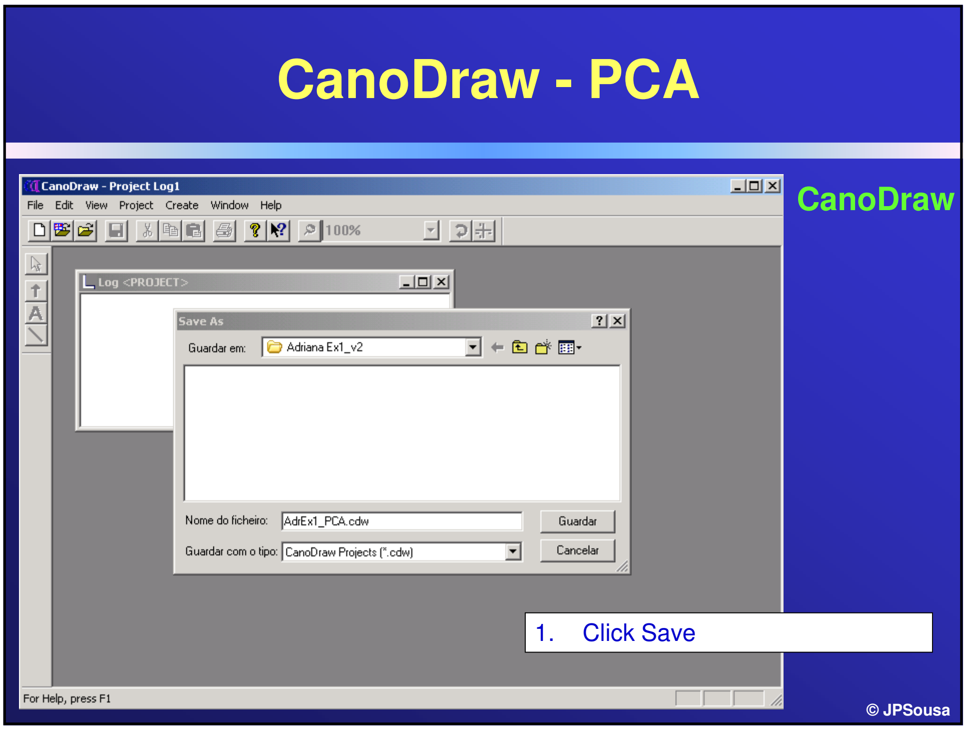

CANOCO for Windows - PCA
Canoco
Finish Options
All options for CANOCO are now selected:
Click the BACK button to modify the options
Click the FINISH button to confirm the settings
You can then save the settings in
a CON project file and select Analyse ...
1
< Retroceder
Concluir
Cancelar
Ajuda
1. Click Finnish
© JPSousa

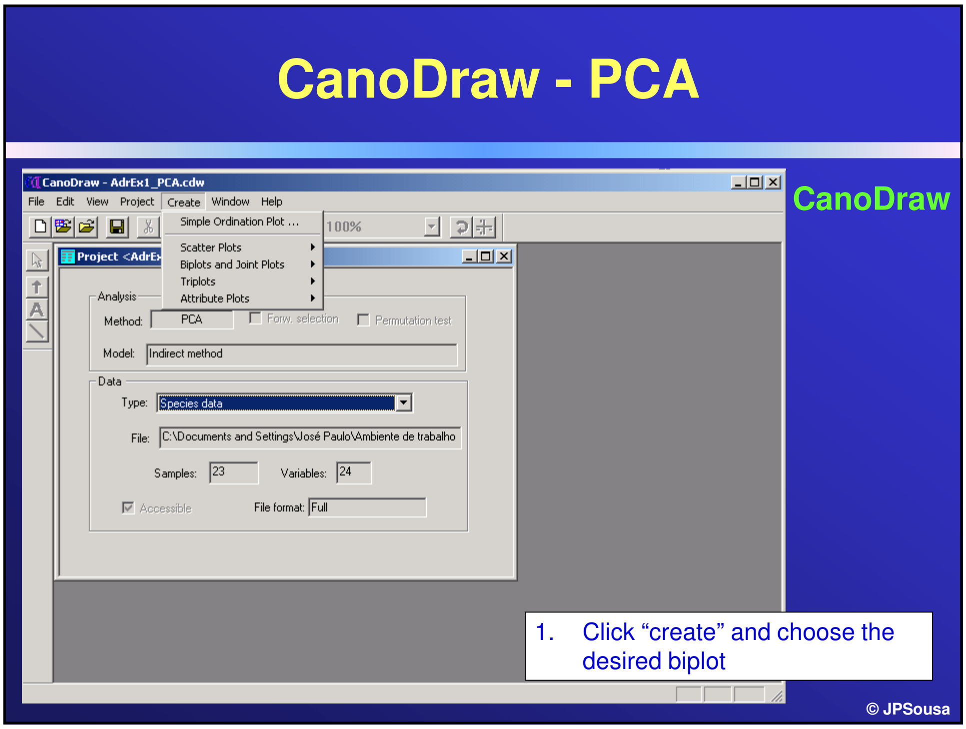

CANOCO for Windows - PCA
Canoco
Canoco for Windows
File Edit Search Project Window Help
Canoco Project
Guardar em: Adriana Ex1_v2
AdrEx1_DCA.con
Nome do ficheiro: AdrEx1_PCA.con
Guardar
Guardar com o tipo: Canoco projects (*.con)
Cancelar
3 4 Total inertia
0.060 0.046 0.967
1.062 1.022
37.4 42.2
0.967
Stand.by species:centered+normed
Stand.by samples:none
Help
1. Do not forget to "save as" the PCA project with another name (e.g., AdrEx1_PCA.con)
CAPS NUM SCRL
© JPSousa

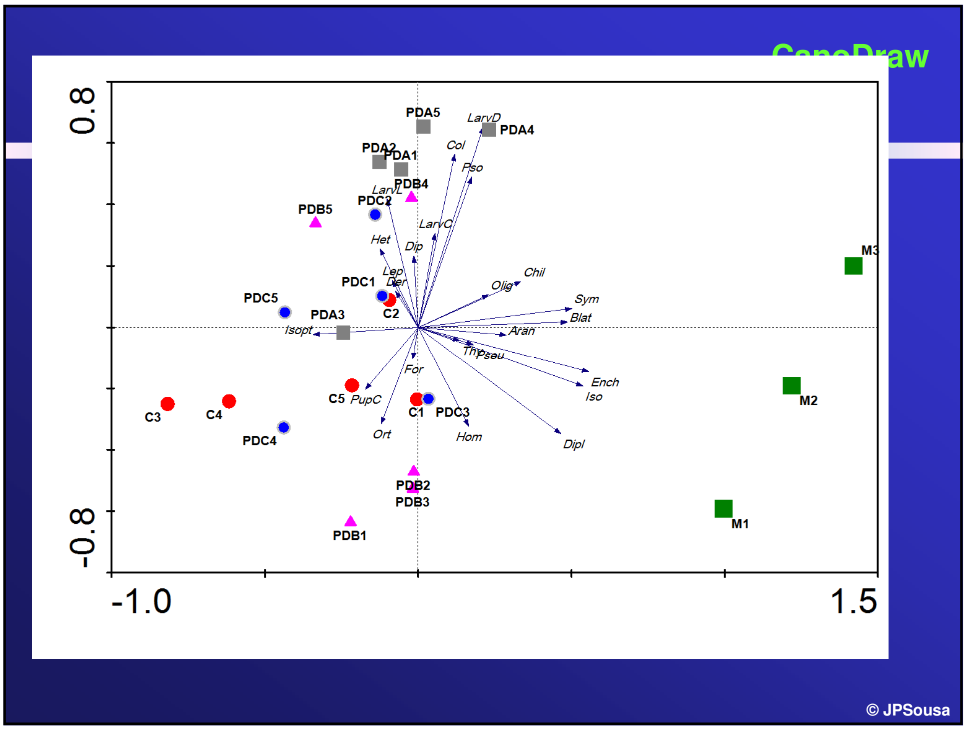

CANOCO for Windows - PCA
Canoco for Windows
File Edit Search Project Window Help
Project: Adriana_PCA.con
Input Data
Data: Species Environment Covariables
Path: E:\PAULO\CURSOS AVANÇADOS\CURSC
Samples: 23 Variables: 24
Analysis
Type: PCA
Forward selection
Permutation test:
Other options:
Species data log-transformed
Scaling:Inter-sample distances (corr.)
Stand.by species:centered
Stand.by samples:none
Commands
Options...
Analyze...
CanoDraw...
Save log...
FS summary
Help
1
2
TAU = 1.49339
1 2 3 4 Total variance
0.274 0.154 0.110 0.090 1.000
27.4 42.9 53.9 62.8
succeeded
1. Click Analyse
2. Click "Save Log" and name the file (e.g., AdrEx1_PCA.log)
© JPSousa

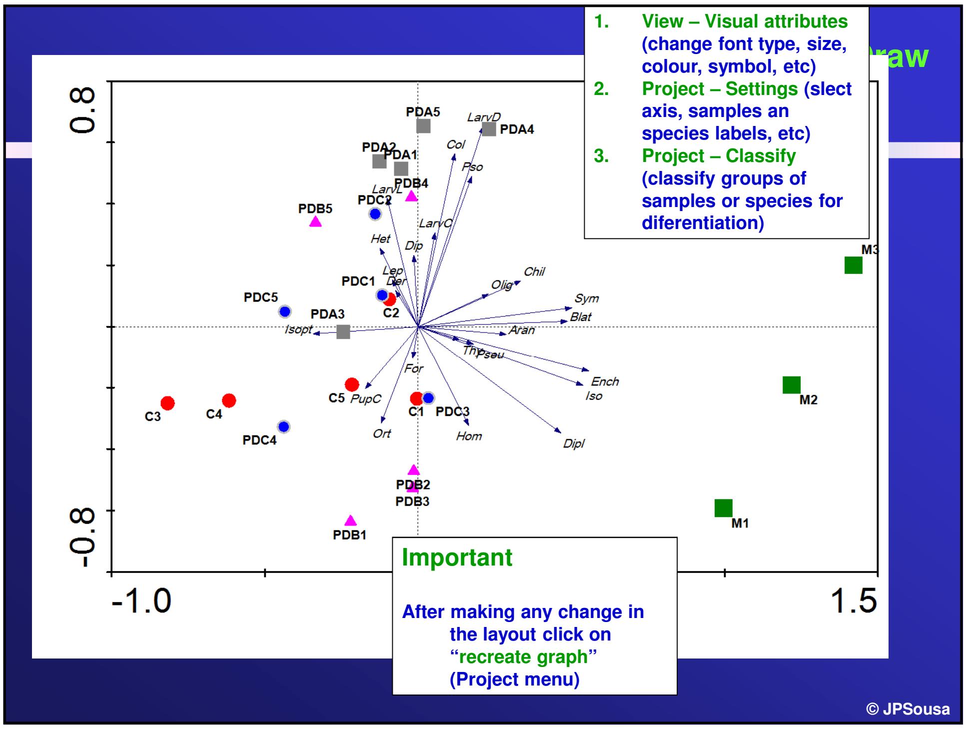

CANOCO for Windows - PCA
Canoco
Log: Adriana_PCA.con
**** Summary ****
Axes 1 2 3 4 Total variance
Eigenvalues : 0.274 0.154 0.110 0.090 1.000
Cumulative percentage variance
of species data : 27.4 42.9 53.9 62.8
Sum of all eigenvalues 1.000
[Sat Oct 13 18:18:32 2012] CANOCO call succeeded
1. Axis 1 (or PC1) explains 27.4% of total variation
2. Axis 2 explains 15,5% of total variation
© JPSousa

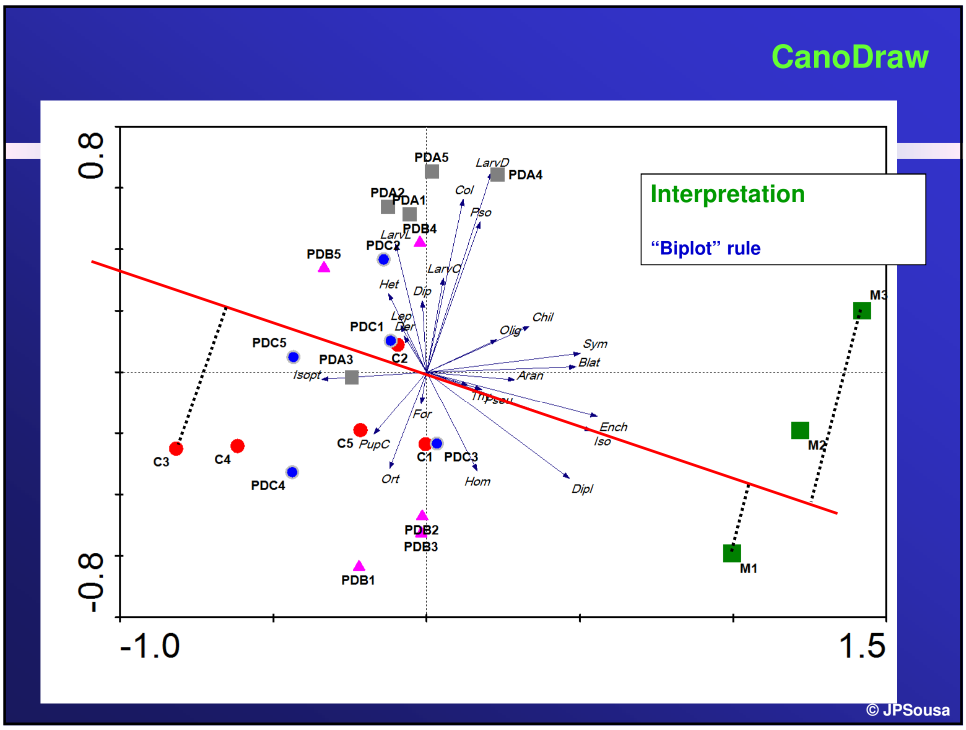

CANOCO for Windows - PCA
Canoco for Windows
File Edit Search Project Window Help
Project: Adriana_PCA.con
Input Data
Data: Species Environment Covariables
Path: E:\PAULO\CURSOS AVANÇADOS\CURSC
Samples: 23 Variables: 24
Analysis
Type: PCA
Forward selection
Permutation test:
Other options:
Species data log-transformed
Scaling:Inter-sample distances (corr.)
Stand.by species:centered
Stand.by samples:none
Commands
Options...
Analyze...
CanoDraw...
Save log...
FS summary
Help
TAU = 1.49339
1
1 2 3 4 Total variance
0.274 0.154 0.110 0.090 1.000
27.4 42.9 53.9 62.8
cceeded
1. Click CanoDraw
© JPSousa

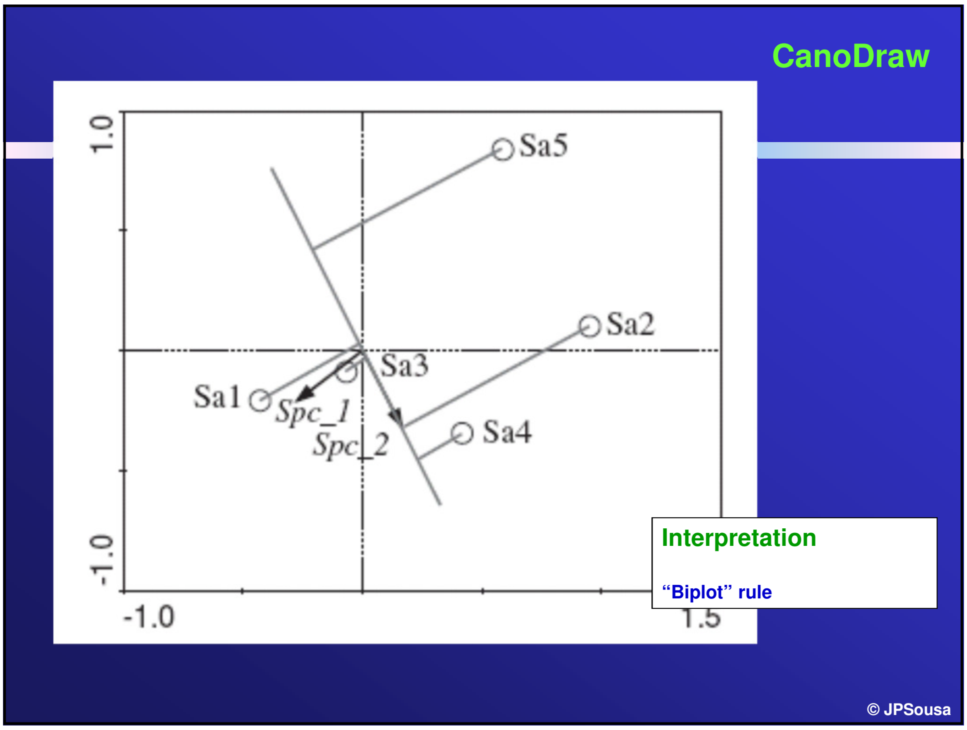

CanoDraw - PCA
CanoDraw - Project Log1
File Edit View Project Create Window Help
100%
Log <PROJECT>
Save As
Guardar em: Adriana Ex1_v2
Nome do ficheiro: AdrEx1_PCA.cdw
Guardar
Guardar com o tipo: CanoDraw Projects (*.cdw)
Cancelar
For Help, press F1
CanoDraw
1. Click Save
© JPSousa

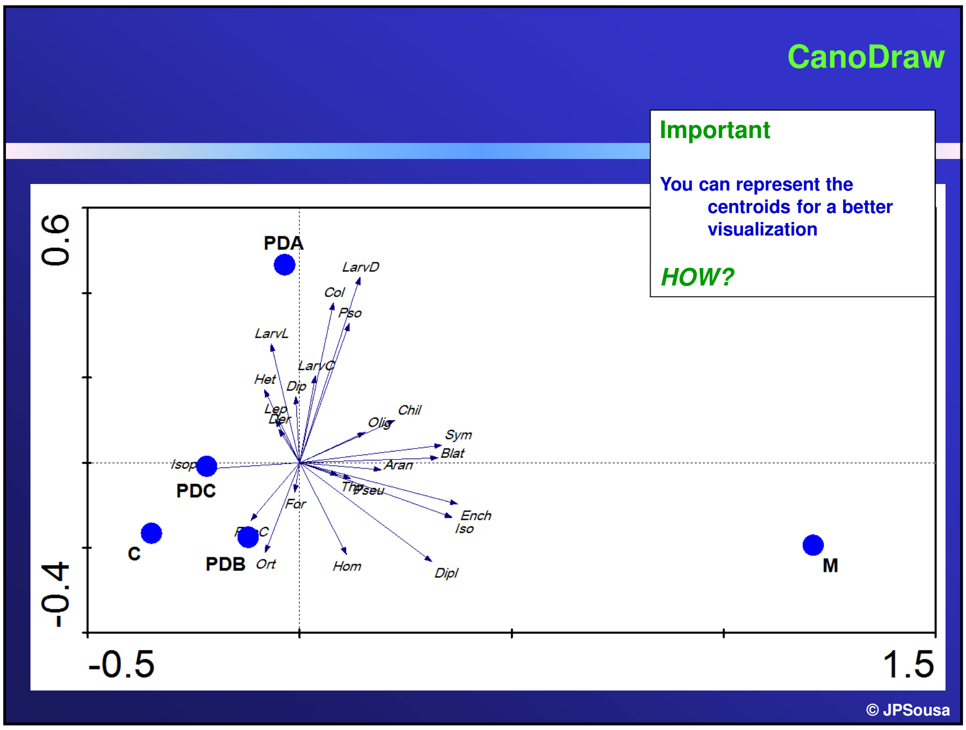

CanoDraw - PCA
CanoDraw
1. Click “create” and choose the desired biplot
© JPSousa

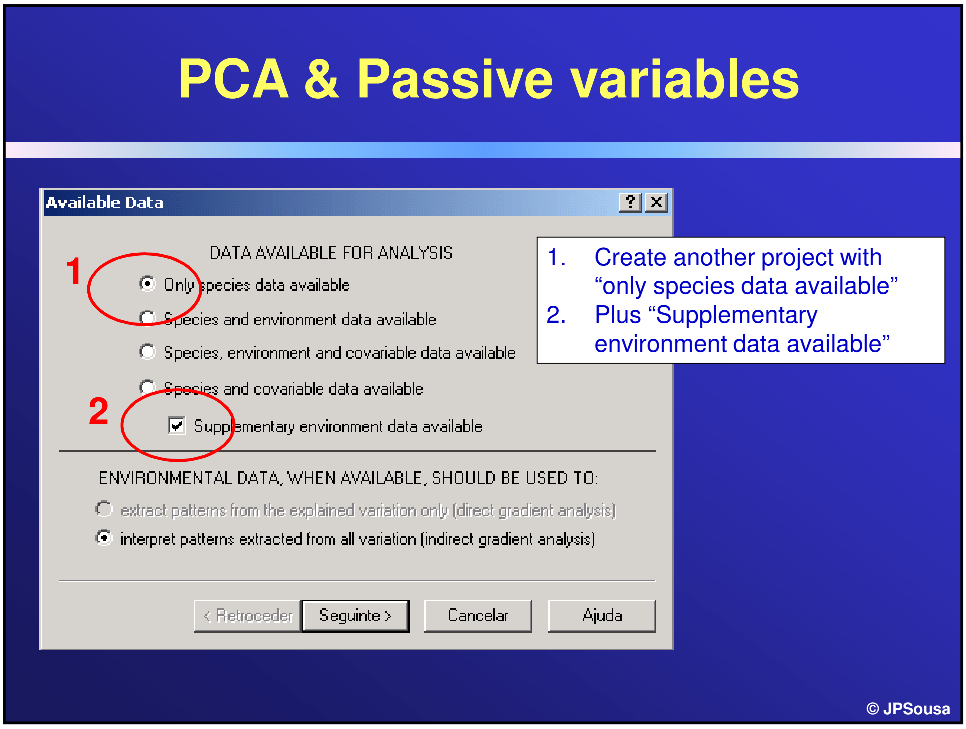

CanoDraw
0.8
-0.8
-1.0
1.5
© JPSousa

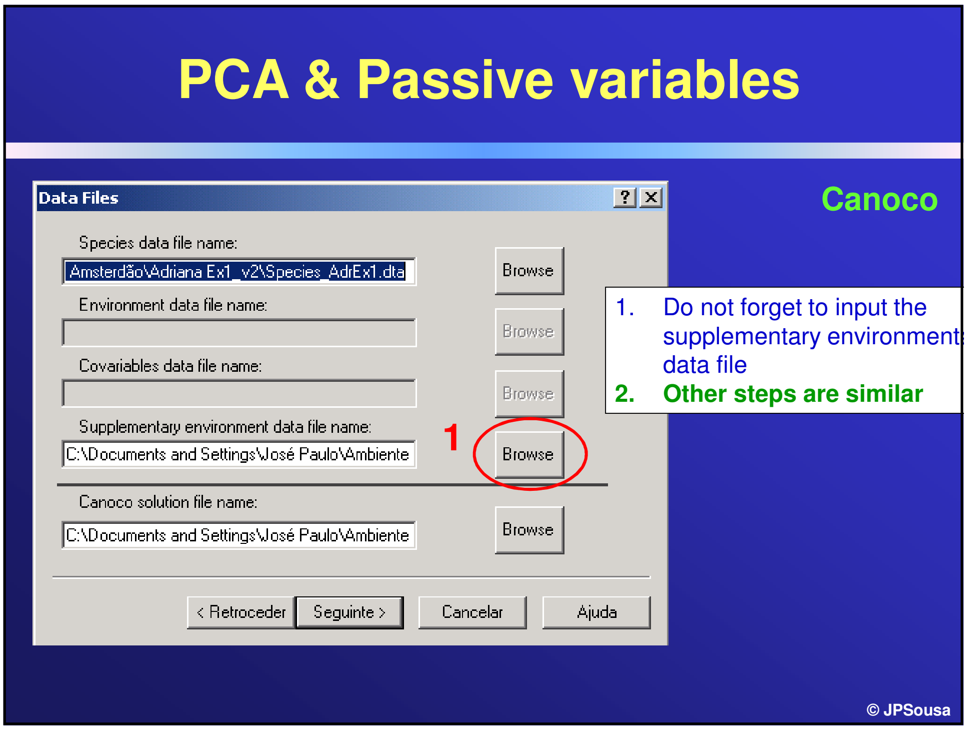

1. View – Visual attributes (change font type, size, colour, symbol, etc)
2. Project – Settings (slect axis, samples an species labels, etc)
3. Project – Classify (classify groups of samples or species for diferentiation)
Important
After making any change in the layout click on "recreate graph" (Project menu)
0.8
-0.8
-1.0
1.5
PDA5
LarvD
PDA4
PDA2
PDA1
Col
Pso
PDB4
PDC2
PDB5
LarvC
Het
Dip
PDC1
Chil
Olig
Sym
Blat
PDC5
PDA3
C2
Isopt
Aran
Pseu
For
Ench
Iso
C5
PupC
C1
PDC3
C3
C4
PDC4
Ort
Hom
Dipl
PDB2
PDB3
PDB1
M1
M2
M3
© JPSousa


CanoDraw
Interpretation
"Biplot" rule
0.8
-0.8
-1.0
1.5
PDA5
LarvD
PDA4
PDA2
PDA1
Col
Pso
PDB4
PDC2
PDB5
LarvC
Het
Dip
PDC1
Chil
Olig
Sym
Blat
PDC5
PDA3
C2
Isopt
Aran
Pseu
For
Ench
Iso
C5
PupC
C1
PDC3
C3
C4
PDC4
Ort
Hom
Dipl
PDB2
PDB3
PDB1
M1
M2
M3
© JPSousa


CanoDraw
1.0
-1.0
Sa5
Sa2
Sa3
Sa1
Spc_1
Spc_2
Sa4
-1.0
1.5
Interpretation
"Biplot" rule
© JPSousa


CanoDraw
Important
You can represent the centroids for a better visualization
HOW?
0.6
-0.4
PDA
LarvD
Col
Pso
LarvL
LarvC
Het
Dip
Chil
Olig
Sym
Blat
Isop
PDC
Aran
For
Ench
Iso
C
PDB
Ort
Hom
Dipl
M
-0.5
1.5
© JPSousa


PCA & Passive variables
Available Data
DATA AVAILABLE FOR ANALYSIS
1
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
2
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Create another project with "only species data available"
2. Plus "Supplementary environment data available"
© JPSousa


PCA & Passive variables
Canoco
Data Files
Species data file name:
Amsterdão\Adriana Ex1_v2\Species_AdrEx1.dta
Browse
Environment data file name:
Browse
Covariables data file name:
Browse
Supplementary environment data file name:
C:\Documents and Settings\José Paulo\Ambiente
1
Browse
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Do not forget to input the supplementary environment data file
2. Other steps are similar
© JPSousa

# PCA & Passive variables

**Canoco**

Log: Adriana_PCA_Cent.con

```
**** Summary ****

Axes                                     1       2       3       4  Total variance

Species-environment correlations  :   0.920   0.648   0.461   0.575
Cumulative percentage variance
   of species data                :    27.4    42.9    53.9    62.8
   of species-environment relation:    56.3    72.0    77.7    84.9

Sum of all            eigenvalues                                         1.000
Sum of all canonical  eigenvalues                                         0.412
```

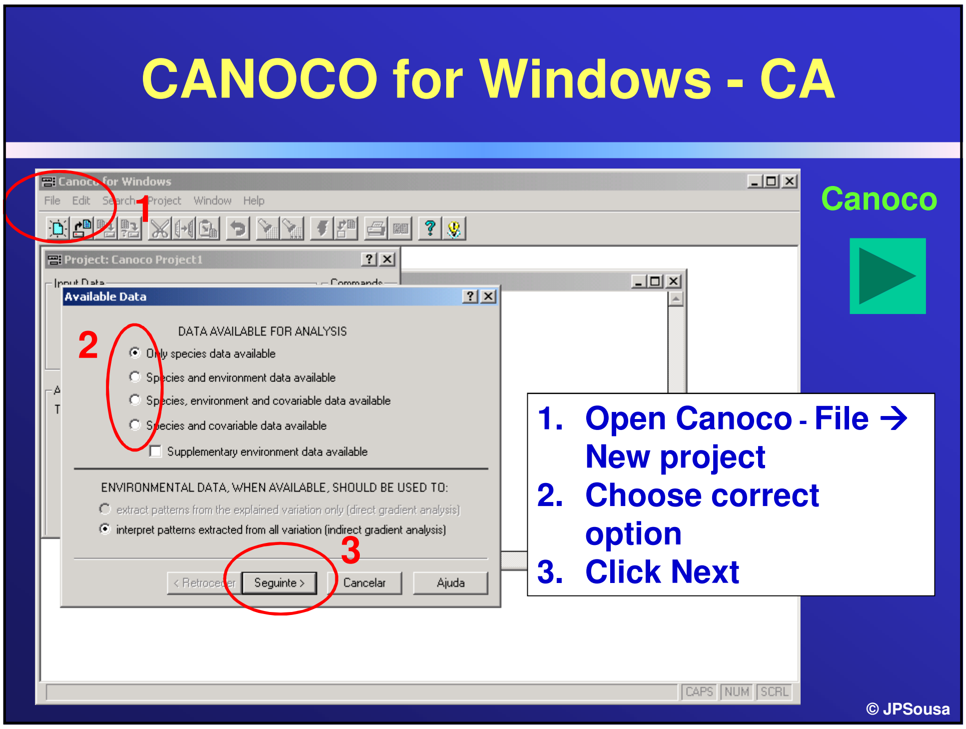
1. Treatment (suppl. Envir. variable) explain 41.2% of total variation
2. From these, 56.3% is explained in axis 1, etc

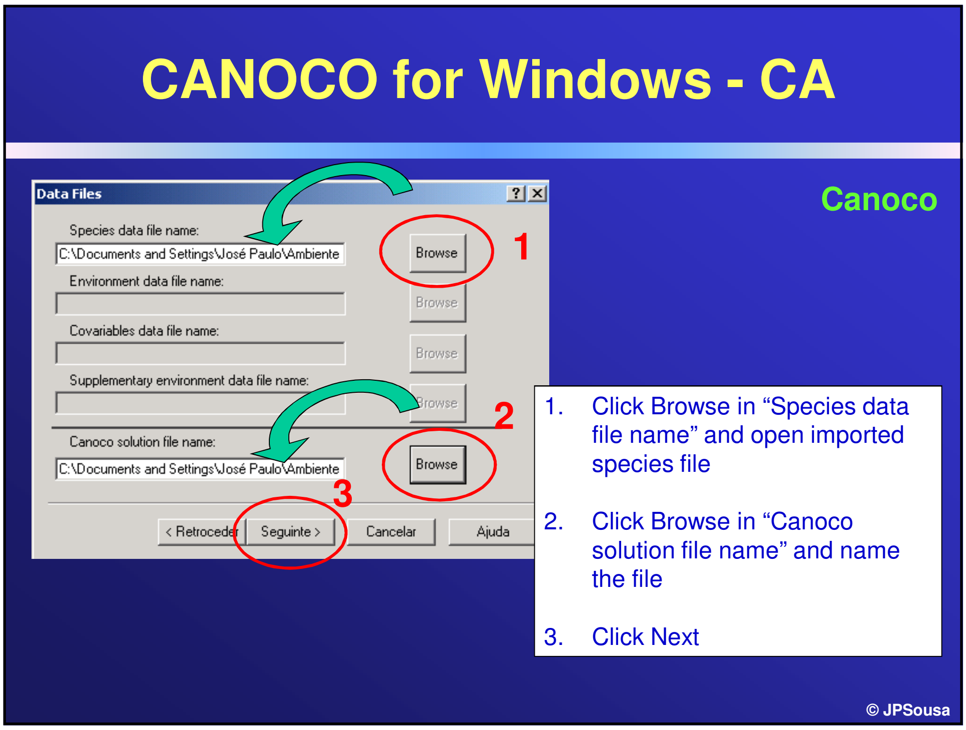


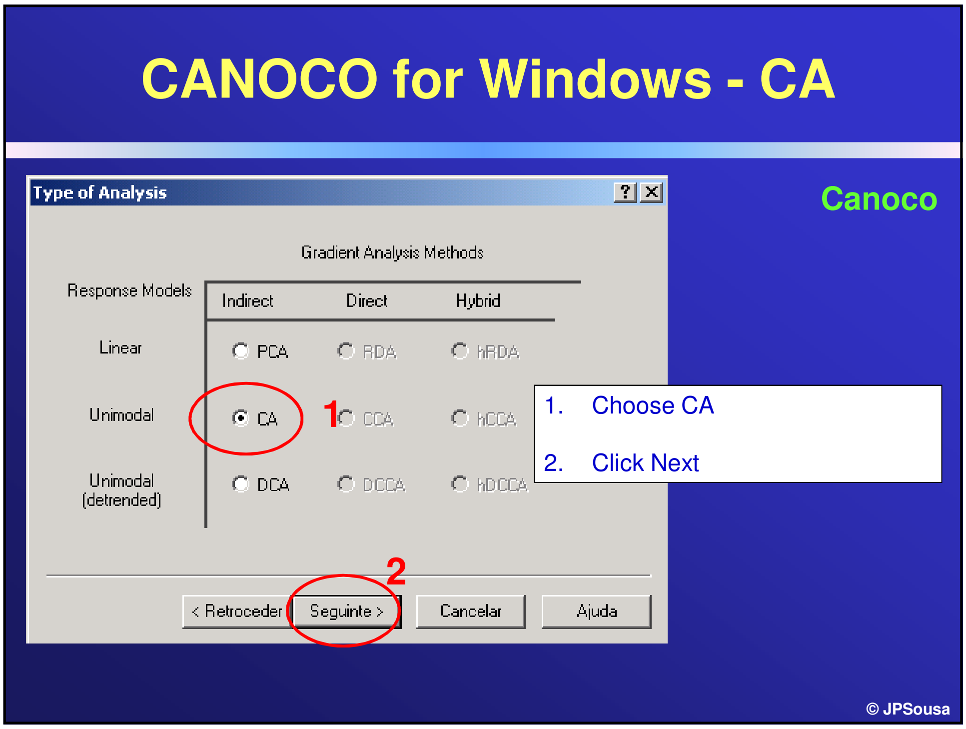
# CANOCO for Windows - CA

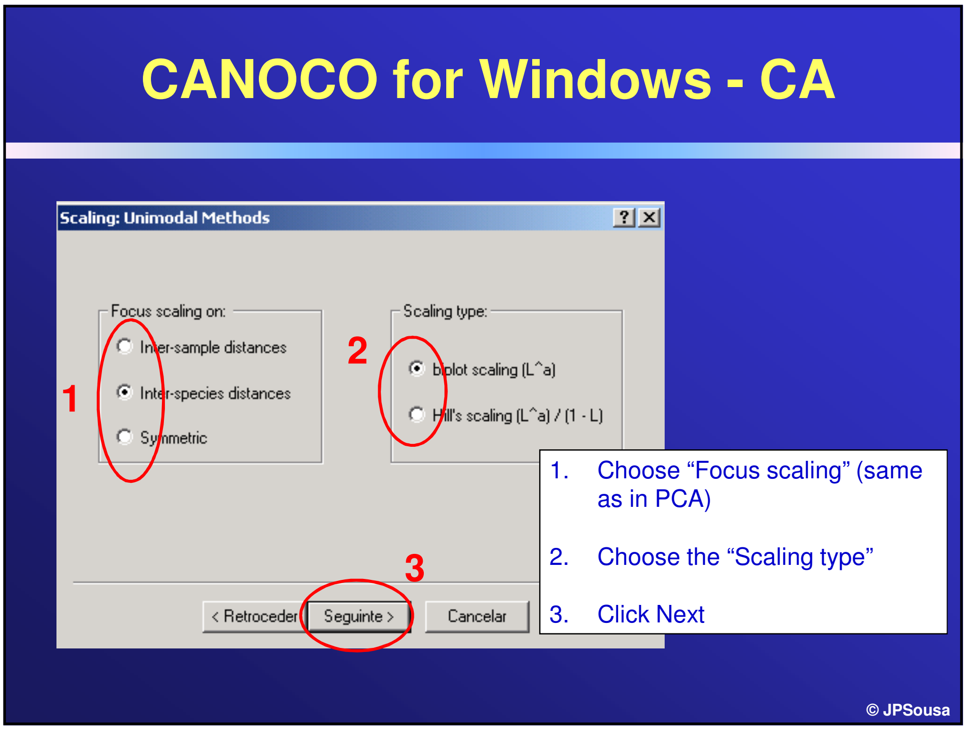
Soil fauna from Cork Oak (*Quercus suber*) and Eucalyptus (*Eucalyptus globulus*) stands (Sousa et al., 2003)

- Soil mesofauna data; soil pedological parameters (File Matrizes_CA_CCA.xls)
- 2 sites (Q e E) with four plots each (A, B, G, M) and each plot with 4 soil cores ;
- 32 samples in total with 45 collembola species identified;
- Objective 1: verify the association between species and sites


CANOCO for Windows - CA
Canoco
Canoco for Windows
File Edit Search Project Window Help
Project: Canoco Project1
Available Data
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
1
2
3
1. Open Canoco - File → New project
2. Choose correct option
3. Click Next
© JPSousa


CANOCO for Windows - CA
Canoco
Data Files
Species data file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
Environment data file name:
Browse
Covariables data file name:
Browse
Supplementary environment data file name:
Browse
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
< Retroceder
Seguinte >
Cancelar
Ajuda
1
2
3
1. Click Browse in “Species data file name” and open imported species file
2. Click Browse in “Canoco solution file name” and name the file
3. Click Next
© JPSousa


CANOCO for Windows - CA
Canoco
Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
1
2
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Choose CA
2. Click Next
© JPSousa


CANOCO for Windows - CA
Scaling: Unimodal Methods
Focus scaling on:
Inter-sample distances
Inter-species distances
Symmetric
Scaling type:
biplot scaling (L^a)
Hill's scaling (L^a) / (1 - L)
1
2
3
< Retroceder
Seguinte >
Cancelar
1. Choose "Focus scaling" (same as in PCA)
2. Choose the "Scaling type"
3. Click Next
© JPSousa

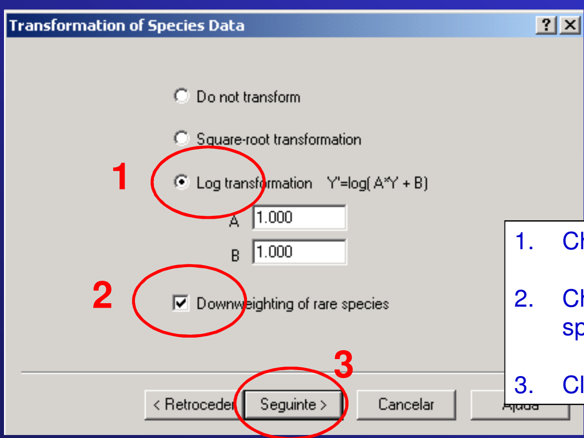
# CANOCO for Windows

## Scaling in CA

- Hill's scaling: large gradients (>4 SD) – interpretation via the centroid rule
- Biplot scaling: shorter gradients (±3 SD) – interpretation via the biplot rule

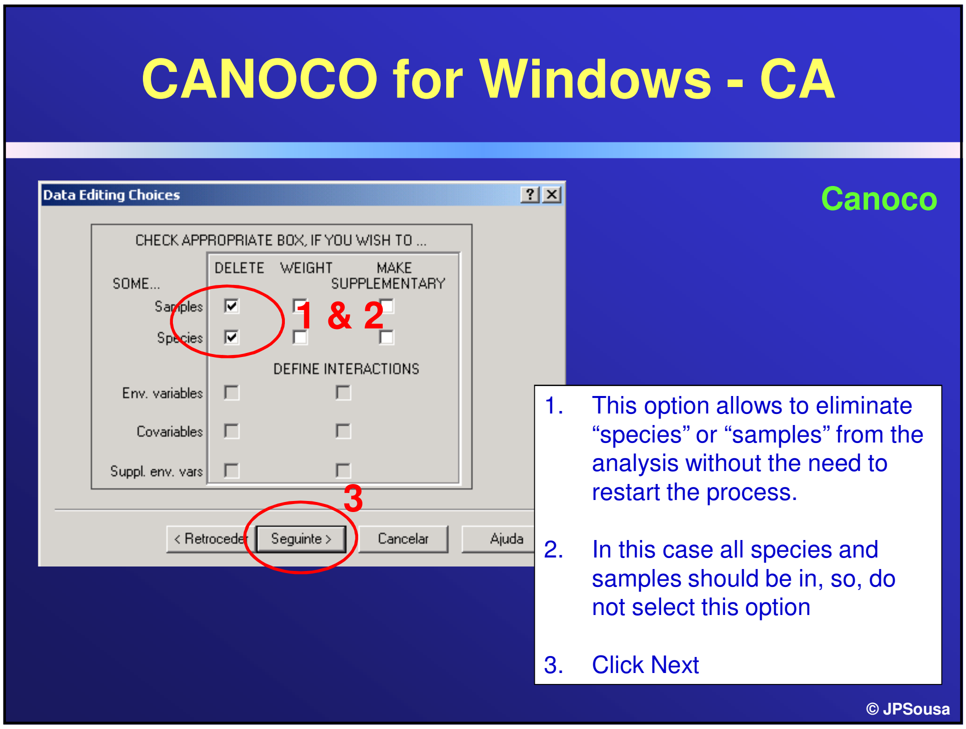


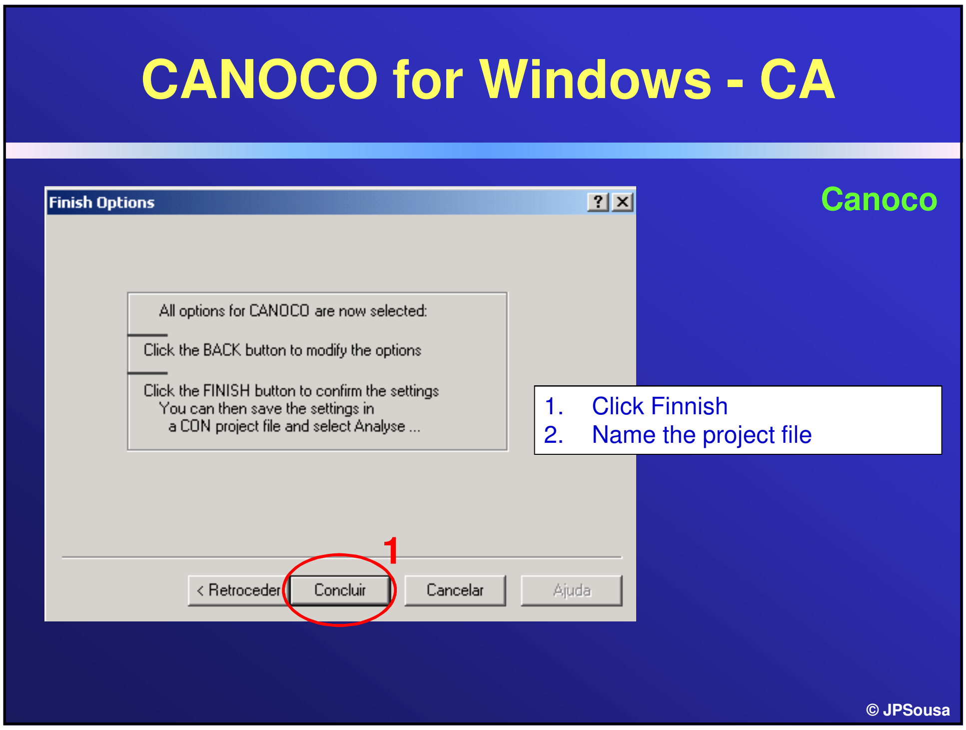
# CANOCO for Windows - CA




Canoco

1. Choose "log transformation"
2. Chooser "downweighting rare species"
3. Click Next


CANOCO for Windows - CA
Canoco
Data Editing Choices
CHECK APPROPRIATE BOX, IF YOU WISH TO ...
DELETE
WEIGHT
MAKE SUPPLEMENTARY
SOME...
Samples
Species
DEFINE INTERACTIONS
Env. variables
Covariables
Suppl. env. vars
1 & 2
3
< Retroceder
Seguinte >
Cancelar
Ajuda
1. This option allows to eliminate “species” or “samples” from the analysis without the need to restart the process.
2. In this case all species and samples should be in, so, do not select this option
3. Click Next
© JPSousa


CANOCO for Windows - CA
Canoco
Finish Options
All options for CANOCO are now selected:
Click the BACK button to modify the options
Click the FINISH button to confirm the settings
You can then save the settings in
a CON project file and select Analyse ...
1
< Retroceder
Concluir
Cancelar
Ajuda
1. Click Finnish
2. Name the project file
© JPSousa

# CANOCO for Windows - CA

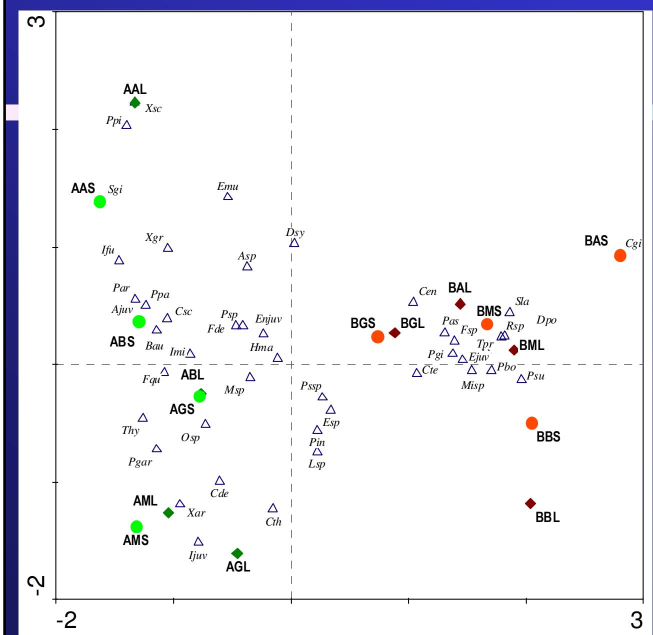
Log: PauloEuc_CA.con

```
**** Summary ****

Axes                                 1       2       3       4   Total inertia

Eigenvalues                     :   0.505   0.198   0.175   0.158          1.682
Cumulative percentage variance
   of species data              :    30.0    41.8    52.2    61.6

Sum of all              eigenvalues                                          1.682
[Thu Jul 28 15:41:57 2005] CANOCO call succeeded
[Thu Jul 28 15:42:33 2005] Settings change cancelled
```

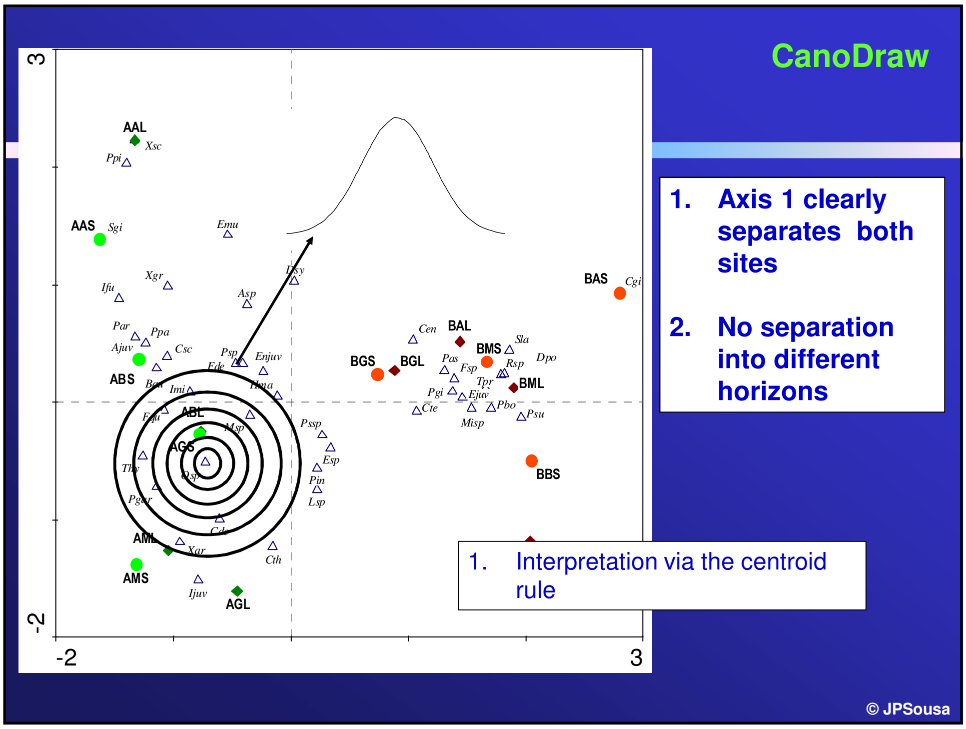
1. Axis 1 explains 30% of total variation
2. Axis 2 explains 11,8% of total variation

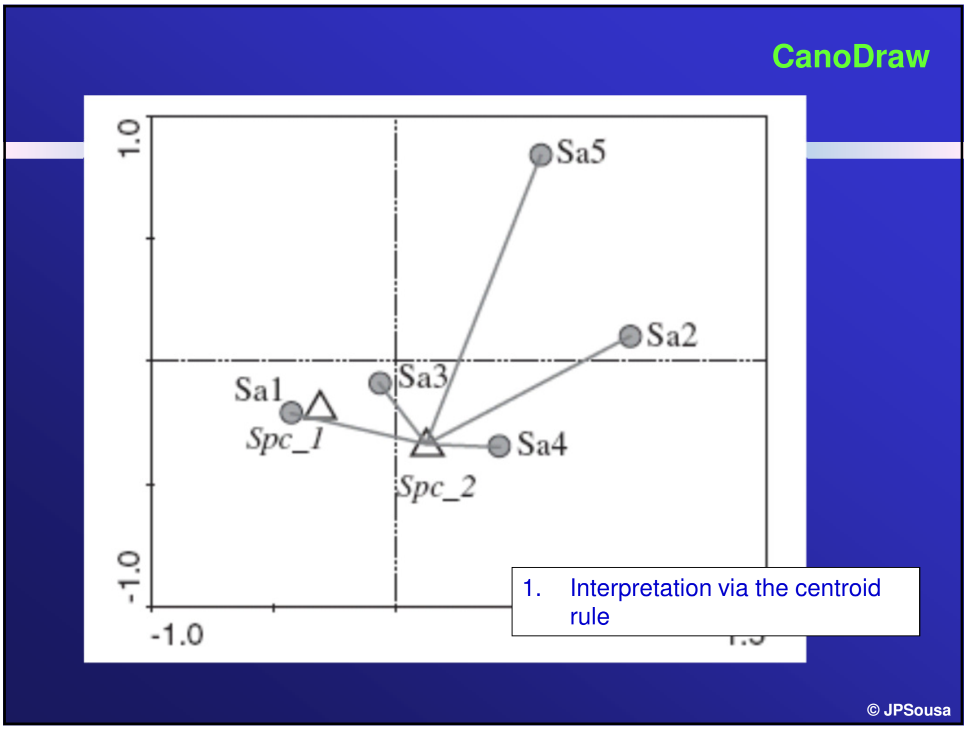


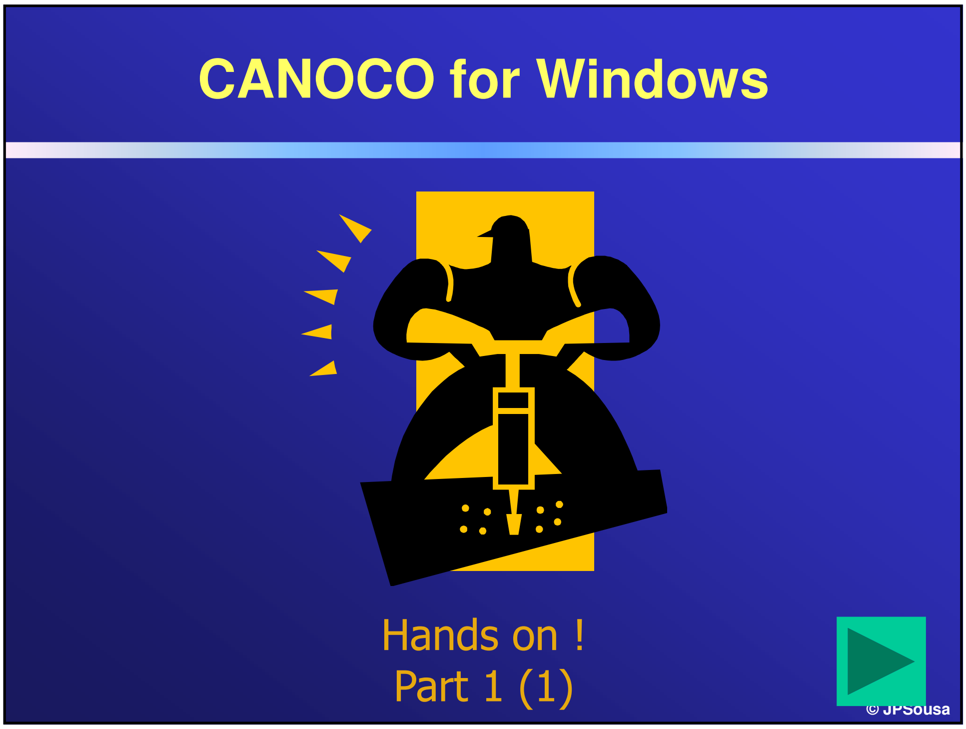
## CanoDraw

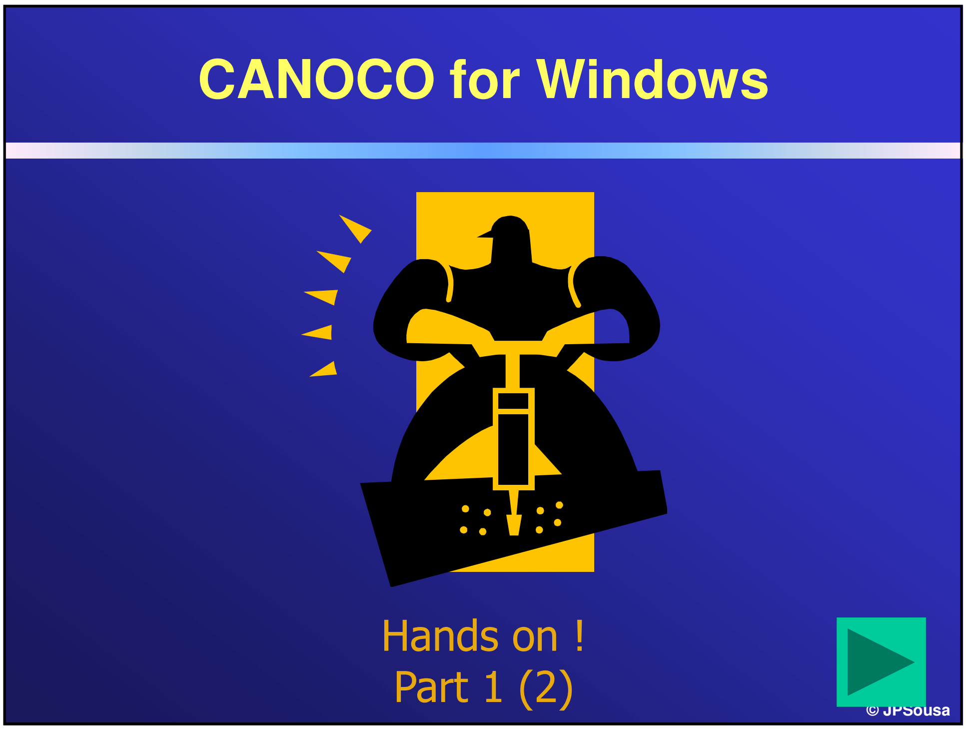



1. Axis 1 clearly separates both sites
2. No separation into different horizons


CanoDraw
1. Axis 1 clearly separates both sites
2. No separation into different horizons
1. Interpretation via the centroid rule
© JPSousa


CanoDraw
Sa5
Sa2
Sa3
Sa1
Spc_1
Spc_2
Sa4
1.0
-1.0
-1.0
1. Interpretation via the centroid rule
© JPSousa


CANOCO for Windows
Hands on !
Part 1 (1)
© JPSousa


CANOCO for Windows
Hands on !
Part 1 (2)
© JPSousa

Multivariate Statistical Tools in Ecology

ISCED, Lubango, March 2016

# Ordination Tools I: Representing the underlying structure of a dataset Part 2



# "Non-Metric Multidimentional Scaling"

## What is the purpose ?

- Represent samples in a ordination space
- Advantage: you can choose the metric (similarity/distance index) used to evaluate the distance among samples
- Advantage: it preserves the distances in the multidimentional space
- It constructs a configuration map of the samples in the *m* dimentions based on the relative similarity between samples

## "Non-Metric Multidimentional Scaling"

### How does it work ?

- Constructs a representation of the samples
- Compare the distances among them (in the diagram) with the values on the (di)similarity matrix
- Evaluates the relation between these two measures with a regression
- Evaluates the reliability of the regression (stress)
- Changes representation to reduce stress
- Repeats process until convergence



## "Non-Metric Multidimentional Scaling"

### Stress values ➔

Stress < 0.05 – excellent representation (low possibility of a wrong interpretation)

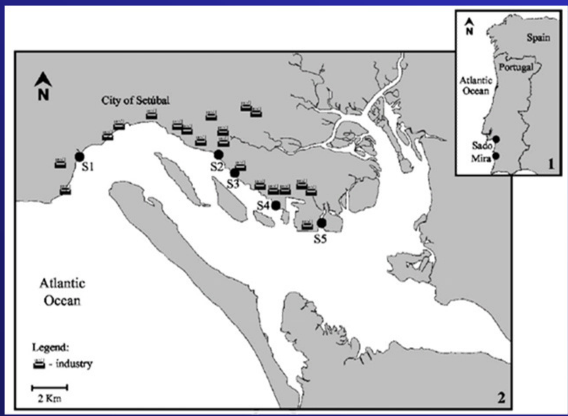
Stress < 0.1 – good representation (3D diagrams do not bring any additional information)

Stress < 0.2 – 2D diagram of certain utility (advisable to complement interpretation with other method)

Stress > 0.3 – non acceptable representation (samples are randomly placed in the diagram)

## NMDS (WinKyst - CANOCO)

Example




Moreira et.al (2006) Aquatic Toxicology

**Physiological functions in the polychaete *Hediste diversicolor***

Measurements of several enzyme biomarkers (neurotransmission, metabolic condition, detoxification process, antioxidant defences)

Reference estuary – Rio Mira

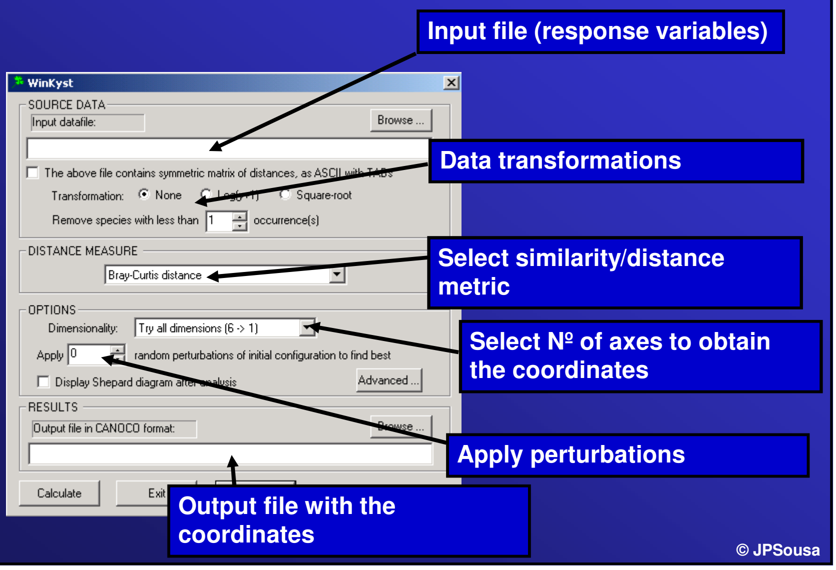
Impacted estuary - Rio Sado

Several sampling sites with several animals at each one of them

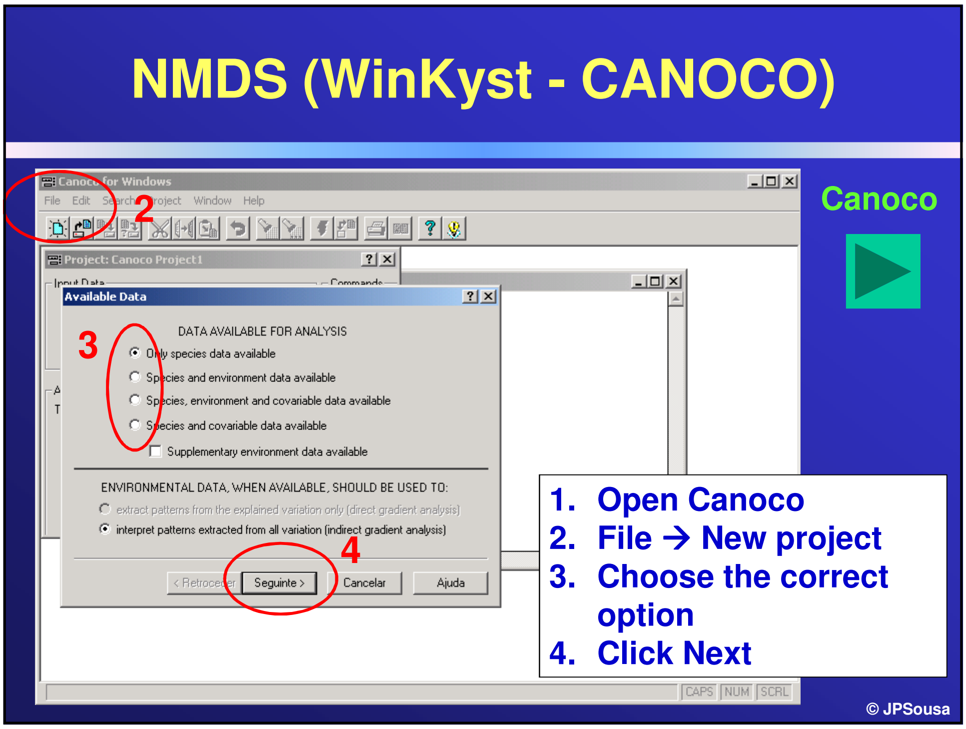
© JPSousa

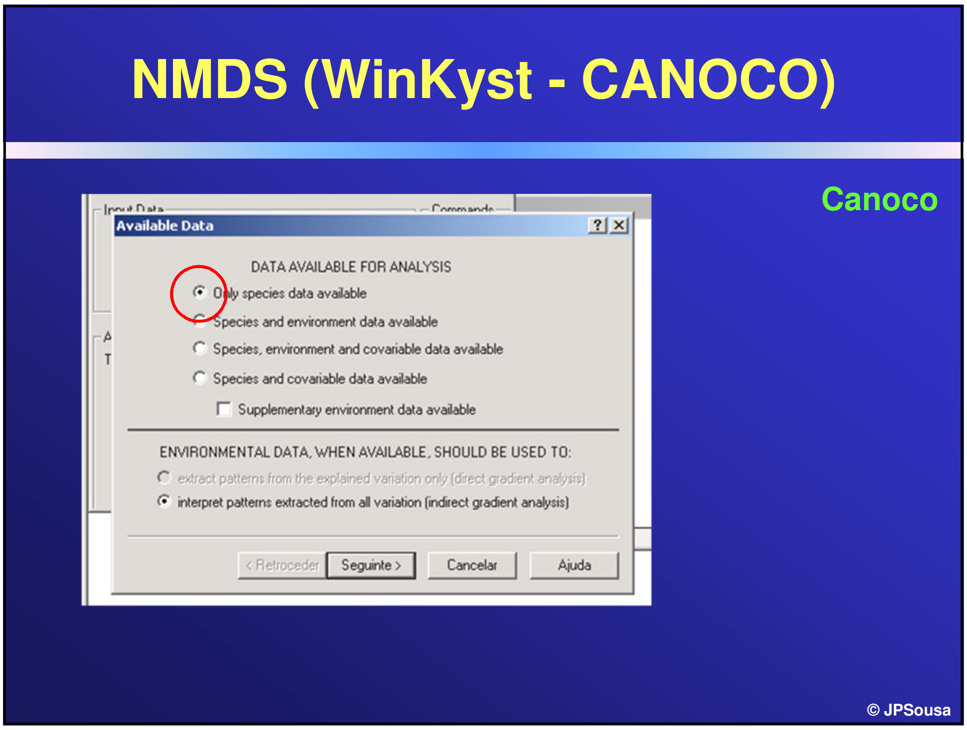
## NMDS (WinKyst - CANOCO)

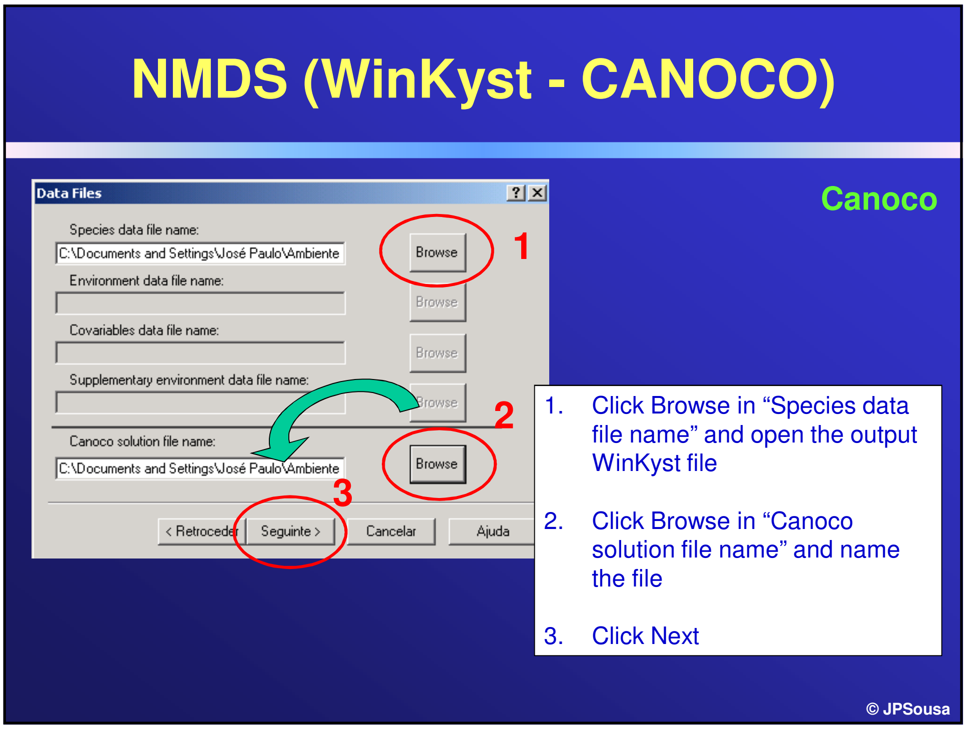
WinKyst

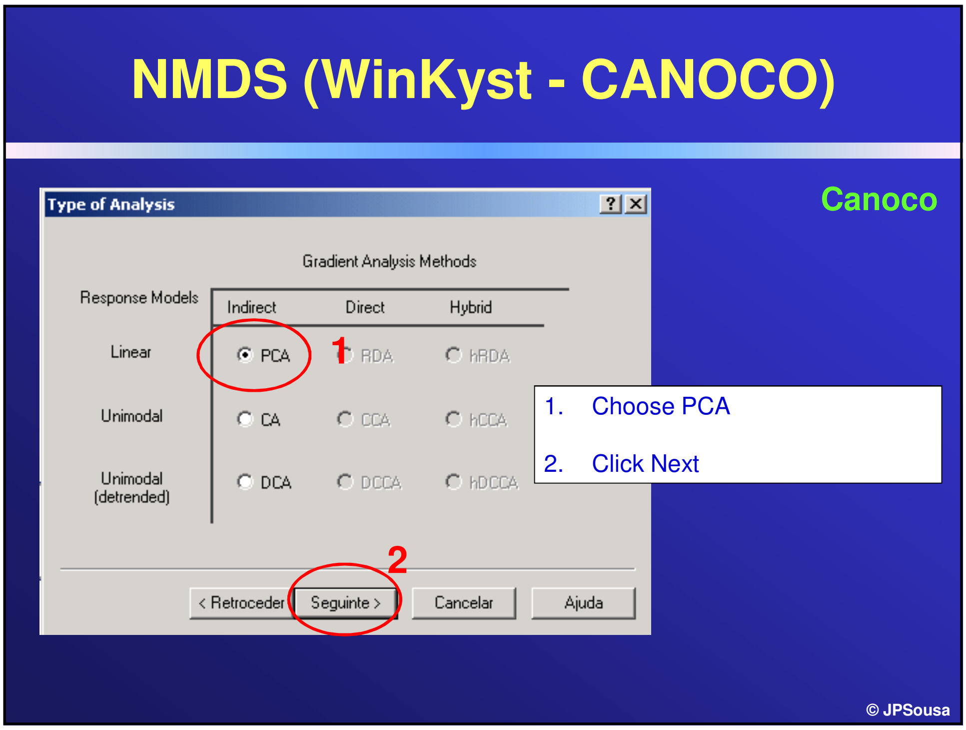


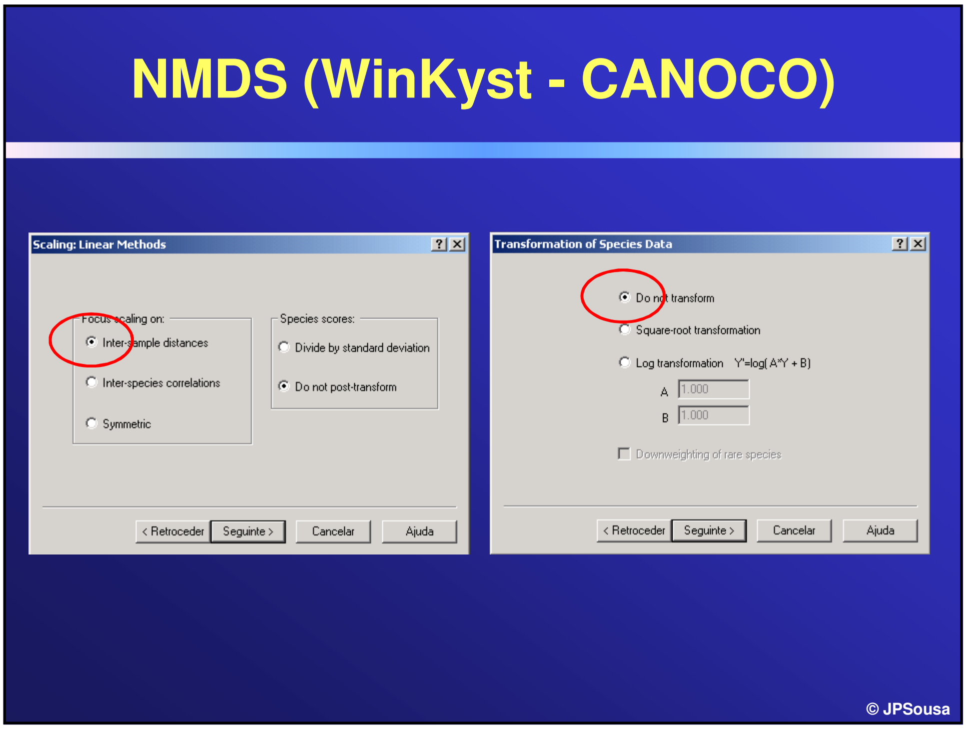

© JPSousa

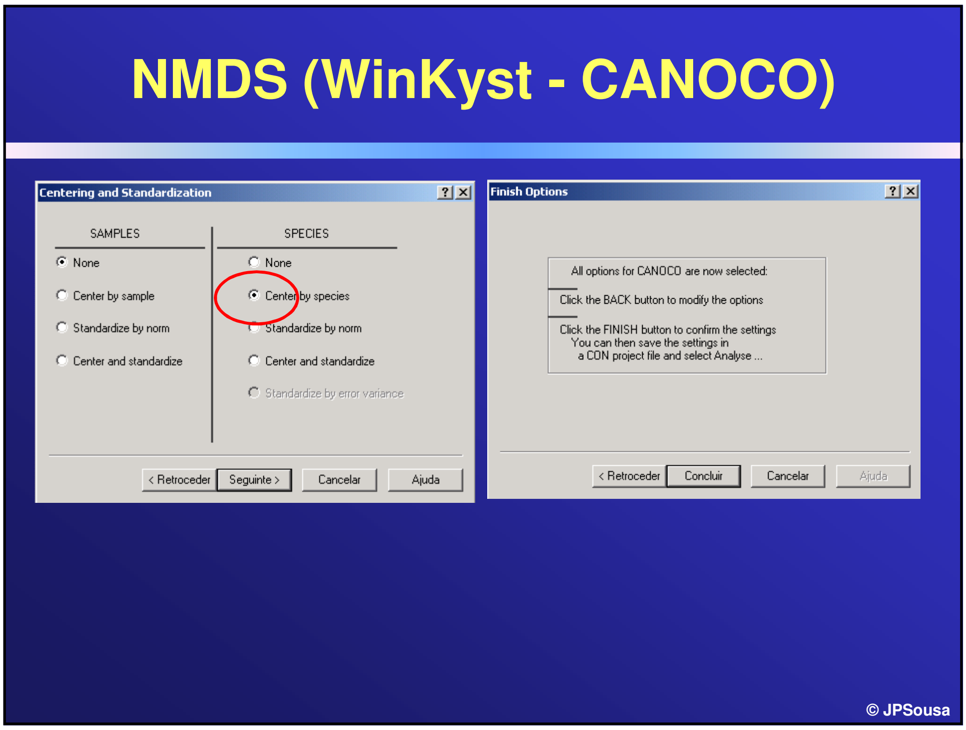

NMDS (WinKyst - CANOCO)
Canoco
Canoco for Windows
File Edit Search Project Window Help
2
Project: Canoco Project1
Available Data
3
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
4
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Open Canoco
2. File → New project
3. Choose the correct option
4. Click Next
CAPS NUM SCRL
© JPSousa

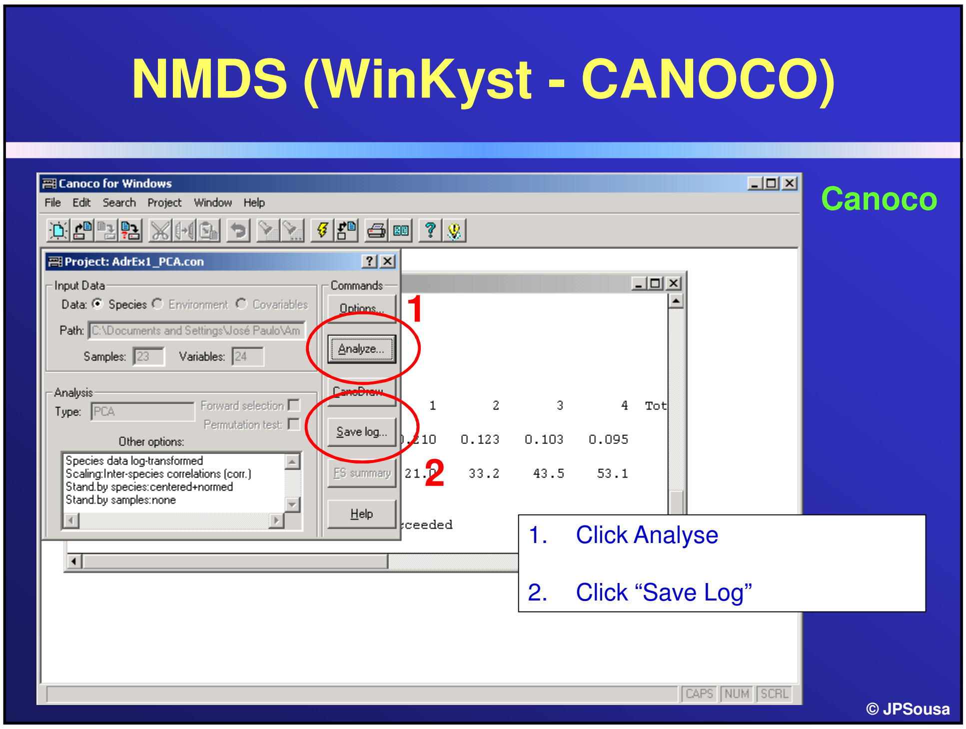

NMDS (WinKyst - CANOCO)
Canoco
Available Data
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

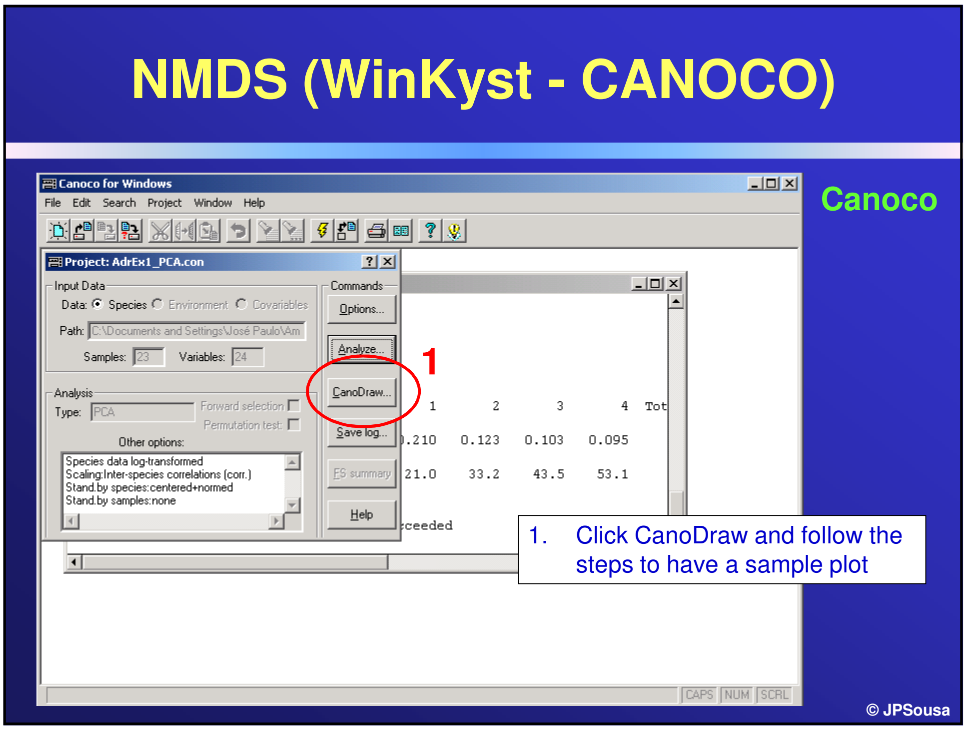

NMDS (WinKyst - CANOCO)
Canoco
Data Files
Species data file name:
C:\Documents and Settings\José Paulo\Ambiente
Environment data file name:
Covariables data file name:
Supplementary environment data file name:
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
1
2
3
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Click Browse in "Species data file name" and open the output WinKyst file
2. Click Browse in "Canoco solution file name" and name the file
3. Click Next
© JPSousa

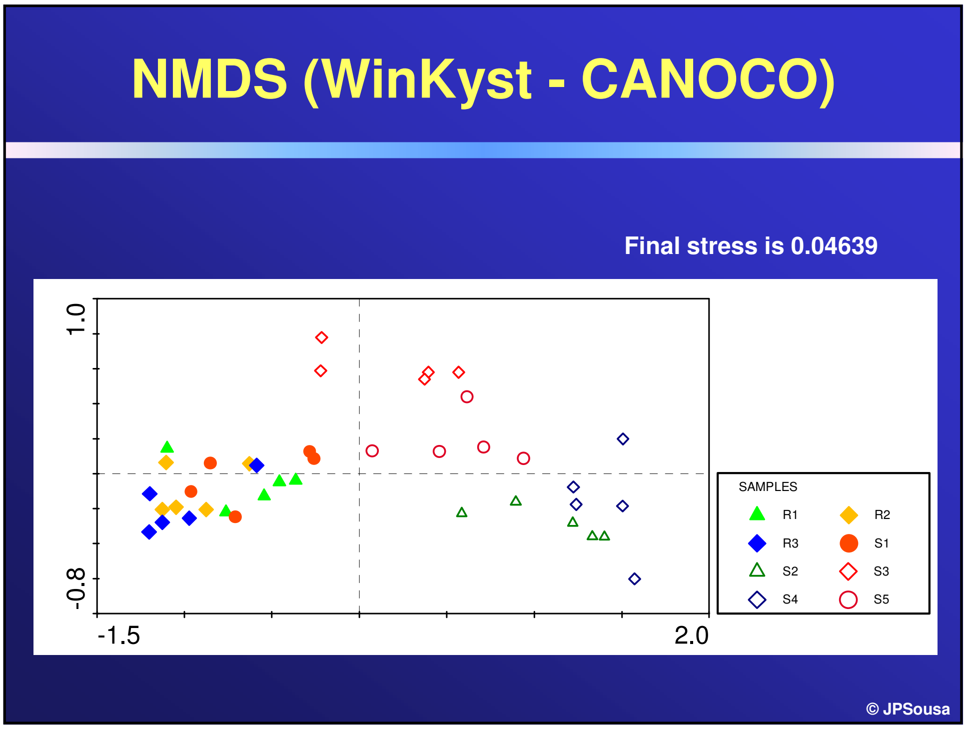

NMDS (WinKyst - CANOCO)
Canoco
Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
1
2
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Choose PCA
2. Click Next
© JPSousa

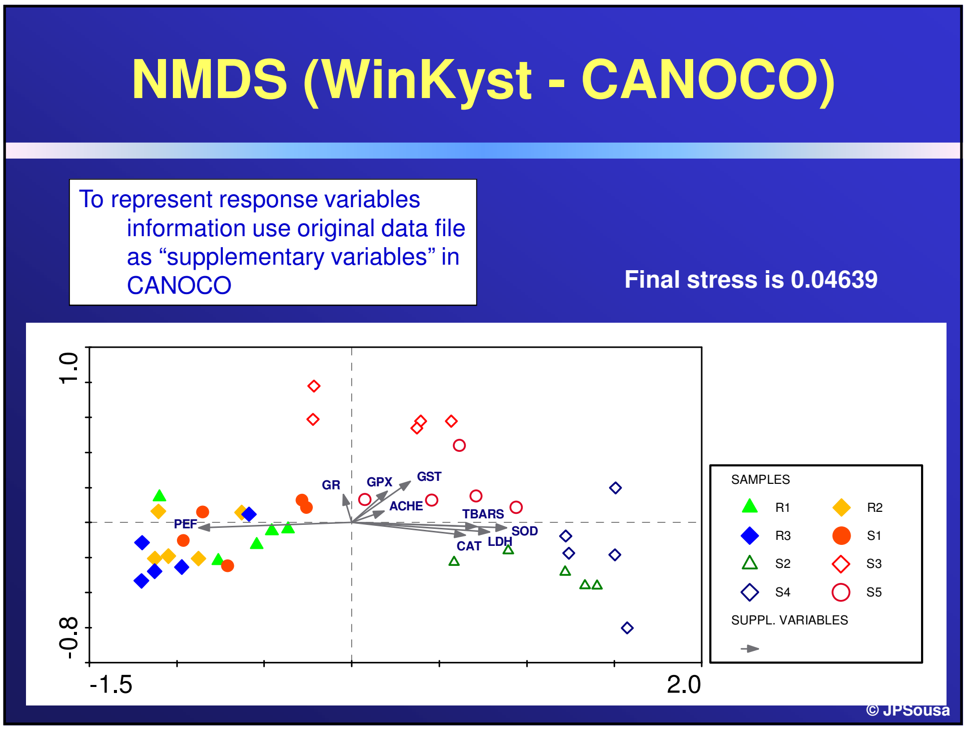

NMDS (WinKyst - CANOCO)
Scaling: Linear Methods
Focus scaling on:
Inter-sample distances
Inter-species correlations
Symmetric
Species scores:
Divide by standard deviation
Do not post-transform
< Retroceder
Seguinte >
Cancelar
Ajuda
Transformation of Species Data
Do not transform
Square-root transformation
Log transformation Y'=log( A*Y + B)
A 1.000
B 1.000
Downweighting of rare species
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa


NMDS (WinKyst - CANOCO)
Centering and Standardization
SAMPLES
None
Center by sample
Standardize by norm
Center and standardize
SPECIES
None
Center by species
Standardize by norm
Center and standardize
Standardize by error variance
< Retroceder
Seguinte >
Cancelar
Ajuda
Finish Options
All options for CANOCO are now selected:
Click the BACK button to modify the options
Click the FINISH button to confirm the settings
You can then save the settings in
a CON project file and select Analyse ...
< Retroceder
Concluir
Cancelar
Ajuda
© JPSousa


NMDS (WinKyst - CANOCO)
Canoco
Canoco for Windows
Project: AdrEx1_PCA.con
Analyze...
Save log...
1
2
1. Click Analyse
2. Click “Save Log”
© JPSousa


NMDS (WinKyst - CANOCO)
Canoco
Canoco for Windows
Project: AdrEx1_PCA.con
CanoDraw...
1
1. Click CanoDraw and follow the steps to have a sample plot
© JPSousa


NMDS (WinKyst - CANOCO)
Final stress is 0.04639
1.0
-0.8
-1.5
2.0
SAMPLES
R1
R2
R3
S1
S2
S3
S4
S5
© JPSousa


NMDS (WinKyst - CANOCO)
To represent response variables information use original data file as "supplementary variables" in CANOCO
Final stress is 0.04639
1.0
-0.8
-1.5
2.0
GR
GPX
GST
ACHE
PEF
TBARS
SOD
CAT
LDH
SAMPLES
R1
R2
R3
S1
S2
S3
S4
S5
SUPPL. VARIABLES
© JPSousa

## NMDS

Hands On !
Part 2

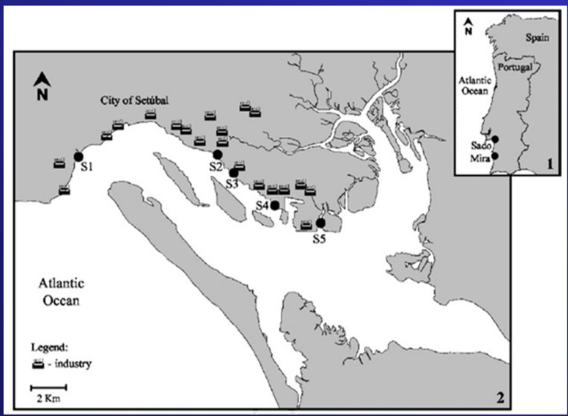


Multivariate Statistical Tools in Ecology
ISCED, Lubango, March 2016

## Ordination Tools II: Discriminating groups of samples/subjects Part 1

# NMDS (using PRIMER v.5)

Example




Moreira et.al (2006) Aquatic Toxicology

**Physiological functions in the polychaete *Hediste diversicolor***

Measurements of several enzyme biomarkers (neurotransmission, metabolic condition, detoxification process, antioxidant defences)

Reference estuary – Rio Mira

Impacted estuary - Rio Sado

Several sampling sites with several animals at each one of them

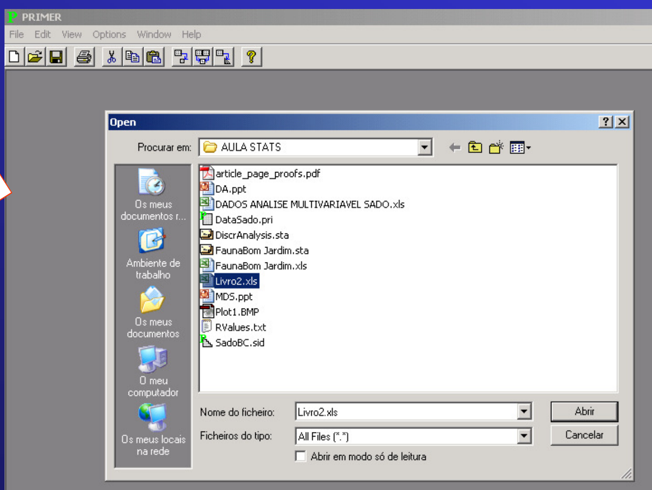


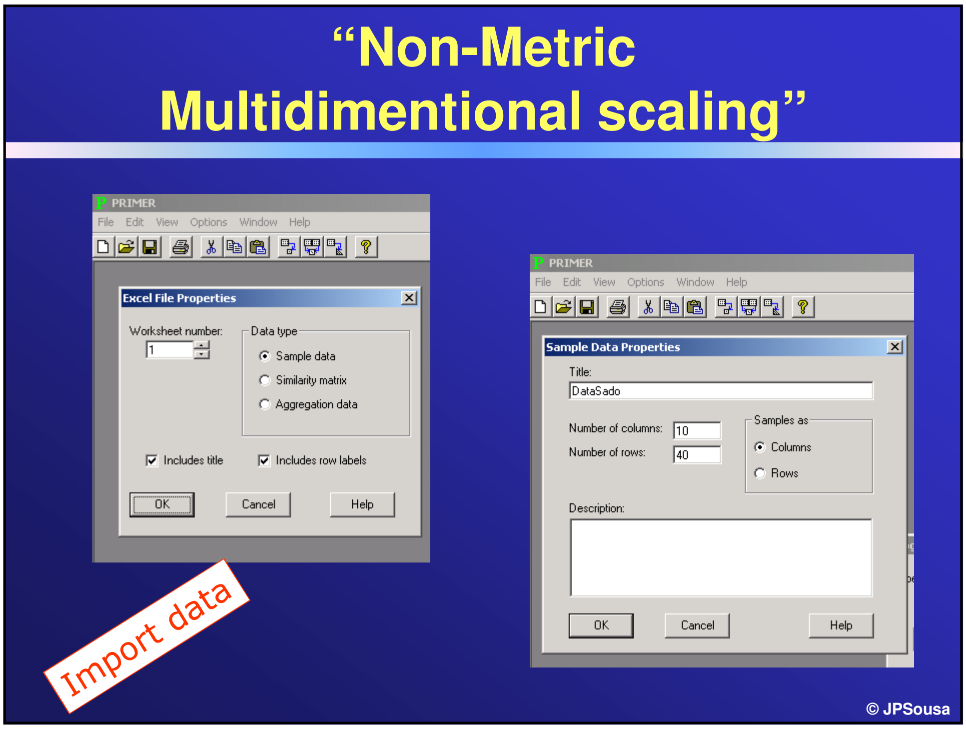
# "Non-Metric Multidimentional scaling"

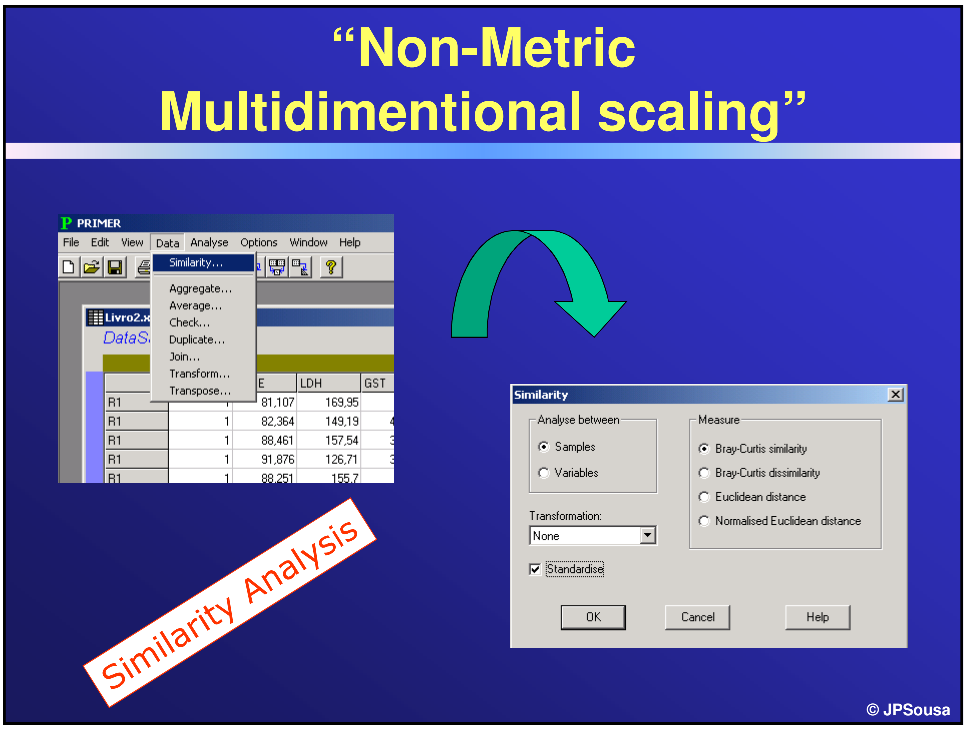
Example

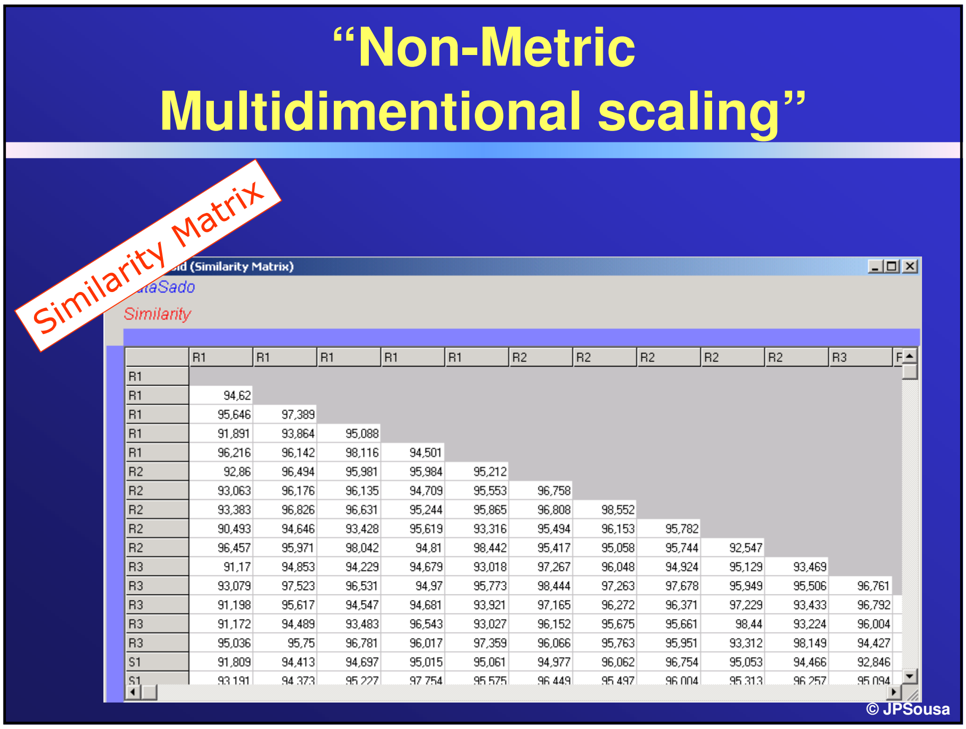
Import data

PRIMER

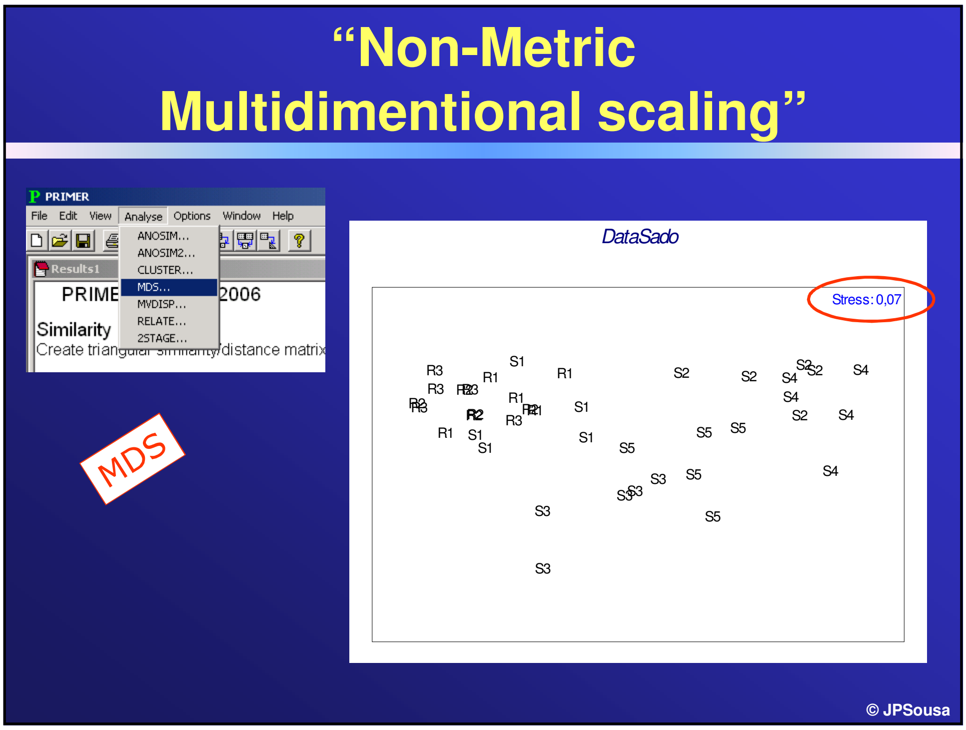



"Non-Metric
Multidimentional scaling"
PRIMER
File Edit View Options Window Help
Excel File Properties
Worksheet number:
1
Data type
Sample data
Similarity matrix
Aggregation data
Includes title
Includes row labels
OK
Cancel
Help
Sample Data Properties
Title:
DataSado
Number of columns:
10
Number of rows:
40
Samples as
Columns
Rows
Description:
OK
Cancel
Help
Import data
© JPSousa


"Non-Metric
Multidimentional scaling"
PRIMER
File Edit View Data Analyse Options Window Help
Similarity...
Aggregate...
Average...
Check...
Duplicate...
Join...
Transform...
Transpose...
LDH
GST
Similarity
Analyse between
Samples
Variables
Measure
Bray-Curtis similarity
Bray-Curtis dissimilarity
Euclidean distance
Normalised Euclidean distance
Transformation:
None
Standardise
OK
Cancel
Help
Similarity Analysis
© JPSousa




| | R1 | R1 | R1 | R1 | R1 | R2 | R2 | R2 | R2 | R2 | R3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| R1 | | | | | | | | | | | |
| R1 | 94,62 | | | | | | | | | | |
| R1 | 95,646 | 97,389 | | | | | | | | | |
| R1 | 91,891 | 93,864 | 95,088 | | | | | | | | |
| R1 | 96,216 | 96,142 | 98,116 | 94,501 | | | | | | | |
| R2 | 92,86 | 96,494 | 95,981 | 95,984 | 95,212 | | | | | | |
| R2 | 93,063 | 96,176 | 96,135 | 94,709 | 95,553 | 96,758 | | | | | |
| R2 | 93,383 | 96,826 | 96,631 | 95,244 | 95,865 | 96,808 | 98,552 | | | | |
| R2 | 90,493 | 94,646 | 93,428 | 95,619 | 93,316 | 95,494 | 96,153 | 95,782 | | | |
| R2 | 96,457 | 95,971 | 98,042 | 94,81 | 98,442 | 95,417 | 95,058 | 95,744 | 92,547 | | |
| R3 | 91,17 | 94,853 | 94,229 | 94,679 | 93,018 | 97,267 | 96,048 | 94,924 | 95,129 | 93,469 | |
| R3 | 93,079 | 97,523 | 96,531 | 94,97 | 95,773 | 98,444 | 97,263 | 97,678 | 95,949 | 95,506 | 96,761 |
| R3 | 91,198 | 95,617 | 94,547 | 94,681 | 93,921 | 97,165 | 96,272 | 96,371 | 97,229 | 93,433 | 96,792 |
| R3 | 91,172 | 94,489 | 93,483 | 96,543 | 93,027 | 96,152 | 95,675 | 95,661 | 98,44 | 93,224 | 96,004 |
| R3 | 95,036 | 95,75 | 96,781 | 96,017 | 97,359 | 96,066 | 95,763 | 95,951 | 93,312 | 98,149 | 94,427 |
| S1 | 91,809 | 94,413 | 94,697 | 95,015 | 95,061 | 94,977 | 96,062 | 96,754 | 95,053 | 94,466 | 92,846 |
| S1 | 93,191 | 94,373 | 95,227 | 97,754 | 95,575 | 96,449 | 95,497 | 96,004 | 95,313 | 96,257 | 95,094 |



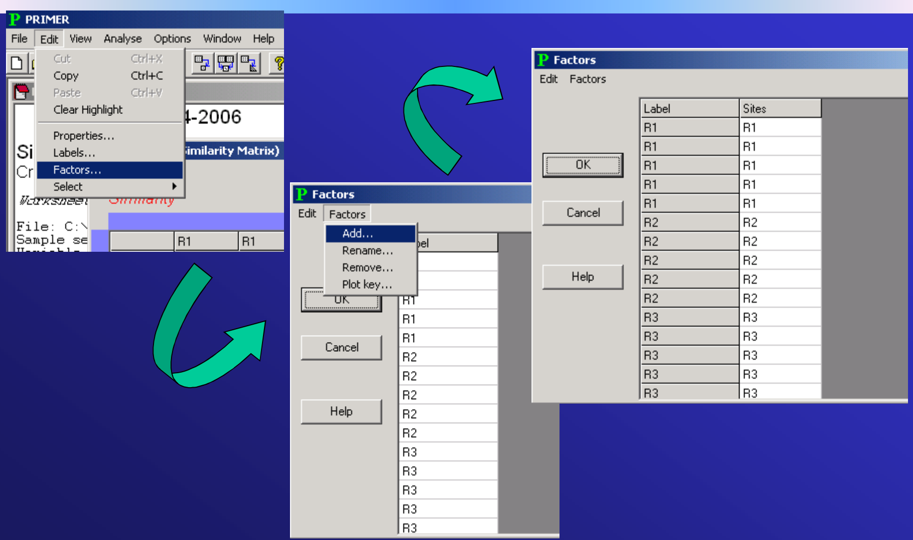
# "ANOSIM" in Primer v.5

## Discriminating groups ➔

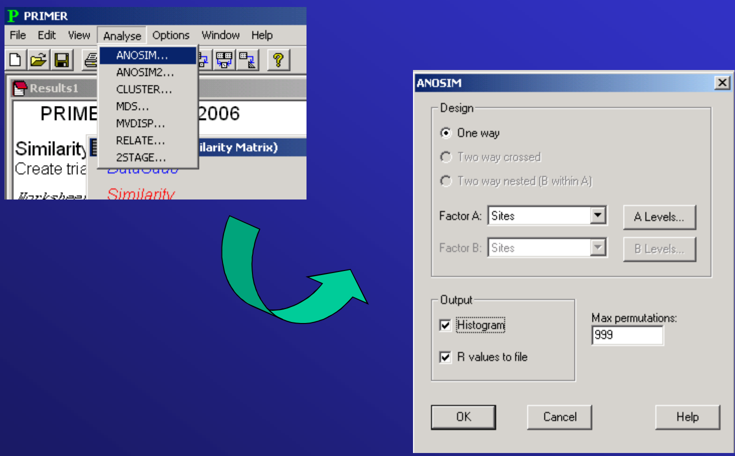
1. Test differences between the average rank similarity within groups ($r_W$) vs. between groups ($r_B$)
2. H0: No differences between groups
3. Calculate R values based on that comparison
4. Recalculate R values after permutations
5. Calculate the probability associated with the global R value



# "ANOSIM" in Primer v.5



# "ANOSIM" in Primer v.5






# "ANOSIM" in Primer v.5

**Global Test**

Global R value: 0,699

Significance value: 0,1% (0,001)

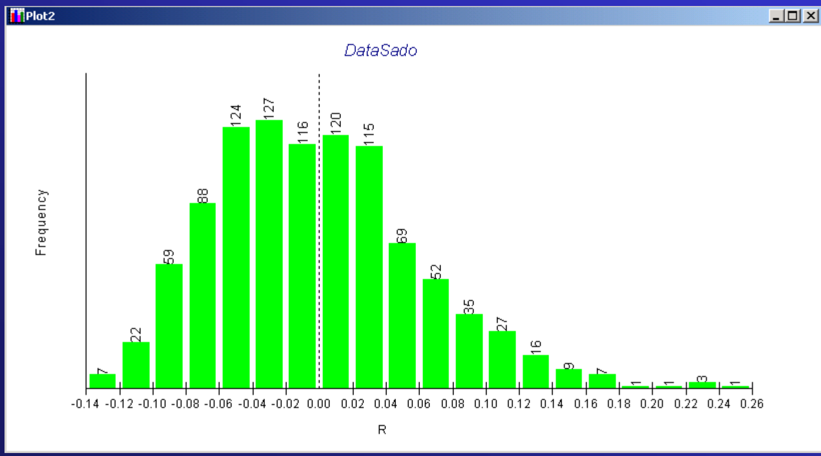
Nº of permutations: 999

Nº of permuted values equal or higher than global R: 0

Reject H0

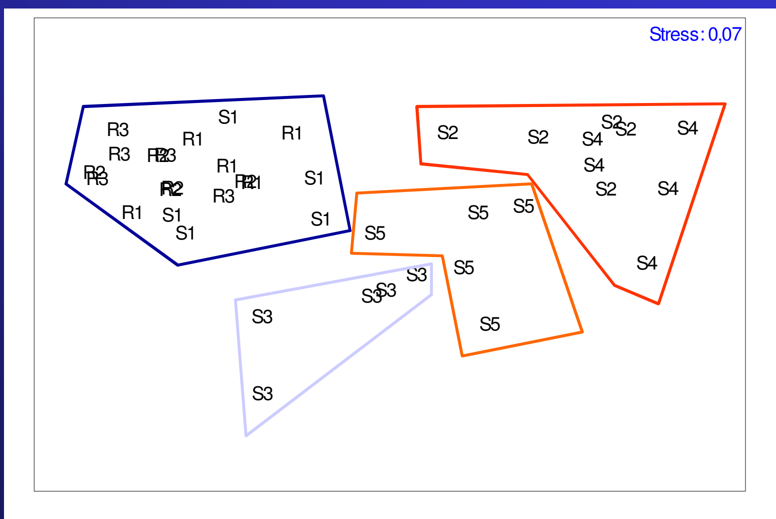
Look at "pairwise" tests !

## "ANOSIM" in Primer v.5




© JPSousa

## "ANOSIM" in Primer v.5




© JPSousa

## Jaccard Coefficient

- only shared presences contribute to similarity
- ignores shared absences or 0-0 matches

$$S^{J} = \frac{a}{a+b+c}$$

| | | SU 2 Present | SU 2 Absent |
|---|---|---|---|
| SU 1 | Present | a | b |
| | Absent | c | d |



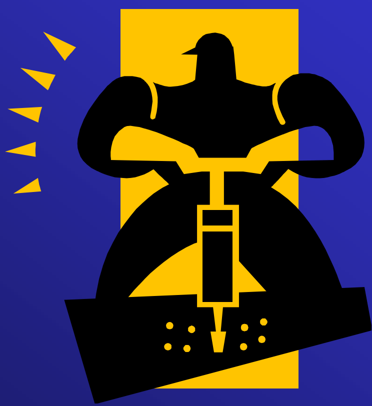
## Jaccard Coefficient

- number of shared species as proportion of total number of species in the two SUs
- ranges from 0 (no species in common) to 1 (the SUs have identical species lists)

$$S^{J} = \frac{a}{a+b+c}$$

| | | SU 2 Present | SU 2 Absent |
|---|---|---|---|
| SU 1 | Present | a | b |
| | Absent | c | d |

## NMDS & ANOSIM



Hands On !
Part 3



Multivariate Statistical Tools in Ecology
ISCED, Lubango, March 2016

## Ordination Tools II:
## Discriminating groups of samples/subjects
## Part 2

# DISCRIMINANT ANALYSIS

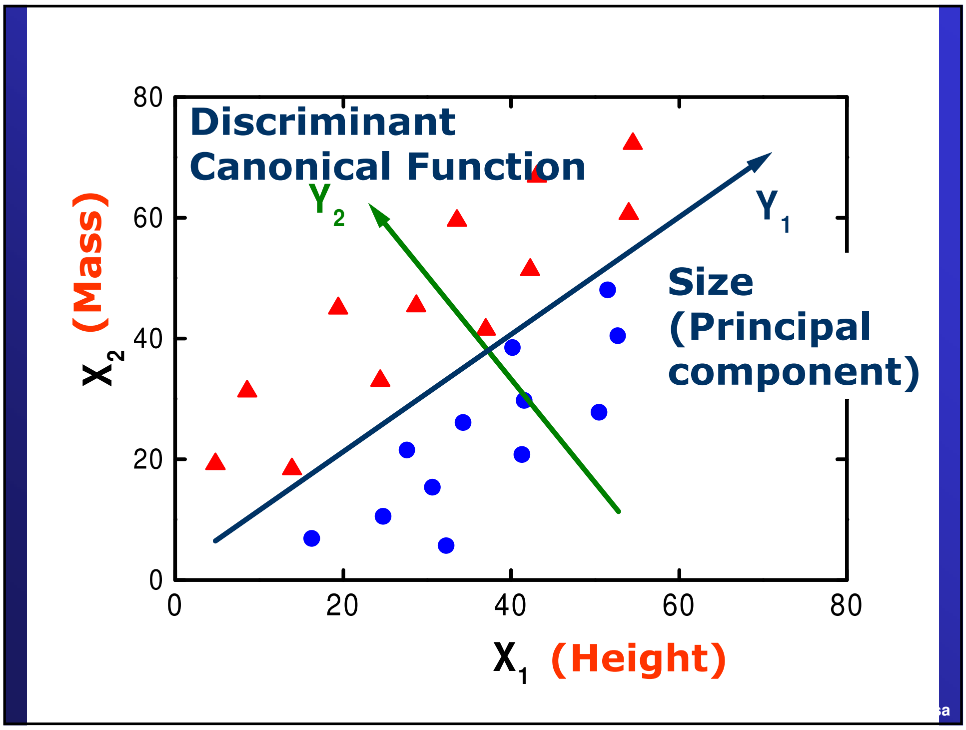
## What for ?

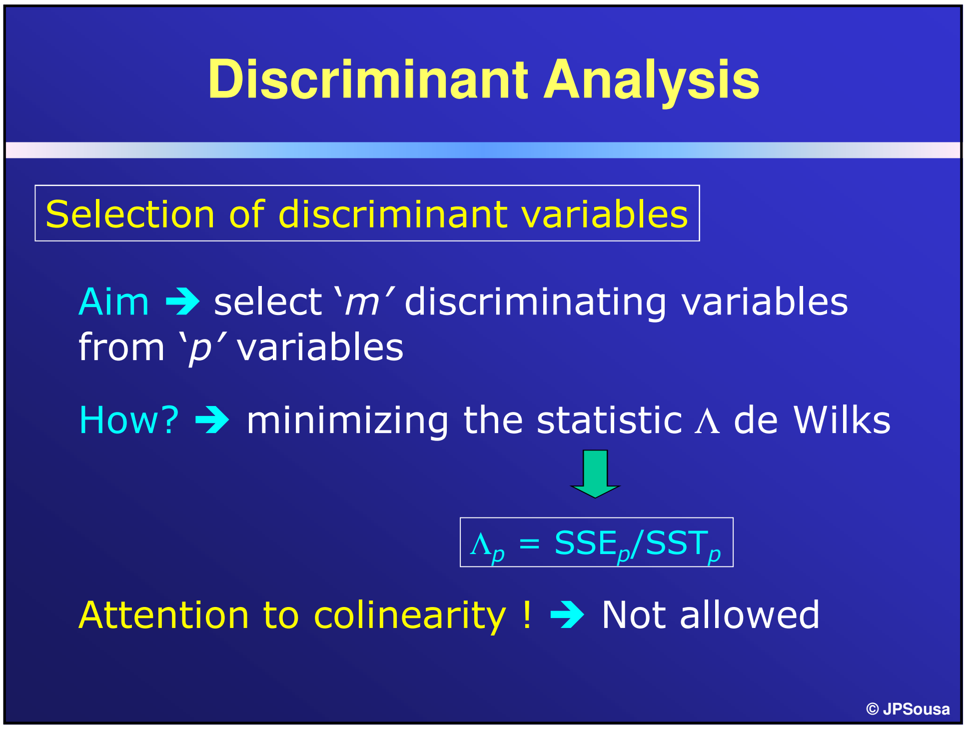
- Discriminate groups of samples (e.g., treatments, sites, etc)

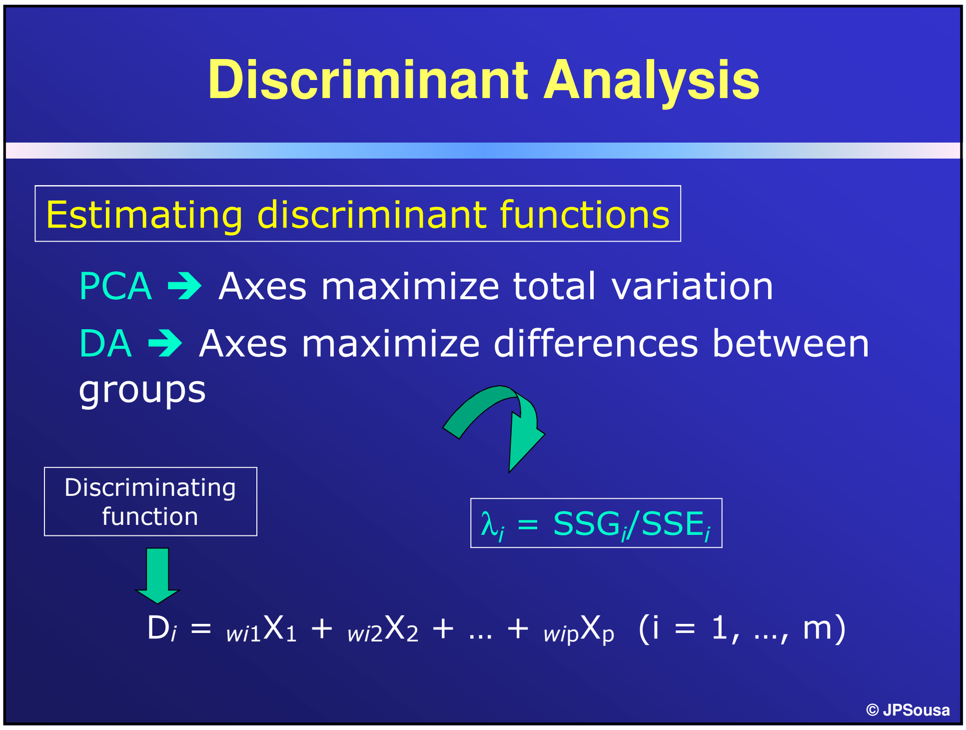
## How does it work ?

- Identification of discriminant variables
- Use of those variables to create "discriminant functions"
- Discriminate groups according to those functions



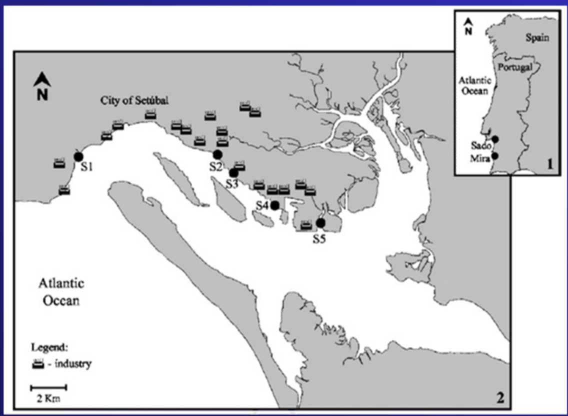



Discriminant Analysis
Selection of discriminant variables
Aim ➔ select 'm' discriminating variables from 'p' variables
How? ➔ minimizing the statistic Λ de Wilks
$\Lambda_p$ = $SSE_p/SST_p$
Attention to colinearity ! ➔ Not allowed
© JPSousa


Discriminant Analysis
Estimating discriminant functions
PCA ➔ Axes maximize total variation
DA ➔ Axes maximize differences between groups
Discriminating function
$\lambda_i$ = $SSG_i/SSE_i$
$D_i = w_{i1}X_1 + w_{i2}X_2 + \ldots + w_{ip}X_p$ (i = 1, ..., m)
© JPSousa

# Discriminant Analysis (Statistica)

Example




Moreira et.al (2006) Aquatic Toxicology

**Physiological functions in the polychaete *Hediste diversicolor***

Measurements of several enzyme biomarkers (neurotransmission, metabolic condition, detoxification process, antioxidant defences)

Reference estuary – Rio Mira

Impacted estuary - Rio Sado

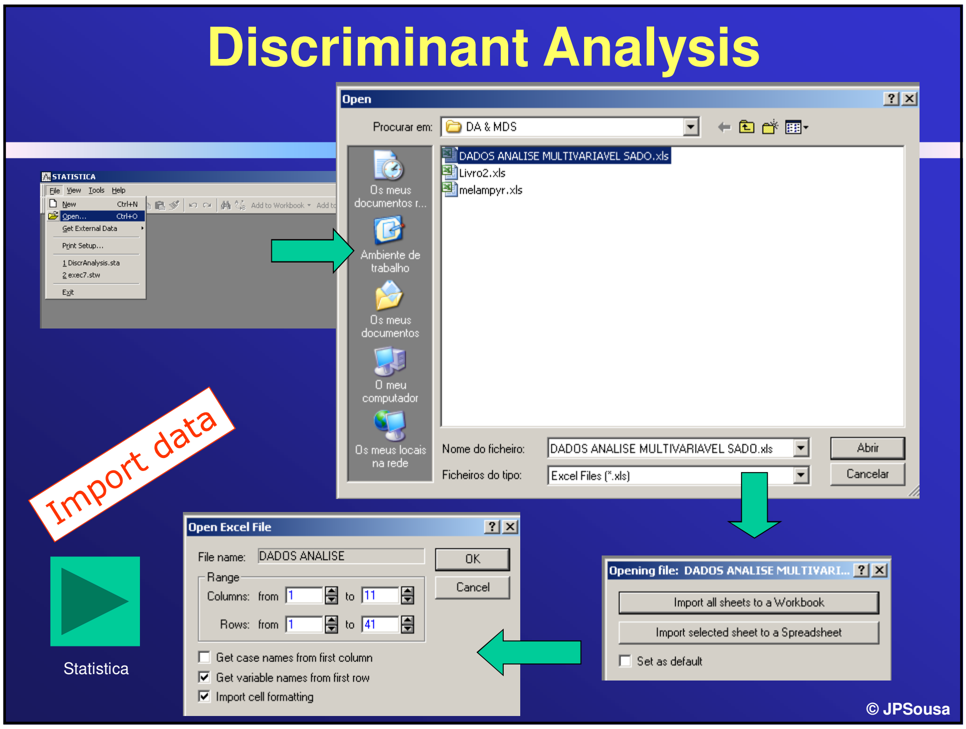
Several sampling sites with several animals at each one of them

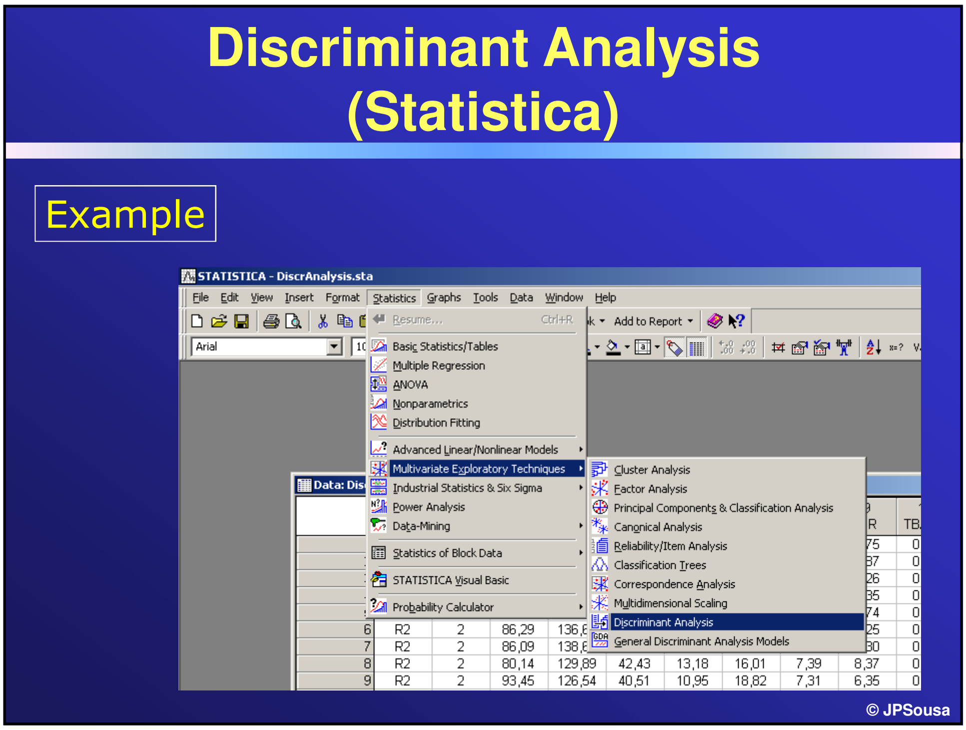


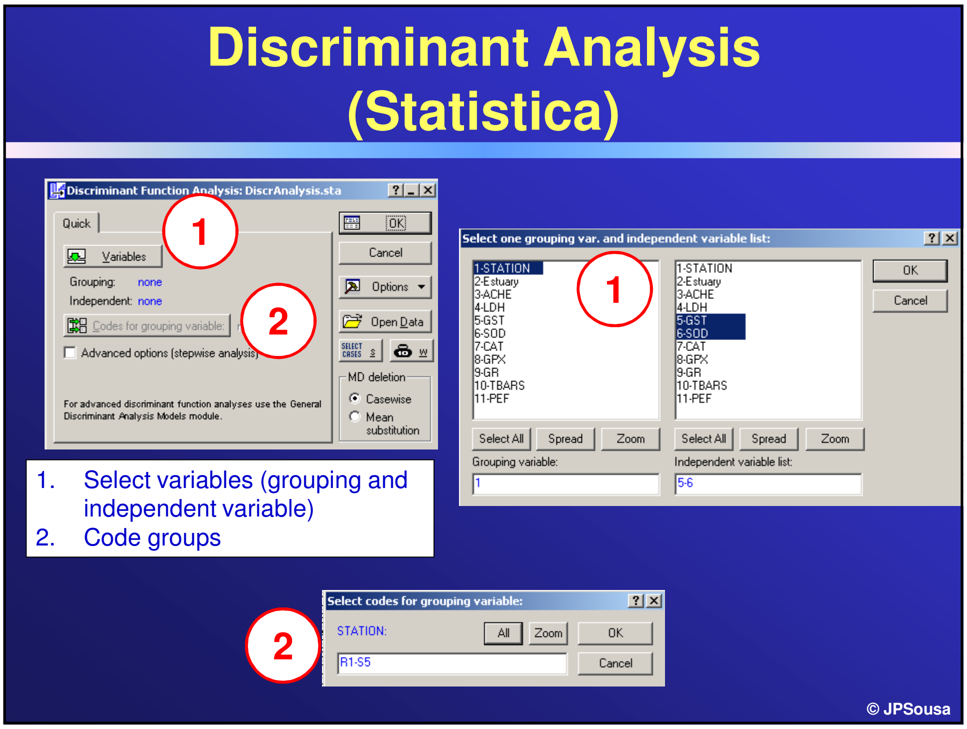
# Discriminant Analysis (Statistica)

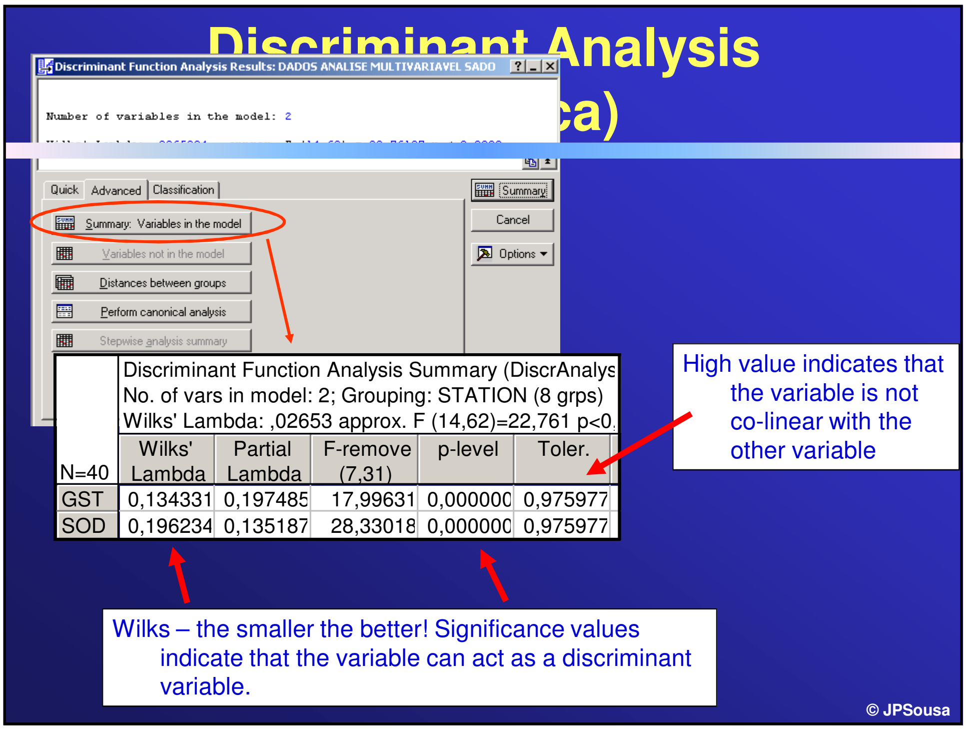
Example

| STATION | Estuary | ACHE | LDH | GST | SOD | CAT | GPX | GR | TBARS |
|---|---|---|---|---|---|---|---|---|---|
| R1 | 1 | 81,11 | 169,95 | 42,53 | 14,98 | 13,79 | 7,37 | 7,75 | 0,59 |
| R1 | 1 | 82,36 | 149,19 | 40,73 | 14,68 | 19,21 | 8,67 | 9,87 | 0,58 |
| R1 | 1 | 88,46 | 157,54 | 39,95 | 15,46 | 18,29 | 8,95 | 9,26 | 0,54 |
| R1 | 1 | 91,88 | 126,71 | 34,33 | 17,90 | 13,57 | 7,16 | 7,35 | 0,32 |
| R1 | 1 | 88,25 | 155,70 | 42,14 | 17,60 | 18,97 | 7,76 | 7,74 | 0,52 |
| R2 | 2 | 86,29 | 136,63 | 36,62 | 15,28 | 14,96 | 6,09 | 7,25 | 0,56 |
| R2 | 2 | 86,09 | 138,67 | 45,02 | 11,20 | 16,82 | 8,35 | 6,30 | 0,65 |
| R2 | 2 | 80,14 | 129,89 | 42,43 | 13,18 | 16,01 | 7,39 | 8,37 | 0,62 |
| R2 | 2 | 93,45 | 126,54 | 40,51 | 10,95 | 18,82 | 7,31 | 6,35 | 0,32 |
| R2 | 2 | 84,54 | 151,77 | 39,88 | 17,17 | 14,31 | 7,34 | 9,36 | 0,24 |
| R3 | 3 | 88,36 | 136,63 | 35,70 | 7,88 | 13,70 | 8,02 | 6,10 | 0,42 |
| R3 | 3 | 84,79 | 138,67 | 38,56 | 13,23 | 17,65 | 6,80 | 7,84 | 0,31 |
| R3 | 3 | 87,81 | 129,89 | 39,45 | 12,96 | 17,37 | 8,16 | 6,88 | 0,55 |
| R3 | 3 | 93,63 | 126,54 | 39,14 | 11,95 | 13,76 | 7,65 | 6,62 | 0,54 |
| R3 | 3 | 90,06 | 149,45 | 41,22 | 20,14 | 13,83 | 7,74 | 9,27 | 0,41 |
| S1 | 4 | 86,23 | 136,11 | 45,15 | 23,75 | 19,04 | 6,43 | 6,43 | 0,44 |
| S1 | 4 | 92,65 | 135,83 | 39,04 | 17,60 | 13,78 | 6,54 | 8,01 | 0,46 |
| S1 | 4 | 79,78 | 164,13 | 43,07 | 19,80 | 14,51 | 8,69 | 9,00 | 0,46 |
| S1 | 4 | 76,52 | 148,61 | 36,09 | 12,99 | 14,27 | 5,92 | 6,40 | 0,51 |
| S1 | 4 | 88,83 | 152,64 | 38,52 | 23,05 | 13,34 | 8,83 | 7,24 | 0,59 |
| S2 | 5 | 91,02 | 214,79 | 42,19 | 46,06 | 23,85 | 9,12 | 6,44 | 0,74 |
| S2 | 5 | 93,88 | 178,86 | 39,69 | 41,72 | 22,26 | 11,31 | 8,21 | 0,77 |
| S2 | 5 | 83,59 | 213,75 | 45,53 | 44,91 | 21,79 | 10,96 | 8,75 | 1,15 |
| S2 | 5 | 89,84 | 179,34 | 43,09 | 40,18 | 20,25 | 9,70 | 5,66 | 0,75 |


Discriminant Analysis
Open
Procurar em: DA & MDS
DADOS ANALISE MULTIVARIAVEL SADO.xls
Livro2.xls
melampyr.xls
Os meus documentos r...
Ambiente de trabalho
Os meus documentos
O meu computador
Os meus locais na rede
Nome do ficheiro: DADOS ANALISE MULTIVARIAVEL SADO.xls
Ficheiros do tipo: Excel Files (*.xls)
Abrir
Cancelar
Import data
Open Excel File
File name: DADOS ANALISE
OK
Cancel
Range
Columns: from 1 to 11
Rows: from 1 to 41
Get case names from first column
Get variable names from first row
Import cell formatting
Opening file: DADOS ANALISE MULTIVARI...
Import all sheets to a Workbook
Import selected sheet to a Spreadsheet
Set as default
Statistica
© JPSousa

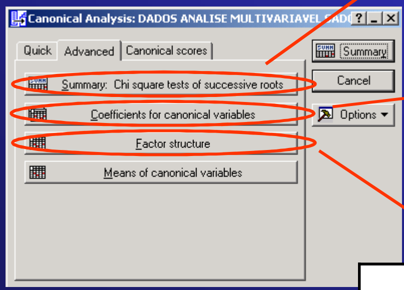

Discriminant Analysis (Statistica)
Example
STATISTICA - DiscrAnalysis.sta
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Resume... Ctrl+R
Add to Report
Basic Statistics/Tables
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
Advanced Linear/Nonlinear Models
Multivariate Exploratory Techniques
Industrial Statistics & Six Sigma
Power Analysis
Data-Mining
Statistics of Block Data
STATISTICA Visual Basic
Probability Calculator
Cluster Analysis
Factor Analysis
Principal Components & Classification Analysis
Canonical Analysis
Reliability/Item Analysis
Classification Trees
Correspondence Analysis
Multidimensional Scaling
Discriminant Analysis
General Discriminant Analysis Models
Data: Dis
© JPSousa


Discriminant Analysis (Statistica)
Discriminant Function Analysis: DiscrAnalysis.sta
Quick
OK
Cancel
Variables
Grouping: none
Independent: none
Codes for grouping variable:
Advanced options (stepwise analysis)
Options
Open Data
MD deletion
Casewise
Mean substitution
For advanced discriminant function analyses use the General Discriminant Analysis Models module.
1
2
Select one grouping var. and independent variable list:
1-STATION
2-Estuary
3-ACHE
4-LDH
5-GST
6-SOD
7-CAT
8-GPX
9-GR
10-TBARS
11-PEF
Select All
Spread
Zoom
Grouping variable:
Independent variable list:
5-6
1. Select variables (grouping and independent variable)
2. Code groups
Select codes for grouping variable:
STATION:
All
R1-S5
© JPSousa


Discriminant Analysis (Statistica)
Discriminant Function Analysis Results: DADOS ANALISE MULTIVARIAVEL SADO
Number of variables in the model: 2
Quick
Advanced
Classification
Summary
Cancel
Options
Summary: Variables in the model
Variables not in the model
Distances between groups
Perform canonical analysis
Stepwise analysis summary
Discriminant Function Analysis Summary (DiscrAnalys
No. of vars in model: 2; Grouping: STATION (8 grps)
Wilks' Lambda: ,02653 approx. F (14,62)=22,761 p<0.
N=40
Wilks' Lambda
Partial Lambda
F-remove (7,31)
p-level
Toler.
GST 0,134331 0,197485 17,99631 0,000000 0,975977
SOD 0,196234 0,135187 28,33018 0,000000 0,975977
High value indicates that the variable is not co-linear with the other variable
Wilks – the smaller the better! Significance values indicate that the variable can act as a discriminant variable.
© JPSousa

# Discriminant Analysis (Statistica)

| Roots Removed | Chi-Square Tests with Successive Roots Removed (DAD Eigen-value | Canonicl R | Wilks' Lambda | Chi-Sqr. | df | p-level |
|---|---|---|---|---|---|---|
| 0 | 6,447271 | 0,930442 | 0,026528 | 123,4044 | 14 | 0,000000 |
| 1 | 4,061649 | 0,895788 | 0,197564 | 55,1375 | 6 | 0,000000 |




| Variable | Standardized Coeffic for Canonical Variab Root 1 | Root 2 |
|---|---|---|
| GST | 0,03584 | -1,01160 |
| SOD | -1,00493 | 0,12138 |
| Eigenval | 6,44727 | 4,06165 |
| Cum.Prop | 0,61350 | 1,00000 |

| Variable | Factor Structure Matrix Correlations Variables (Pooled-within-groups Root 1 | Root 2 |
|---|---|---|
| GST | -0,119915 | -0,992784 |
| SOD | -0,999373 | -0,035410 |

© JPSousa

# Discriminant Analysis (Statistica)

| Roots removed | Eigen- | Canonicl | Wilks' | Chi-Sqr. | df | p-level |
|---|---|---|---|---|---|---|
| 0 | 6,447271 | 0,930442 | 0,026528 | 123,4044 | 14 | 0,000000 |
| 1 | 4,061649 | 0,895788 | 0,197564 | 55,1375 | 6 | 0,000000 |

High value indicates that the discriminant variables can discriminate the groups well

Wilks – the smallest the better! Significance values indicate that the groups along the two discriminant functions are really discriminated (H0 rejected). Each discriminating function is significant

© JPSousa

# Discriminant Analysis (Statistica)

| Variable | Standardized Coeffic for Canonical Variab Root 1 | Root 2 |
|---|---|---|
| GST | 0,03584 | -1,01160 |
| SOD | -1,00493 | 0,12138 |
| Eigenval | 6,44727 | 4,06165 |
| Cum.Prop | 0,61350 | 1,00000 |

**Standardized Coefficients**
Indicate the contribution of each variable for the definition of the discriminant function

SOD is more important along discriminant function 1
GST is more important along discriminant function 2

| Variable | Factor Structure Matrix Correlations Variables (Pooled-within-groups Root 1 | Root 2 |
|---|---|---|
| GST | -0,119915 | -0,992784 |
| SOD | -0,999373 | -0,035410 |

**Canonical Correlation Coefficients**
Indicate the correlation of each variable with the discriminant function



# Discriminant analysis (S

Discriminant Function Analysis Results: DADOS ANA

```
Number of variables in the model: 2

Wilks' Lambda: ,0265284   approx. F (14,62)
```

Quick | Advanced | Classification

Summary: Variables in the model
Variables not in the model
Distances between groups
Perform canonical analysis
Stepwise analysis summary

Squared Mahalanobis Distances (DADOS ANALISE MULTIVARIAVEL SADO)

| STATION | R1 | R2 | R3 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|---|
| R1 | 0,00000 | 0,61053 | 0,61851 | 0,73247 | 43,30727 | 38,07450 | 27,63473 | 9,61372 |
| R2 | 0,61053 | 0,00000 | 0,43714 | 2,47374 | 52,74130 | 36,16661 | 34,79658 | 9,68986 |
| R3 | 0,61851 | 0,43714 | 0,00000 | 2,63411 | 53,71437 | 44,22431 | 36,31717 | 13,59249 |
| S1 | 0,73247 | 2,47374 | 2,63411 | 0,00000 | 32,77863 | 35,69392 | 19,40099 | 8,02383 |
| S2 | 43,30727 | 52,74130 | 53,71437 | 32,77863 | 0,00000 | 59,30965 | 2,15763 | 35,94739 |
| S3 | 38,07450 | 36,16661 | 44,22431 | 35,69392 | 59,30965 | 0,00000 | 41,43731 | 9,87520 |
| S4 | 27,63473 | 34,79658 | 36,31717 | 19,40099 | 2,15763 | 41,43731 | 0,00000 | 20,56692 |
| S5 | 9,61372 | 9,68986 | 13,59249 | 8,02383 | 35,94739 | 9,87520 | 20,56692 | 0,00000 |

F-values; df = 2,31 (DADOS ANALISE MULTIVARIAVEL SADO)

| STATION | R1 | R2 | R3 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|---|
| R1 | | 0,59145 | 0,59918 | 0,70958 | 41,95391 | 36,88467 | 26,77114 | 9,31329 |
| R2 | 0,59145 | | 0,42348 | 2,39643 | 51,09314 | 35,03641 | 33,70918 | 9,38705 |
| R3 | 0,59918 | 0,42348 | | 2,55180 | 52,03579 | 42,84230 | 35,18225 | 13,16772 |
| S1 | 0,70958 | 2,39643 | 2,55180 | | 31,75429 | 34,57849 | 18,79471 | 7,77309 |
| S2 | 41,95391 | 51,09314 | 52,03579 | 31,75429 | | 57,45622 | 2,09021 | 34,82404 |
| S3 | 36,88467 | 35,03641 | 42,84230 | 34,57849 | 57,45622 | | 40,14239 | 9,56660 |
| S4 | 26,77114 | 33,70918 | 35,18225 | 18,79471 | 2,09021 | 40,14239 | | 19,92420 |
| S5 | 9,31329 | 9,38705 | 13,16772 | 7,77309 | 34,82404 | 9,56660 | 19,92420 | |

p-levels (DADOS ANALISE MULTIVARIAVEL SADO)

| STATION | R1 | R2 | R3 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|---|
| R1 | | 0,559647 | 0,555500 | 0,499664 | 0,000000 | 0,000000 | 0,000000 | 0,000680 |
| R2 | 0,559647 | | 0,658498 | 0,107711 | 0,000000 | 0,000000 | 0,000000 | 0,000649 |
| R3 | 0,555500 | 0,658498 | | 0,094205 | 0,000000 | 0,000000 | 0,000000 | 0,000073 |
| S1 | 0,499664 | 0,107711 | 0,094205 | | 0,000000 | 0,000000 | 0,000005 | 0,001836 |
| S2 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | | 0,000000 | 0,140748 | 0,000000 |
| S3 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | | 0,000000 | 0,000581 |
| S4 | 0,000000 | 0,000000 | 0,000000 | 0,000005 | 0,140748 | 0,000000 | | 0,000003 |
| S5 | 0,000680 | 0,000649 | 0,000073 | 0,001836 | 0,000000 | 0,000581 | 0,000003 | |

# Discriminant Analysis (Statistica)

Squared Mahalanobis Distances (DiscrAnalysis.sta)

| | R1 | R2 | R3 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|---|
| R1 | 0,00000 | 0,61053 | 0,61851 | 0,73247 | 43,30727 | 38,07450 | 27,63473 | 9,61372 |
| R2 | 0,61053 | 0,00000 | 0,43714 | 2,47374 | 52,74130 | 36,16661 | 34,79658 | 9,68986 |
| R3 | 0,61851 | 0,43714 | 0,00000 | 2,63411 | 53,71437 | 44,22431 | 36,31717 | 13,59249 |
| S1 | 0,73247 | 2,47374 | 2,63411 | 0,00000 | 32,77863 | 35,69392 | 19,40099 | 8,02383 |
| S2 | 43,30727 | 52,74130 | 53,71437 | 32,77863 | 0,00000 | 59,30965 | 2,15763 | 35,94739 |
| S3 | 38,07450 | 36,16661 | 44,22431 | 35,69392 | 59,30965 | 0,00000 | 41,43731 | 9,87520 |
| S4 | 27,63473 | 34,79658 | 36,31717 | 19,40099 | 2,15763 | 41,43731 | 0,00000 | 20,56692 |
| S5 | 9,61372 | 9,68986 | 13,59249 | 8,02383 | 35,94739 | 9,87520 | 20,56692 | 0,00000 |

F-values; df = 2,31 (DiscrAnalysis.sta)

| | R1 | R2 | R3 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|---|
| R1 | | 0,59145 | 0,59918 | 0,70958 | 41,95391 | 36,88467 | 26,77114 | 9,31329 |
| R2 | 0,59145 | | 0,42348 | 2,39643 | 51,09314 | 35,03641 | 33,70918 | 9,38705 |
| R3 | 0,59918 | 0,42348 | | 2,55180 | 52,03579 | 42,84230 | 35,18225 | 13,16772 |
| S1 | 0,70958 | 2,39643 | 2,55180 | | 31,75429 | 34,57849 | 18,79471 | 7,77309 |
| S2 | 41,95391 | 51,09314 | 52,03579 | 31,75429 | | 57,45622 | 2,09021 | 34,82404 |
| S3 | 36,88467 | 35,03641 | 42,84230 | 34,57849 | 57,45622 | | 40,14239 | 9,56660 |
| S4 | 26,77114 | 33,70918 | 35,18225 | 18,79471 | 2,09021 | 40,14239 | | 19,92420 |
| S5 | 9,31329 | 9,38705 | 13,16772 | 7,77309 | 34,82404 | 9,56660 | 19,92420 | |

p-levels (DiscrAnalysis.sta)

| | R1 | R2 | R3 | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|---|---|---|
| R1 | | 0,559647 | 0,555500 | 0,499664 | 0,000000 | 0,000000 | 0,000000 | 0,000680 |
| R2 | 0,559647 | | 0,658498 | 0,107711 | 0,000000 | 0,000000 | 0,000000 | 0,000649 |
| R3 | 0,555500 | 0,658498 | | 0,094205 | 0,000000 | 0,000000 | 0,000000 | 0,000073 |
| S1 | 0,499664 | 0,107711 | 0,094205 | | 0,000000 | 0,000000 | 0,000005 | 0,001836 |
| S2 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | | 0,000000 | 0,140748 | 0,000000 |
| S3 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | | 0,000000 | 0,000581 |
| S4 | 0,000000 | 0,000000 | 0,000000 | 0,000005 | 0,140748 | 0,000000 | | 0,000003 |
| S5 | 0,000680 | 0,000649 | 0,000073 | 0,001836 | 0,000000 | 0,000581 | 0,000003 | |

**Group 1:**
R1=R2=R3=S1

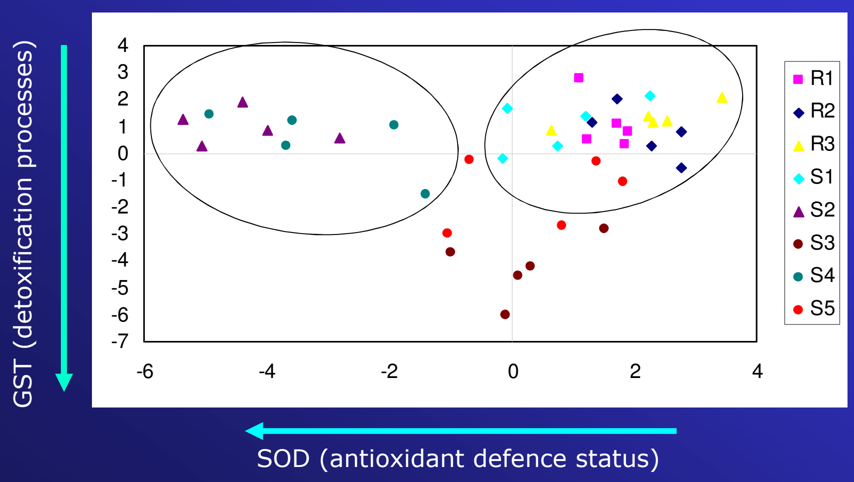
**Group 2:**
S2=S4

**Group 3:**
S3

**Group 4:**
S5

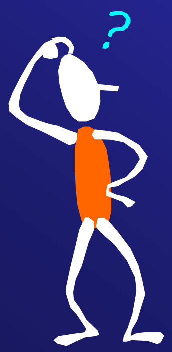


# Discriminant Analysis (Statistica)



## Discriminant Analysis (Statistica)

Perform a setpwise analysis using all variables



Can we have a better discrimination?



## Discriminant Analysis (Statistica)

| | Wilks' | Partial | F-remove | p-level | Toler. | 1-Toler. |
|---|---|---|---|---|---|---|
| PEF | 0,002117 | 0,427767 | 4,777580 | 0,001608 | 0,704428 | 0,295572 |
| GST | 0,002235 | 0,405169 | 5,243232 | 0,000888 | 0,864383 | 0,135617 |
| SOD | 0,002240 | 0,404305 | 5,262073 | 0,000867 | 0,809208 | 0,190792 |
| GPX | 0,002373 | 0,381509 | 5,789902 | 0,000456 | 0,821138 | 0,178862 |
| TBARS | 0,001204 | 0,752326 | 1,175754 | 0,351494 | 0,873572 | 0,126428 |
| ACHE | 0,001425 | 0,635340 | 2,049863 | 0,088144 | 0,507946 | 0,492054 |
| LDH | 0,001274 | 0,710864 | 1,452639 | 0,229325 | 0,605079 | 0,394921 |
| CAT | 0,001190 | 0,760971 | 1,121823 | 0,380890 | 0,846854 | 0,153146 |

These 4 variables are not significant

# Discriminant Analysis (Statistica)

| Roots removed | Eigen- | Canonicl | Wilks' | Chi-Sqr. | df | p-level |
|---|---|---|---|---|---|---|
| 0 | 22,38807 | 0,978388 | 0,000905 | 217,2191 | 56 | 0,000000 |
| 1 | 10,32528 | 0,954831 | 0,021177 | 119,5001 | 42 | 0,000000 |
| 2 | 1,69329 | 0,792910 | 0,239835 | 44,2620 | 30 | 0,045147 |
| 3 | 0,28793 | 0,472819 | 0,645944 | 13,5483 | 20 | 0,852640 |
| 4 | 0,15967 | 0,371062 | 0,831928 | 5,7043 | 12 | 0,930249 |
| 5 | 0,02912 | 0,168211 | 0,964764 | 1,1120 | 6 | 0,981007 |
| 6 | 0,00719 | 0,084518 | 0,992857 | 0,2222 | 2 | 0,894832 |

Axis 4 and higher are not significant



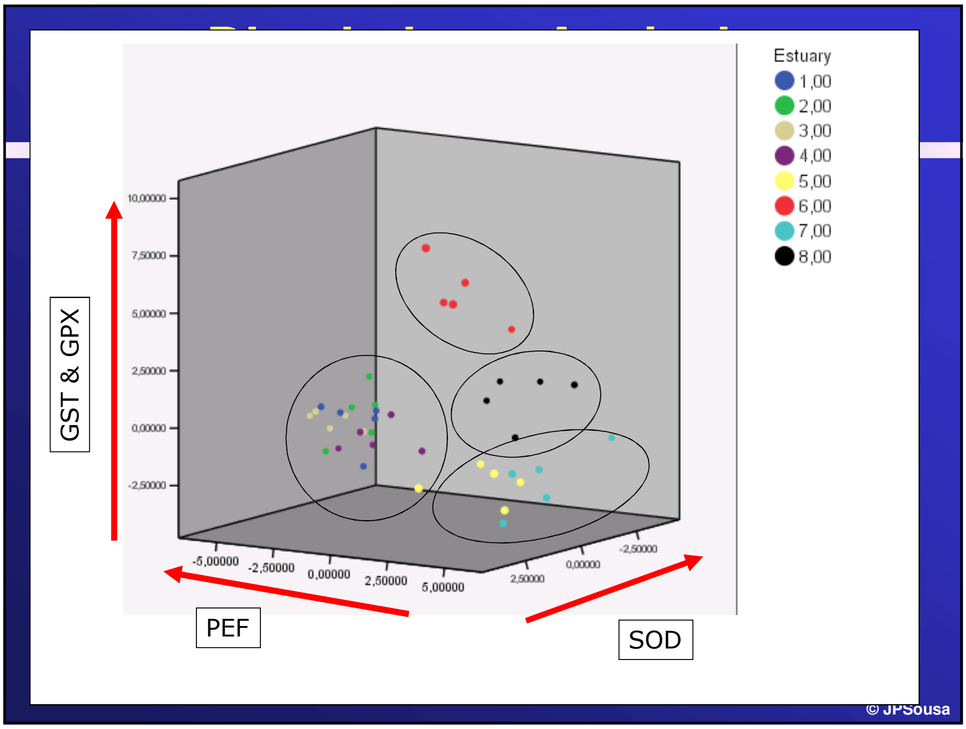
# Discriminant Analysis (Statistica)

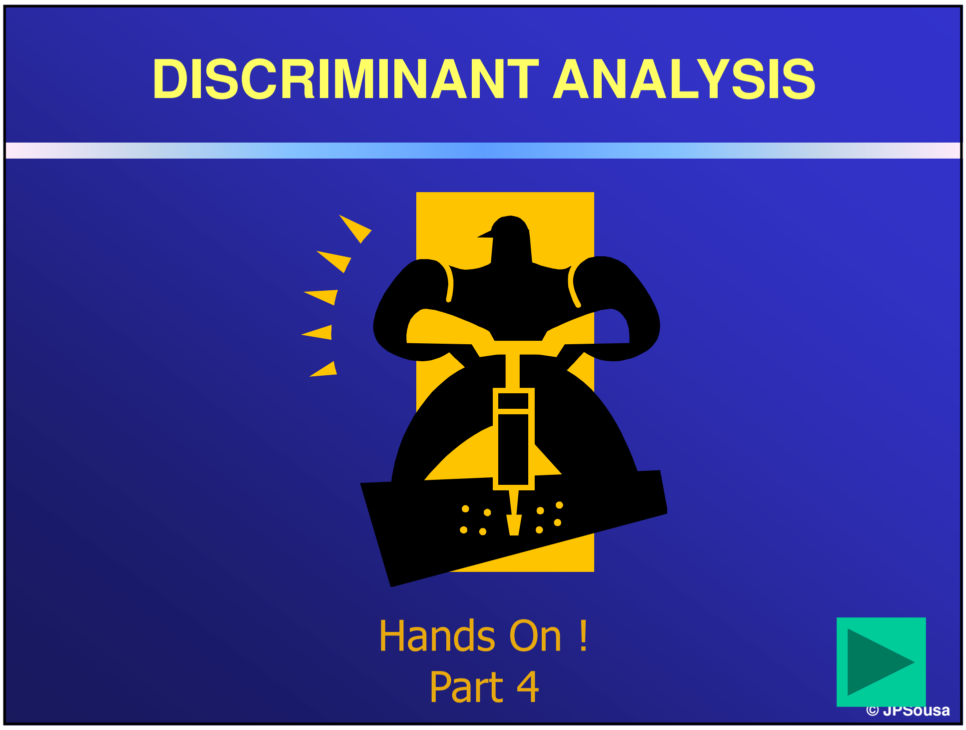
**Std coefficients**

| | Root 1 | Root 2 | Root 3 |
|---|---|---|---|
| PEF | **-0,70350** | 0,30259 | -0,582558 |
| GST | -0,00158 | **0,81686** | 0,306950 |
| SOD | 0,48437 | -0,41628 | **-0,710880** |
| GPX | 0,14087 | **0,80706** | -0,466914 |
| Eigenval | 22,38807 | 10,32528 | 1,693290 |
| Cum.Prop | 0,64167 | 0,93760 | 0,986131 |

**Factor structure matrix**

| | Root 1 | Root 2 | Root 3 |
|---|---|---|---|
| PEF | **-0,631882** | -0,007909 | -0,414397 |
| GST | 0,156296 | **0,570653** | 0,266978 |
| SOD | 0,503298 | -0,101250 | **-0,619335** |
| GPX | 0,106174 | **0,517622** | -0,474907 |


Estuary
1,00
2,00
3,00
4,00
5,00
6,00
7,00
8,00
GST & GPX
PEF
SOD
© JPSousa


DISCRIMINANT ANALYSIS
Hands On !
Part 4
© JPSousa

Multivariate Statistical Tools in Ecology

ISCED, Lubango, March 2016

# Ordination Tools III: Relationship between response variables and explanatory variables



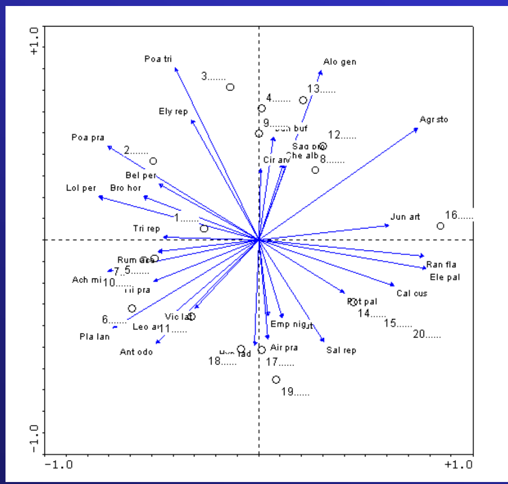
## Relationship between two data sets

### – Indirect Gradient Analysis

- Starts with a normal ordination where the coordinates of a particular axis can be interpreted as an environmental gradient;
- Regression techniques can be used to verify that link between response and explanatory variables;
- No direct input from the explanatory variables in the defining the positions in the ordination plot.

# Indirect Gradient Analysis

(example PCA – vegetation in managed dune systems - Batterink & Wijffels, 1983)

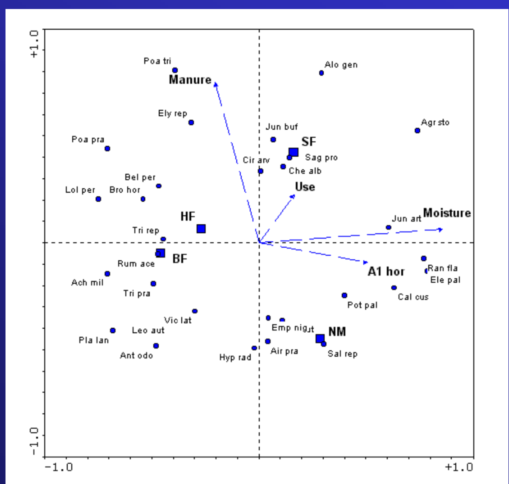





# Indirect Gradient Analysis

(example PCA – vegetation in managed dune systems - Batterink & Wijffels, 1983)

Passive explanatory variables added *a posteriori*



## Indirect Gradient Analysis
(Passive Explanatory Variables)

– Passive explanatory variables help in the interpretation of already extracted axes.

– Passive explanatory variables are projected on top of the ordination plot



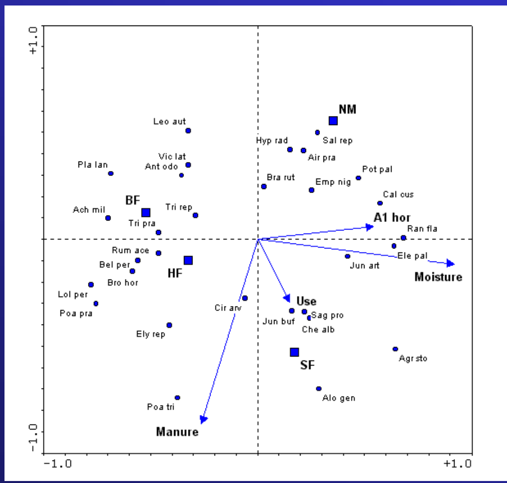
## Relationship between two data sets

– Direct Gradient Analysis
- Used to detect, interpret and predict the underlying structure of the data set based on the explanatory variables (e.g., community composition based on management, land-use, vegetation structure, etc = environmental variables);
- Starts with two datasets that are represented simultaneously in the ordination plot; the relationships between the datasets are derived from that diagram, i.e., the diagram represents the variability explained by the explanatory variables;
- There is a direct input of the explanatory variables in the analysis

# Direct Gradient Analysis

**(example RDA – vegetation in managed dune systems - Batterink & Wijffels, 1983)**

**Explanatory variables are incorporated into the model**






# Relationship between response variables and explanatory variables

| Indirect | Direct |
|---|---|
| Analysis of total variation | Analysis of total variation |
| ⇩ | ⇩ |
| Samples ⇔ Response variables (e.g. species) | Analysis of the variation explained by the explanatory variables |
| ⇩ | ⇩ |
| Interpretation with explanatory variables using *a posteriori* regression / correlation | Samples ⇔ Response variables ⇔ Explanatory variables |

## Relationship between response variables and explanatory variables

**Direct**

Analysis of total variation
↓
Analysis of the variation explained by the explanatory variables
↓
Samples ⇔ Response variables ⇔ Explanatory variables

*Linear response* ← **Redundancy Analysis (RDA)**

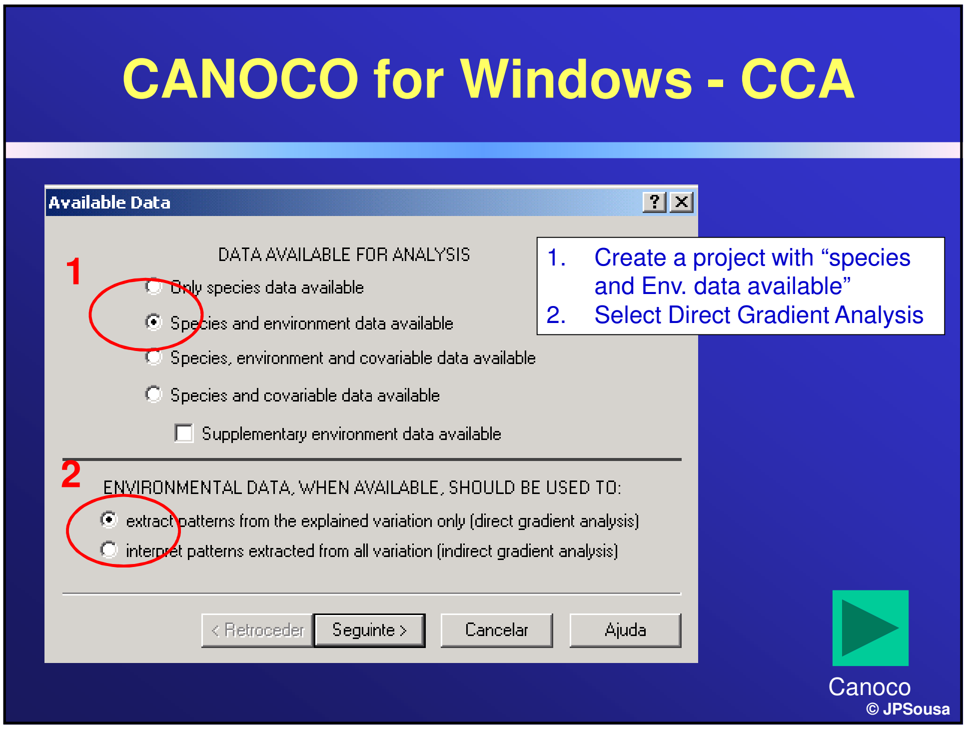
*Unimodal response* ← **Canonical Correspondence Analysis (CCA)**

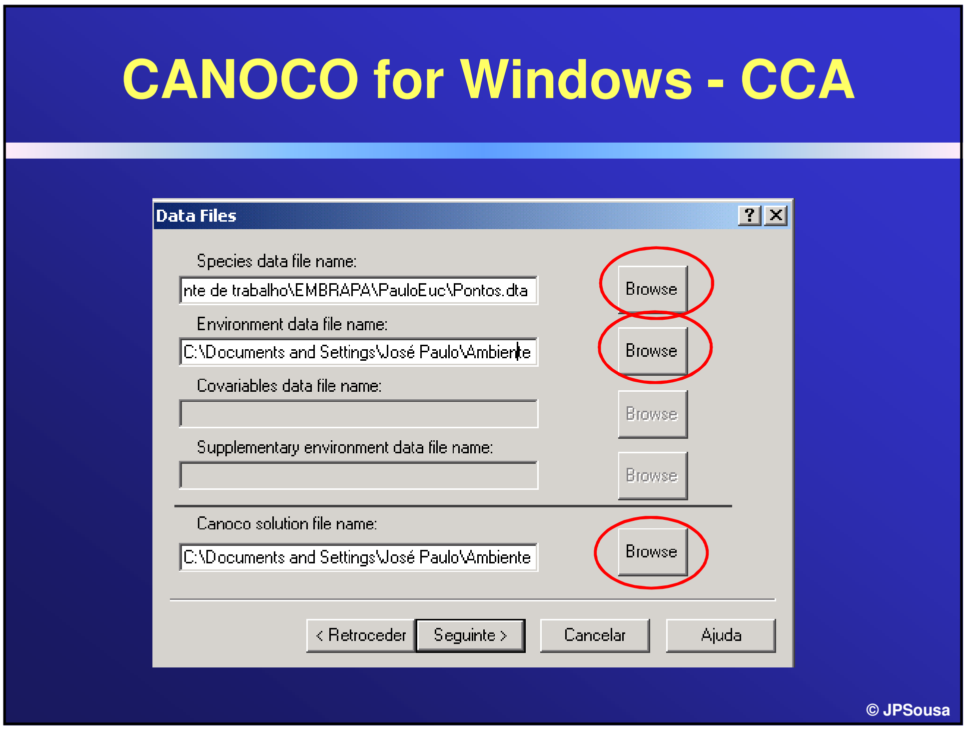


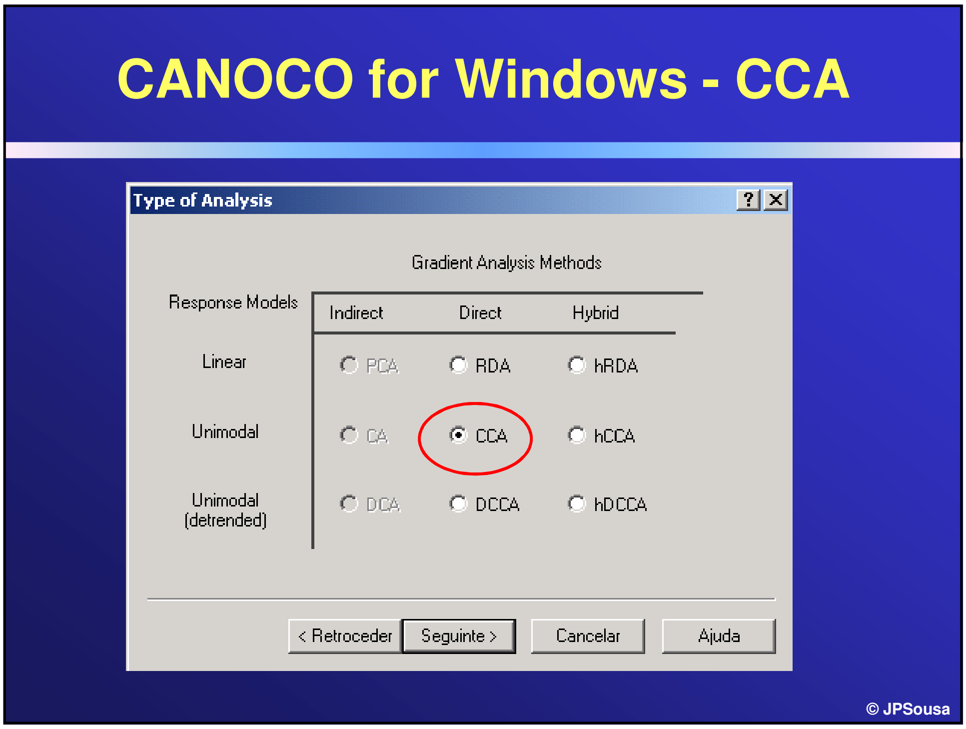
## CANOCO for Windows - CCA

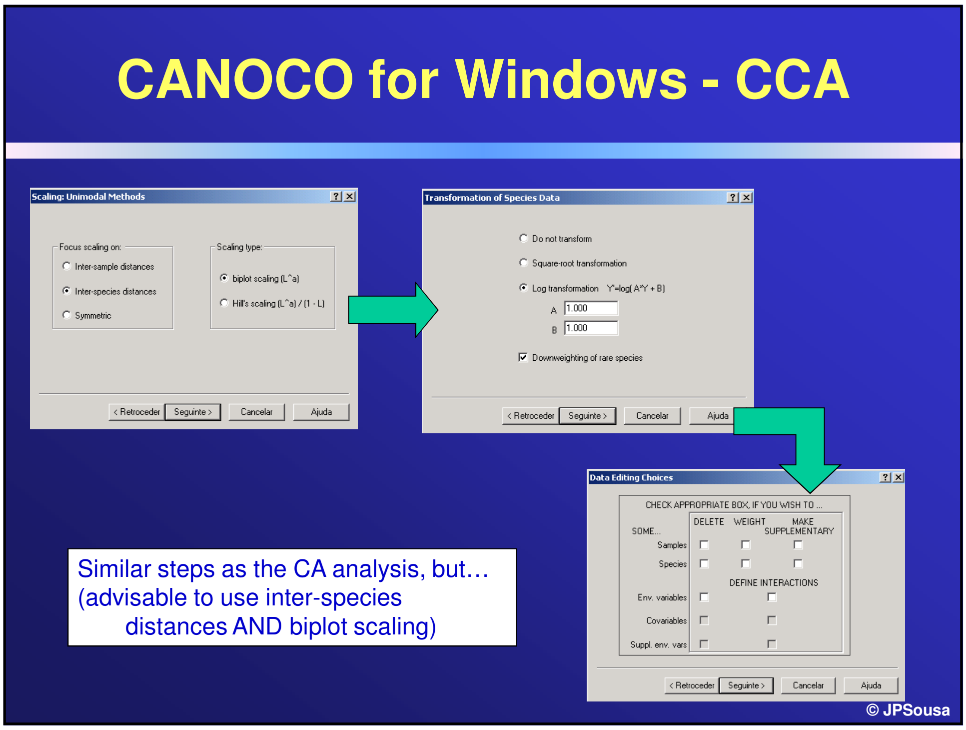
Soil fauna from Cork Oak (*Quercus suber*) and Eucalyptus (*Eucalyptus globulus*) stands (Sousa et al., 2003)

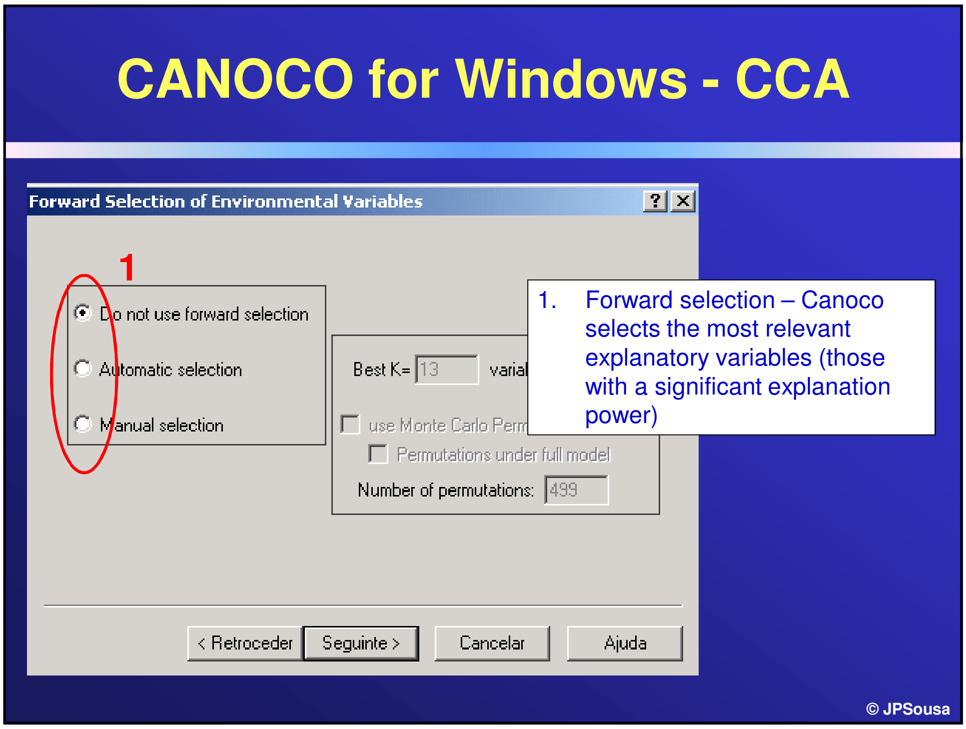
- Soil mesofauna data; soil pedological parameters (File Matrizes_CA_CCA.xls)
- 2 sites (Q e E) with four plots each (A, B, G, M) and each plot with 4 soil cores ;
- 32 samples in total with 45 collembola species identified;
- Objective: to evaluate the association between species and soil parameters

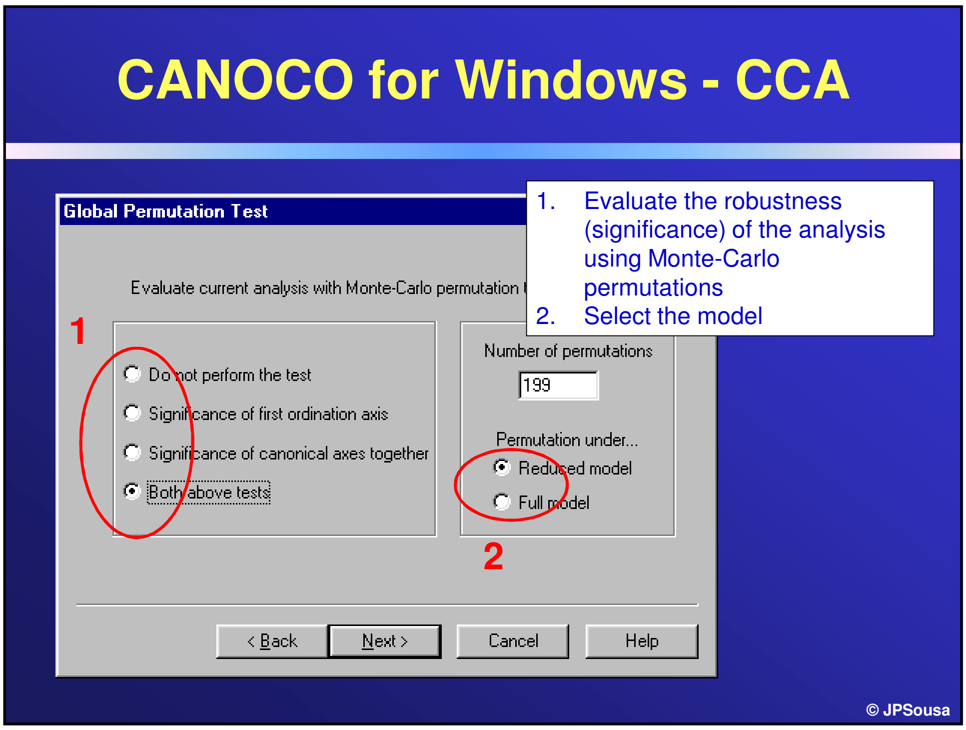

CANOCO for Windows - CCA
Available Data
DATA AVAILABLE FOR ANALYSIS
1
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
2
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Create a project with "species and Env. data available"
2. Select Direct Gradient Analysis
Canoco
© JPSousa


CANOCO for Windows - CCA
Data Files
Species data file name:
nte de trabalho\EMBRAPA\PauloEuc\Pontos.dta
Browse
Environment data file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
Covariables data file name:
Browse
Supplementary environment data file name:
Browse
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa


CANOCO for Windows - CCA
Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa


CANOCO for Windows - CCA
Scaling: Unimodal Methods
Focus scaling on:
Inter-sample distances
Inter-species distances
Symmetric
Scaling type:
biplot scaling (L^a)
Hill's scaling (L^a) / (1 - L)
Transformation of Species Data
Do not transform
Square-root transformation
Log transformation Y'=log( A*Y + B)
A 1.000
B 1.000
Downweighting of rare species
Data Editing Choices
CHECK APPROPRIATE BOX, IF YOU WISH TO ...
DELETE
WEIGHT
MAKE SUPPLEMENTARY
SOME...
Samples
Species
DEFINE INTERACTIONS
Env. variables
Covariables
Suppl. env. vars
< Retroceder
Seguinte >
Cancelar
Ajuda
Similar steps as the CA analysis, but…
(advisable to use inter-species distances AND biplot scaling)
© JPSousa


CANOCO for Windows - CCA
Forward Selection of Environmental Variables
1
Do not use forward selection
Automatic selection
Manual selection
Best K= 13 varial
use Monte Carlo Perm
Permutations under full model
Number of permutations: 499
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Forward selection – Canoco selects the most relevant explanatory variables (those with a significant explanation power)
© JPSousa


CANOCO for Windows - CCA
Global Permutation Test
Evaluate current analysis with Monte-Carlo permutation t
1
Do not perform the test
Significance of first ordination axis
Significance of canonical axes together
Both above tests
Number of permutations
199
Permutation under...
Reduced model
Full model
2
< Back
Next >
Cancel
Help
1. Evaluate the robustness (significance) of the analysis using Monte-Carlo permutations
2. Select the model
© JPSousa

## CANOCO for Windows - CCA

### Test the significance of the first canonical axes

- **Null hypothesis:**
  - Species **ARE NOT** correlated with the environmental variables
- **Is the relation between species and environmental variables stronger than that expected by chance ?**



## CANOCO for Windows - CCA

### Test the significance of the axes: basic ideia

**H0: species not correlated to environment**

- **Calculate F value (F0) for the available data based on the % variance explained**
- **Calculate the reference distribution of F values by permutation (F1…..Fk)**
- **Calculate the significance level:**

**$p$ = (1+ n)/(1+N); n = nº of permutation where F>F0,
N = total nº of permutations**

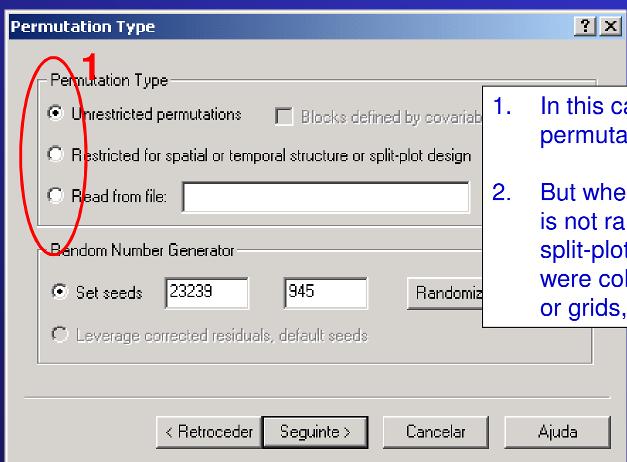
# CANOCO for Windows - CCA

## Full vs. Reduced Model

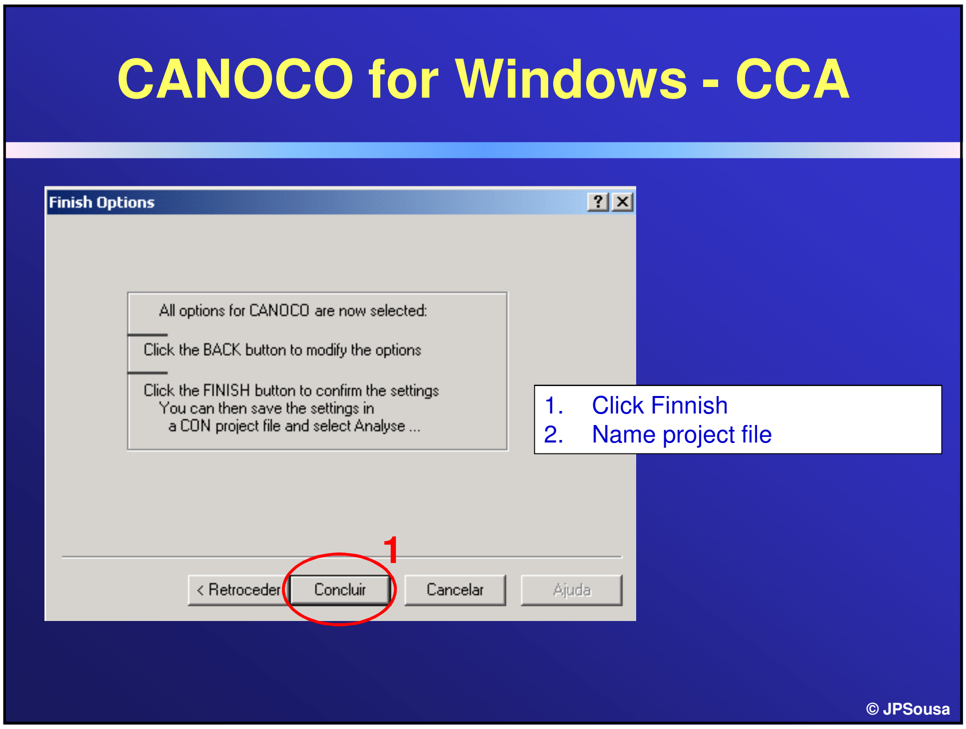
- Use the reduced model:
  - Exact in most situations
  - So powerful as the full-model

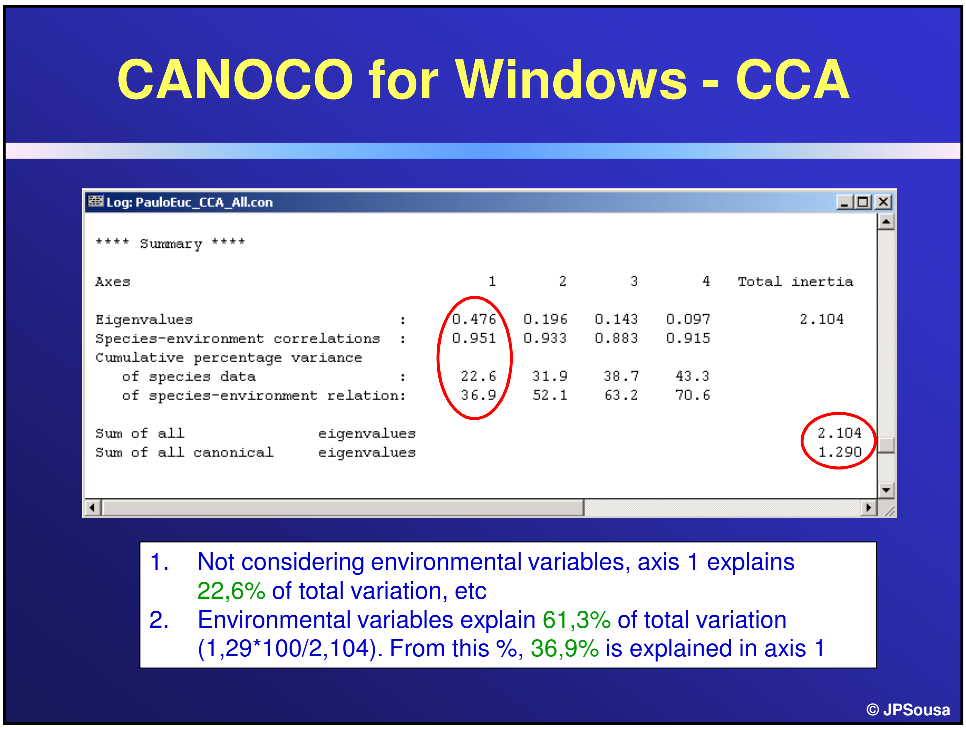


# CANOCO for Windows - CCA




1. In this case use “Unrestricted permutations”
2. But when the sampling design is not ramdom (e.g., blocks, split-plot) or when samples were colected along transepts or grids, use other options.


CANOCO for Windows - CCA
Finish Options
All options for CANOCO are now selected:
Click the BACK button to modify the options
Click the FINISH button to confirm the settings
You can then save the settings in
a CON project file and select Analyse ...
1
< Retroceder
Concluir
Cancelar
Ajuda
1. Click Finnish
2. Name project file
© JPSousa


CANOCO for Windows - CCA
Log: PauloEuc_CCA_All.con
**** Summary ****
Axes 1 2 3 4 Total inertia
Eigenvalues : 0.476 0.196 0.143 0.097 2.104
Species-environment correlations : 0.951 0.933 0.883 0.915
Cumulative percentage variance
of species data : 22.6 31.9 38.7 43.3
of species-environment relation: 36.9 52.1 63.2 70.6
Sum of all eigenvalues 2.104
Sum of all canonical eigenvalues 1.290
1. Not considering environmental variables, axis 1 explains 22,6% of total variation, etc
2. Environmental variables explain 61,3% of total variation (1,29*100/2,104). From this %, 36,9% is explained in axis 1
© JPSousa

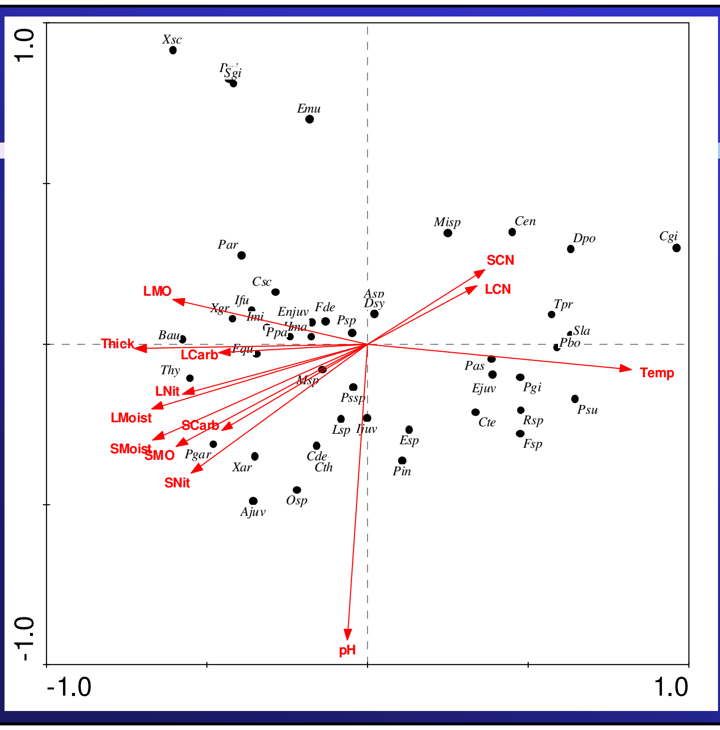
# CANOCO for Windows - CCA

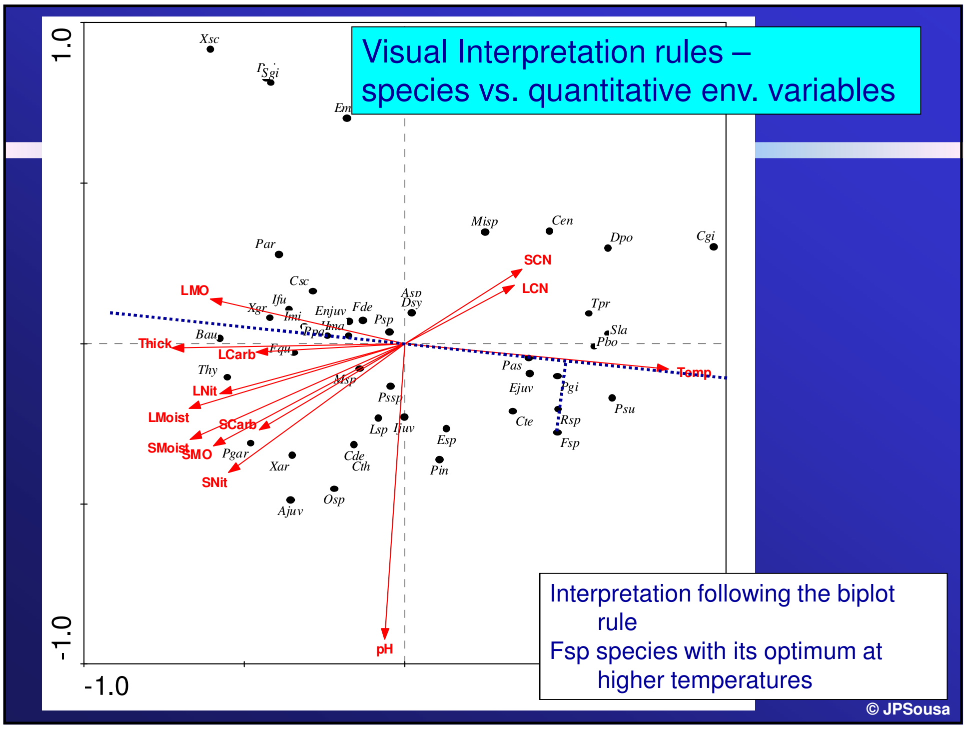
Log: PauloEuc_CCA_All.con

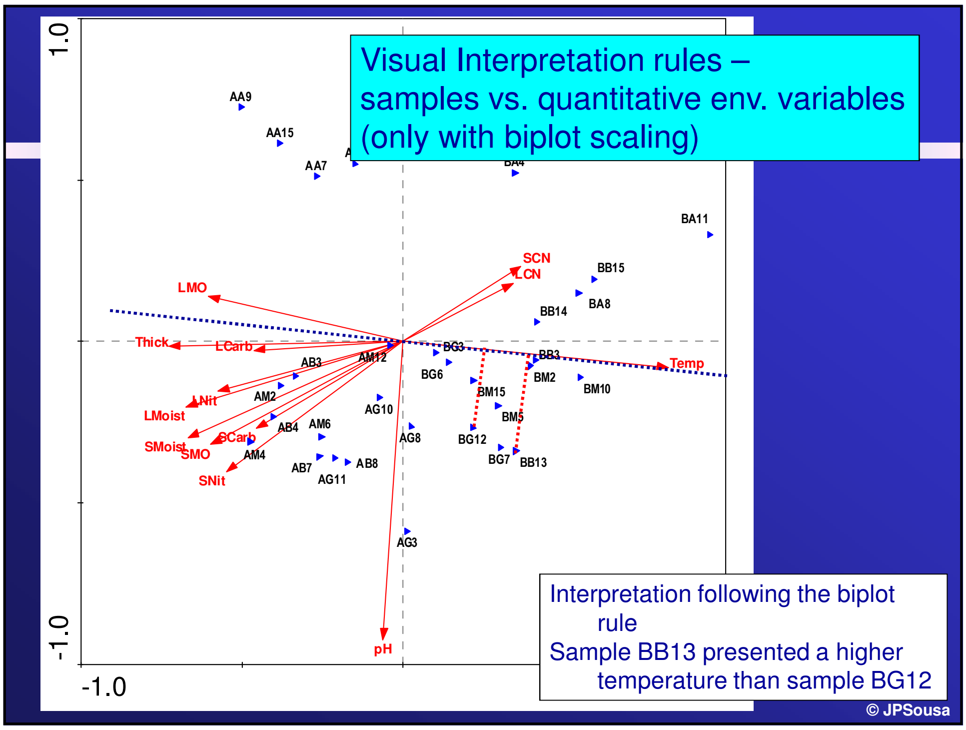
```
**** Summary of Monte Carlo test ****

Test of significance of first canonical axis: eigenvalue =    0.476
                                              F-ratio    =    5.257
                                              P-value    =    0.0020

Test of significance of all canonical axes  : Trace      =    1.290
                                              F-ratio    =    2.192
                                              P-value    =    0.0020

(  499 permutations under reduced model)

[Thu Jul 28 18:14:00 2005] CANOCO call succeeded
```

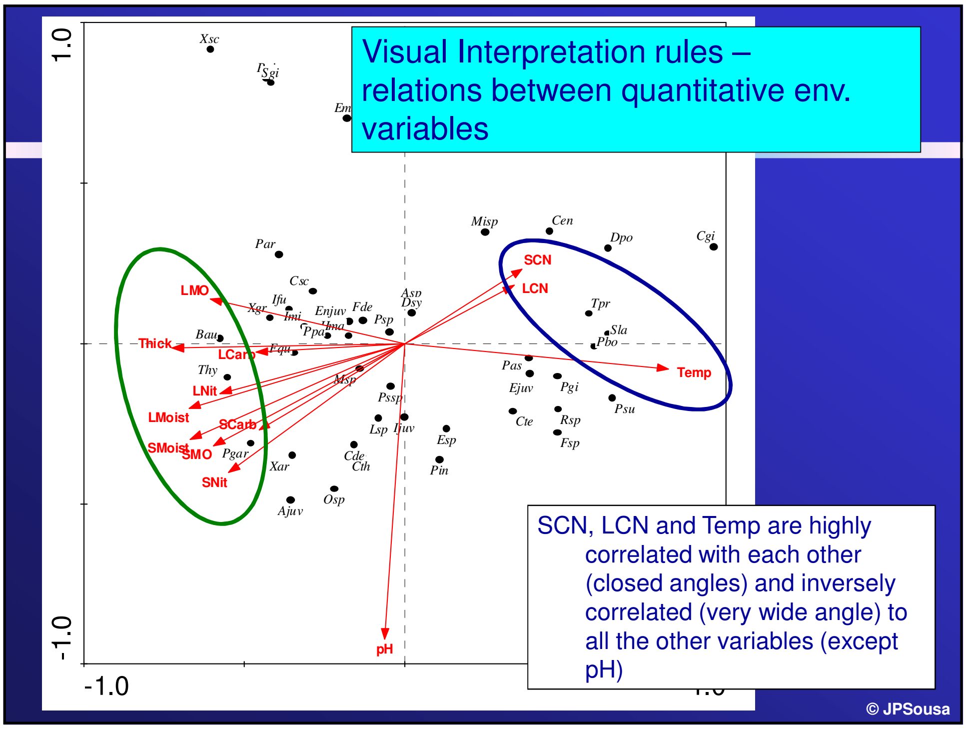
1. Monte-Carlo tests revealed significant relationships between species and environmental variables

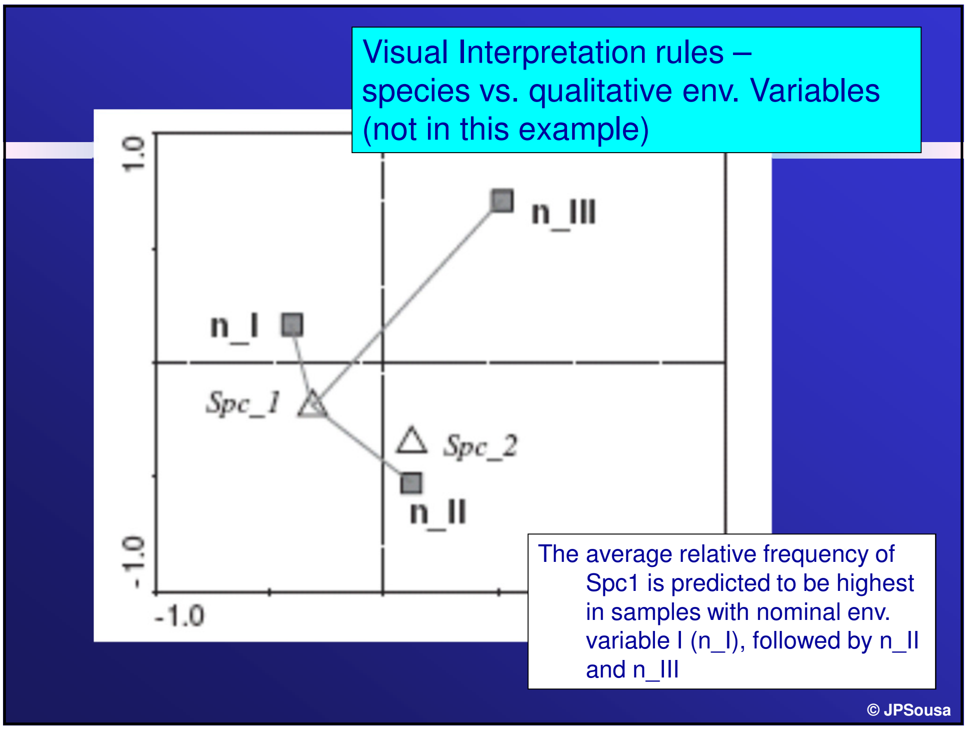


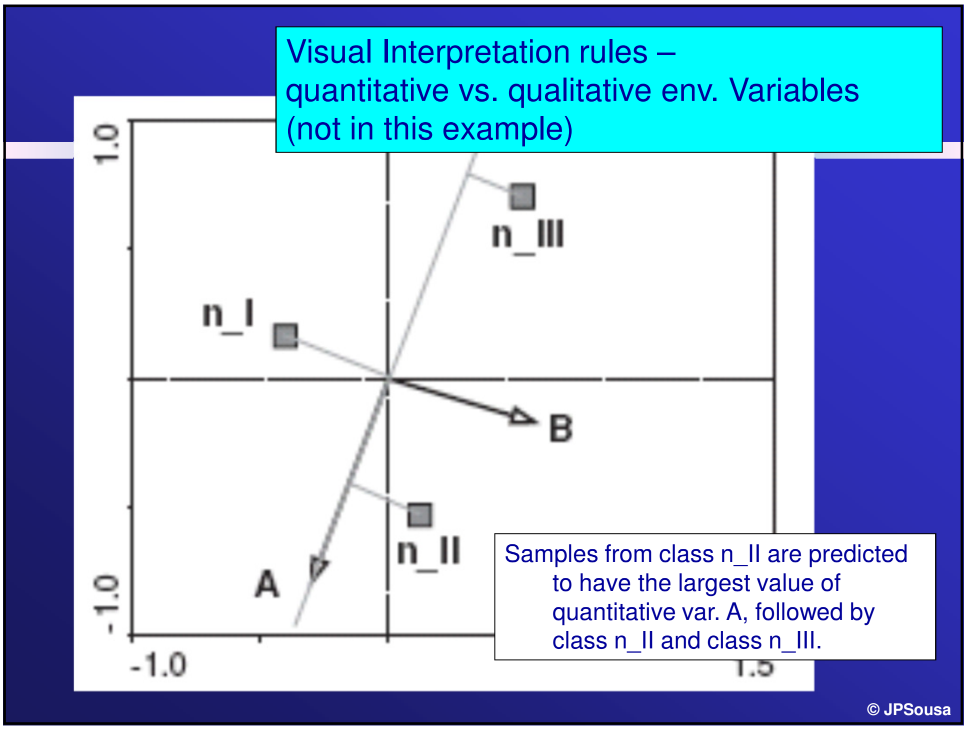



Visual Interpretation rules –
species vs. quantitative env. variables
Interpretation following the biplot rule
Fsp species with its optimum at higher temperatures
SCN
LCN
LMO
Thick
LCarb
LNit
LMoist
SCarb
SMoist
SMO
SNit
Temp
pH
1.0
-1.0
-1.0
© JPSousa


Visual Interpretation rules –
samples vs. quantitative env. variables
(only with biplot scaling)
Interpretation following the biplot rule
Sample BB13 presented a higher temperature than sample BG12
SCN
LCN
LMO
Thick
LCarb
LNit
LMoist
SCarb
SMoist
SMO
SNit
Temp
pH
1.0
-1.0
-1.0
© JPSousa


Visual Interpretation rules – relations between quantitative env. variables
1.0
-1.0
-1.0
SCN
LCN
Temp
LMO
Thick
LCarb
LNit
LMoist
SCarb
SMoist
SMO
SNit
pH
SCN, LCN and Temp are highly correlated with each other (closed angles) and inversely correlated (very wide angle) to all the other variables (except pH)
© JPSousa


Visual Interpretation rules – species vs. qualitative env. Variables (not in this example)
1.0
-1.0
-1.0
n_III
n_I
Spc_1
Spc_2
n_II
The average relative frequency of Spc1 is predicted to be highest in samples with nominal env. variable I (n_I), followed by n_II and n_III
© JPSousa




## Interpretation aids - CCA

| | SPEC AX1 | SPEC AX2 | SPEC AX3 | SPEC AX4 | ENVI AX1 | ENVI AX2 |
|---|---|---|---|---|---|---|
| SPEC AX1 | 1.0000 | | | | | |
| SPEC AX2 | -0.0235 | 1.0000 | | | | |
| SPEC AX3 | 0.0492 | -0.0741 | 1.0000 | | | |
| SPEC AX4 | 0.0596 | -0.0293 | 0.0239 | 1.0000 | | |
| ENVI AX1 | 0.9512 | 0.0000 | 0.0000 | 0.0000 | 1.0000 | |
| ENVI AX2 | 0.0000 | 0.9333 | 0.0000 | 0.0000 | 0.0000 | 1.0000 |
| ENVI AX3 | 0.0000 | 0.0000 | 0.8831 | 0.0000 | 0.0000 | 0.0000 |
| ENVI AX4 | 0.0000 | 0.0000 | 0.0000 | 0.9149 | 0.0000 | 0.0000 |
| Thick | -0.6904 | -0.0140 | -0.2021 | -0.1378 | -0.7259 | -0.0150 |
| LMoist | -0.6398 | -0.1884 | 0.5115 | 0.1612 | -0.6726 | -0.2019 |
| LCarb | -0.4404 | -0.0255 | 0.0894 | 0.5917 | -0.4630 | -0.0273 |
| LNit | -0.5476 | -0.1450 | 0.1201 | 0.2141 | -0.5757 | -0.1554 |
| LCN | 0.3244 | 0.1676 | -0.0211 | 0.5584 | 0.3410 | 0.1796 |
| LMO | -0.5755 | 0.1289 | 0.0501 | 0.3173 | -0.6050 | 0.1381 |
| Temp | 0.7810 | -0.0748 | 0.2749 | 0.0367 | 0.8211 | -0.0801 |
| pH | -0.0597 | -0.8617 | -0.0389 | -0.1225 | -0.0628 | -0.9233 |
| SMoist | -0.6361 | -0.2791 | 0.2624 | 0.0870 | -0.6687 | -0.2991 |
| SCarb | -0.4335 | -0.2517 | -0.3031 | 0.1861 | -0.4558 | -0.2696 |
| SNit | -0.5221 | -0.3766 | -0.1891 | 0.1526 | -0.5489 | -0.4035 |
| SCN | 0.3476 | 0.2165 | -0.0488 | -0.0592 | 0.3655 | 0.2320 |
| SMO | -0.5686 | -0.2985 | -0.1973 | 0.1896 | -0.5978 | -0.3199 |

**"Intraset correlation coefficients" (LOG file):**
**Correlations between environmental variables and the samples scores derived from the environmental variables (SamE)**

# Interpretation aids - CCA

```
WCanoImp produced data file
 CCA    Canonical axes:          4        Covariables:     0      Scaling:        -2
                 Downweight
Log-transformation
Regr: Regression/canonical coefficients for standardized variables
```

| N | NAME | AX1 | AX2 | AX3 | AX4 |
|---|---|---|---|---|---|
| | EIG | 0.4757 | 0.1960 | 0.1429 | 0.0965 |
| 1 | Thick | -0.3416 | 0.0276 | -0.1159 | -0.4806 |
| 2 | LMoist | -0.3719 | -0.6898 | 1.3120 | -0.1122 |
| 3 | LCarb | -0.4146 | 0.3118 | -0.5769 | -0.1800 |
| 4 | LNit | 0.2820 | 0.2243 | -0.0898 | 1.8059 |
| 5 | LCN | 0.4699 | -0.2263 | 0.0883 | 1.3578 |
| 6 | LMO | 0.2363 | -0.3006 | 0.6103 | -0.9634 |
| 7 | Temp | 0.4993 | -0.2010 | 0.5609 | -0.1700 |
| 8 | pH | 0.2919 | -1.1881 | -0.0853 | -0.1194 |
| 9 | SMoist | -0.2500 | 0.8233 | -0.1227 | -0.0853 |
| 10 | SCarb | -1.2368 | -0.6302 | -1.2741 | 2.1764 |
| 11 | SNit | 1.4096 | 0.0068 | 1.2554 | -3.2106 |
| 12 | SCN | 1.2220 | 0.6502 | 0.6278 | -2.0815 |
| 13 | SMO | -0.0678 | 0.2597 | -0.2504 | 0.7503 |

**"Canonical coefficients" (SOL file):**
**Coefficients derived from multiple regression of the species-derived sample scores (Samp) on the standardized environmental variables.**
**Unstable when environmental variables are correlated to each other**



# Interpretation aids - CCA

1. VIF – variance inflation factor (indicator of colinearity)
2. VIF > 20 colinear variables, threfore **redundant**

| | | | and. dev. | inflation factor |
|---|---|---|---|---|
| | | | 1.4519 | |
| 2 | SPEC AX2 | 0.0000 | 1.1950 | |
| 3 | SPEC AX3 | 0.0000 | 1.2232 | |
| 4 | SPEC AX4 | 0.0000 | 1.1499 | |
| 5 | ENVI AX1 | 0.0000 | 1.3810 | |
| 6 | ENVI AX2 | 0.0000 | 1.1153 | |
| 7 | ENVI AX3 | 0.0000 | 1.0802 | |
| 8 | ENVI AX4 | 0.0000 | 1.0521 | |
| 1 | Thick | 2.7135 | 0.9425 | 1.9826 |
| 2 | LMoist | 49.6527 | 34.4516 | 7.0048 |
| 3 | LCarb | 31.1398 | 8.4664 | 33.3312 |
| 4 | LNit | 1.5065 | 0.4436 | 32.5846 |
| 5 | LCN | 21.2830 | 4.5121 | 16.7075 |
| 6 | LMO | 53.5694 | 11.6797 | 10.2575 |
| 7 | Temp | 16.5237 | 1.8305 | 2.5362 |
| 8 | pH | 5.5979 | 0.9529 | 2.0585 |
| 9 | SMoist | 43.6898 | 30.5917 | 7.1993 |
| 10 | SCarb | 15.8552 | 4.4288 | 38.9587 |
| 11 | SNit | 0.6977 | 0.2087 | 69.0495 |
| 12 | SCN | 23.5541 | 4.8449 | 28.2549 |
| 13 | SMO | 24.9756 | 7.0962 | 10.4947 |

## Interpretation aids - CCA

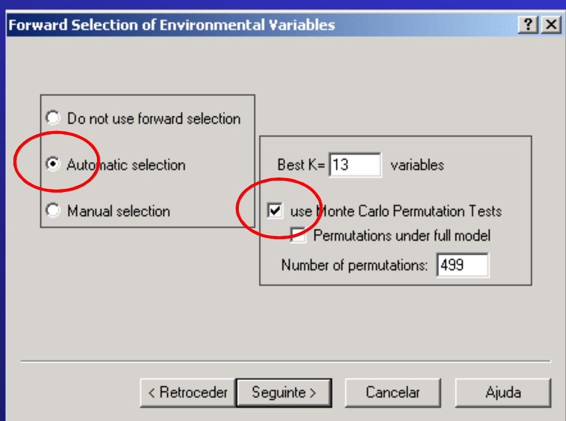
**Auxiliary tables:**
Both the Canonical coefficients (SOL file) and the Intraset correlation coefficients (LOG file) are used in the interpretation of the community structure based on the environmental variables (they measure the contribution of each environmental variable).

**Be careful with MULTICOLINEARITY !**
In case of Correlated environmental variables DO NOT USE THE CANONICAL COEFFICIENTS !



## CCA with Forward selection

1. All steps similar except "Forward selection"



# CCA with Forward selection

Resultados da Forward selection

| Marginal Effects | | |
|---|---|---|
| Variable | Var.N | Lambda1 |
| Temp | 7 | 0.34 |
| LMoist | 2 | 0.28 |
| Thick | 1 | 0.28 |
| SMoist | 9 | 0.26 |
| LMO | 6 | 0.21 |
| SMO | 13 | 0.21 |
| SNit | 11 | 0.21 |
| LNit | 4 | 0.19 |
| pH | 8 | 0.18 |
| LCarb | 3 | 0.16 |
| SCarb | 10 | 0.15 |
| LCN | 5 | 0.12 |
| SCN | 12 | 0.12 |

| Conditional Effects | | | | |
|---|---|---|---|---|
| Variable | Var.N | LambdaA | P | F |
| **Temp** | **7** | **0.34** | **0.002** | **5.86** |
| **LMoist** | **2** | **0.19** | **0.002** | **3.39** |
| **pH** | **8** | **0.17** | **0.002** | **3.51** |
| **Thick** | **1** | **0.09** | **0.016** | **1.77** |
| **LCarb** | **3** | **0.08** | **0.040** | **1.60** |
| **SCN** | **12** | **0.07** | **0.026** | **1.53** |
| SMoist | 9 | 0.06 | 0.168 | 1.26 |
| SCarb | 10 | 0.05 | 0.266 | 1.19 |
| LCN | 5 | 0.05 | 0.236 | 1.18 |
| LNit | 4 | 0.06 | 0.196 | 1.22 |
| SNit | 11 | 0.05 | 0.440 | 1.03 |
| LMO | 6 | 0.04 | 0.386 | 1.05 |
| SMO | 13 | 0.04 | 0.664 | 0.83 |

**Marginal effects**
Variables are ordered according to the variation they explain

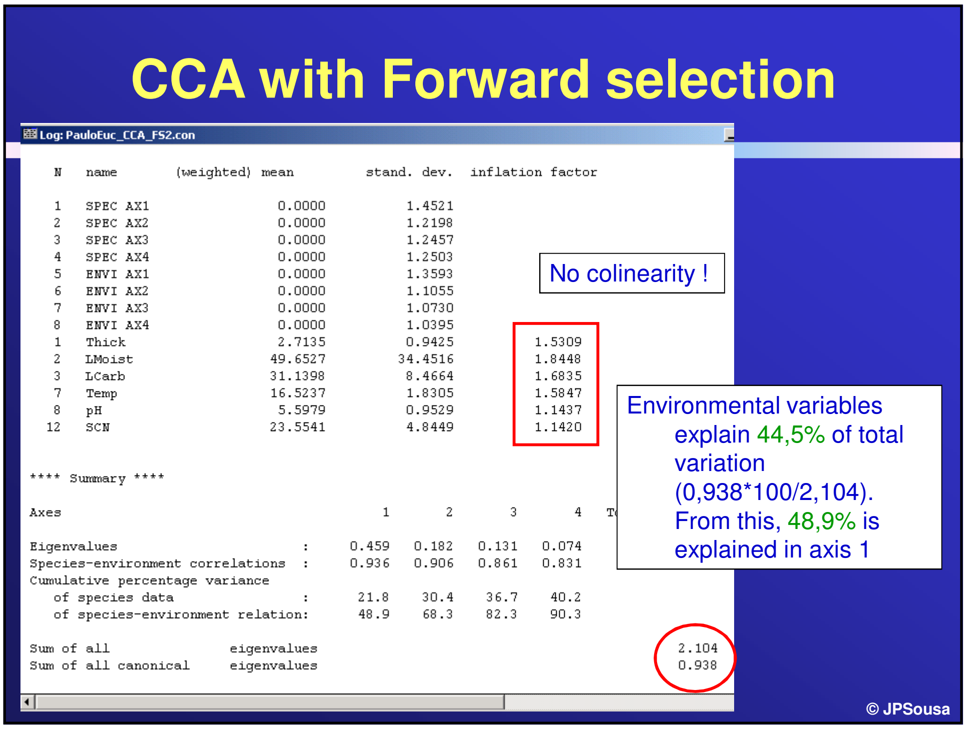
**Conditional effects**
Variables are ordered according to their entrance in the model

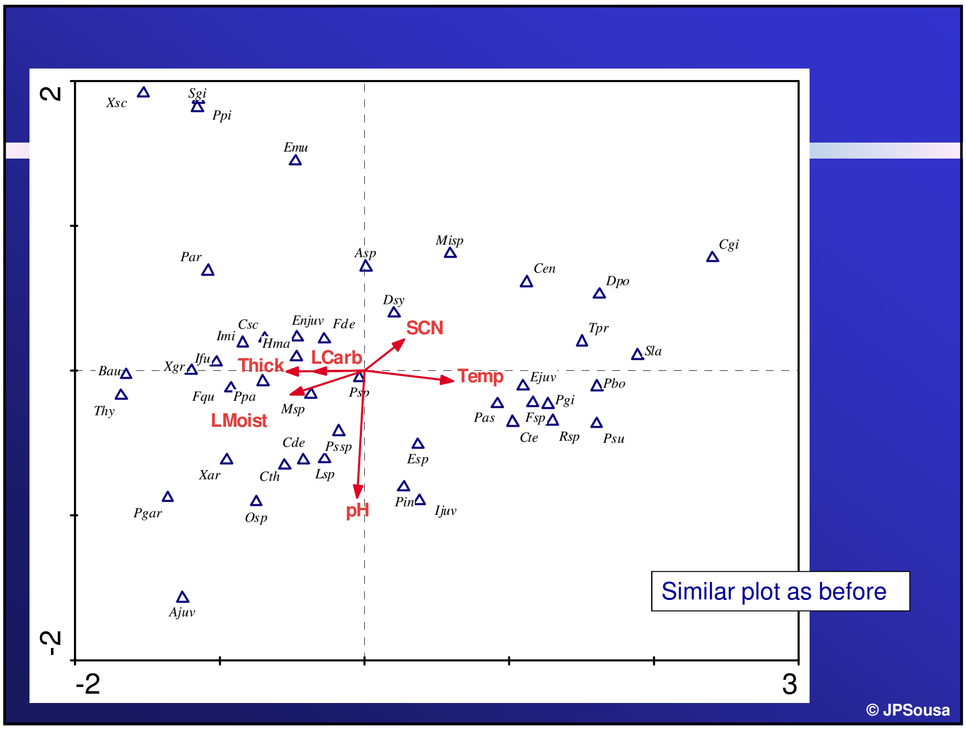
Differences between effects are due to correlations between variables. In case of uncorrelated variables, the result would be the same



# CCA with Forward selection

After sellection of variables a new "normal" CCA is done only with the selected variables


CCA with Forward selection
Log: PauloEuc_CCA_FS2.con
N name (weighted) mean stand. dev. inflation factor
1 SPEC AX1 0.0000 1.4521
2 SPEC AX2 0.0000 1.2198
3 SPEC AX3 0.0000 1.2457
4 SPEC AX4 0.0000 1.2503
5 ENVI AX1 0.0000 1.3593
6 ENVI AX2 0.0000 1.1055
7 ENVI AX3 0.0000 1.0730
8 ENVI AX4 0.0000 1.0395
1 Thick 2.7135 0.9425 1.5309
2 LMoist 49.6527 34.4516 1.8448
3 LCarb 31.1398 8.4664 1.6835
7 Temp 16.5237 1.8305 1.5847
8 pH 5.5979 0.9529 1.1437
12 SCN 23.5541 4.8449 1.1420
No colinearity !
**** Summary ****
Axes 1 2 3 4
Eigenvalues : 0.459 0.182 0.131 0.074
Species-environment correlations : 0.936 0.906 0.861 0.831
Cumulative percentage variance
of species data : 21.8 30.4 36.7 40.2
of species-environment relation: 48.9 68.3 82.3 90.3
Sum of all eigenvalues 2.104
Sum of all canonical eigenvalues 0.938
Environmental variables explain 44,5% of total variation (0,938*100/2,104). From this, 48,9% is explained in axis 1
© JPSousa


2
-2
-2
3
SCN
Temp
Thick
LCarb
LMoist
pH
Xsc
Sgi
Ppi
Emu
Par
Asp
Misp
Cgi
Cen
Dpo
Dsy
Tpr
Sla
Csc
Imi
Hma
Enjuv
Fde
Ifu
Xgr
Bau
Thy
Fqu
Ppa
Msp
Psp
Ejuv
Pbo
Pgi
Pas
Fsp
Cte
Rsp
Psu
Cde
Pssp
Esp
Cth
Lsp
Xar
Pin
Ijuv
Pgar
Osp
Ajuv
Similar plot as before
© JPSousa

# CANOCO for Windows

Hands on !
Part 5



# PRESENCE - ABSENCE DATA

*How to relate "species" data to environmental variables ?*

- The PCoA (Principal Coordinate Analysis) or the NMDS (Non-metric Multidimentional Scaling) methods can be two possible solutions to relate presence-absence data to environmental variables
- The principle is to use the similarity between samples to create a representation in an ordination space and than, to perform an indirect (in the case of an NMDS) or direct (in the case of the PCoA) analysis in CANOCO using these coordinates as "species data".

# PRESENCE - ABSENCE DATA

*How to relate "species" data to environmental variables ?*

1. Calculate the oordination coordinates in PrCoord (for a PCoA) or in WinKyst (for a NMDS).
2. In the case of presence-absence data use qualitative similarity coefficients (Jaccard or Sorensen)
3. In the case of quantitative data use quantitative similarity coefficients (e.g. Bray Curtis, Hellinger, etc).
4. Use the result of that analysis as the "species data" in an indirect PCA (in case of a NMDS) or a dbRDA (case of a PCoA) + explanatory variables as "environmental variables" + species (original matrix) as "supplementary environmental variables"

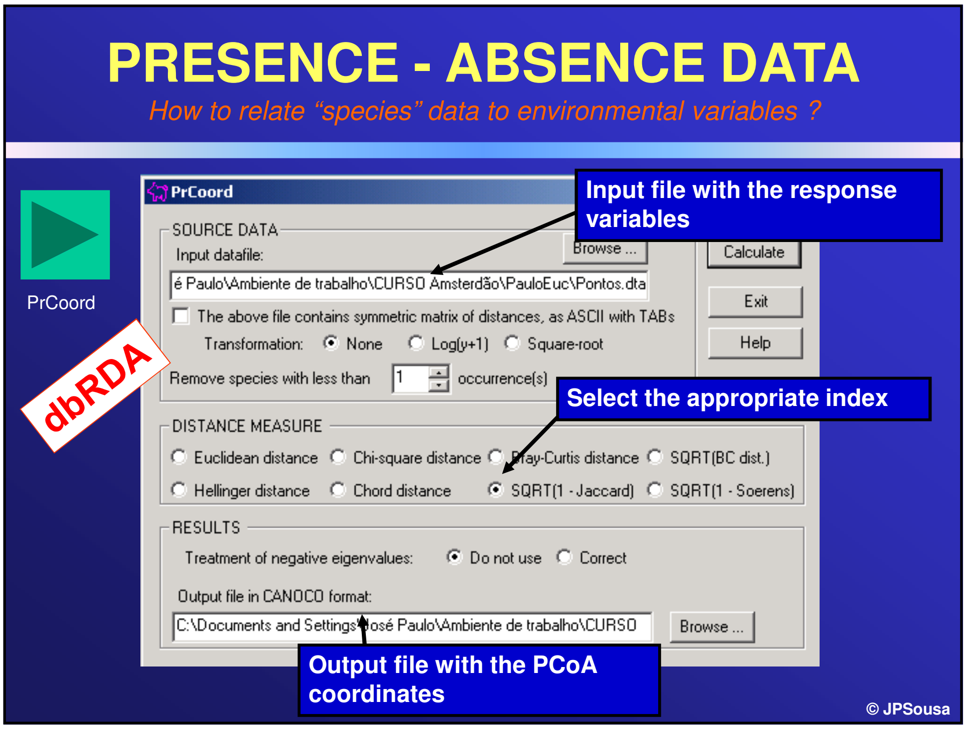
**For the use of WinKyst and how to do a NMDS and how to do an indirect PCA – see example on Lecture 2**

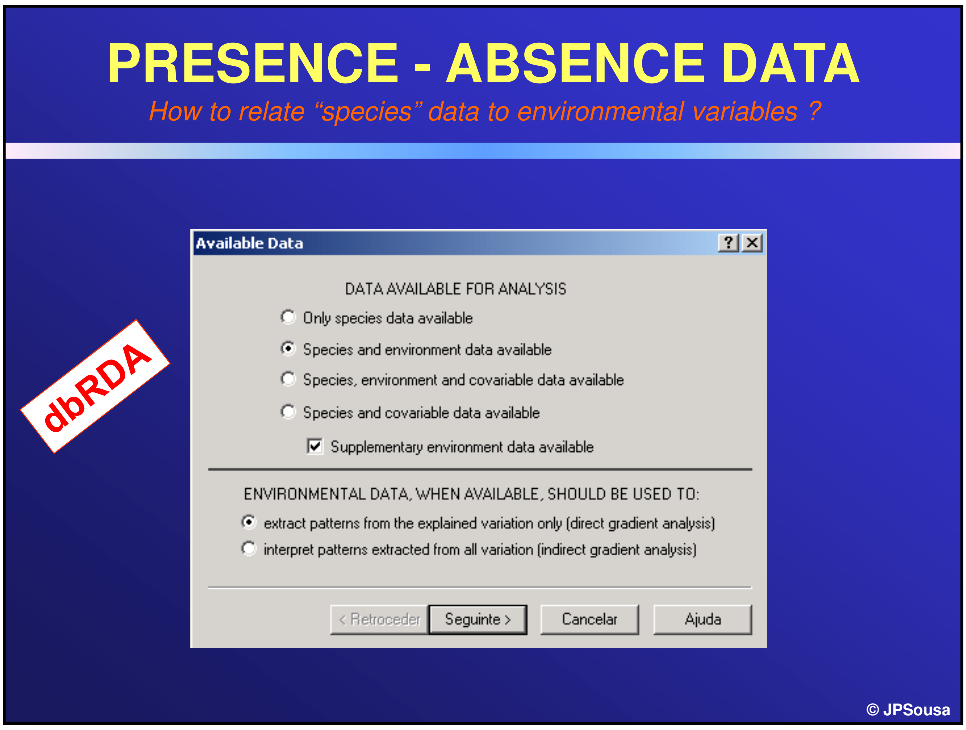


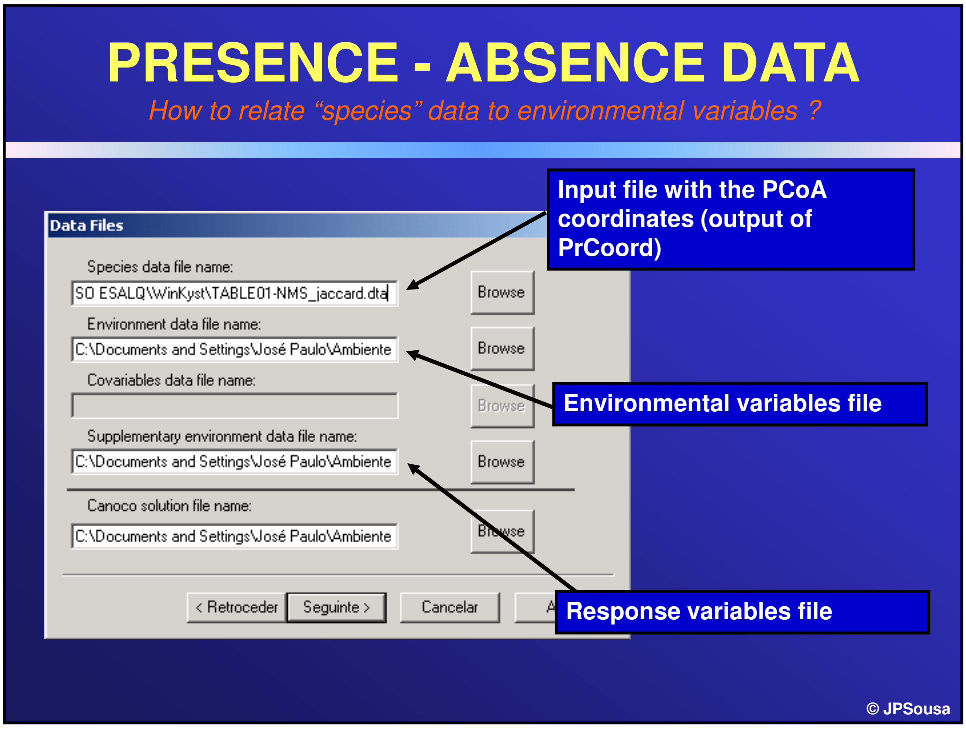
# CANOCO for Windows - dbRDA

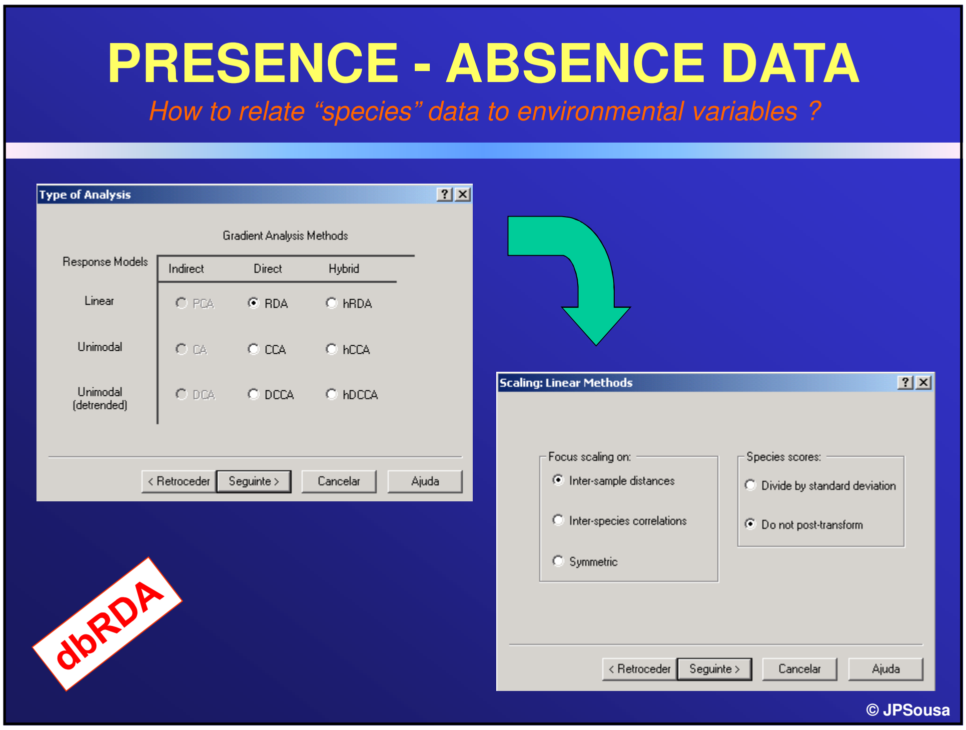
Soil fauna from Cork Oak (*Quercus suber*) and Eucalyptus (*Eucalyptus globulus*) stands (Sousa et al., 2003)

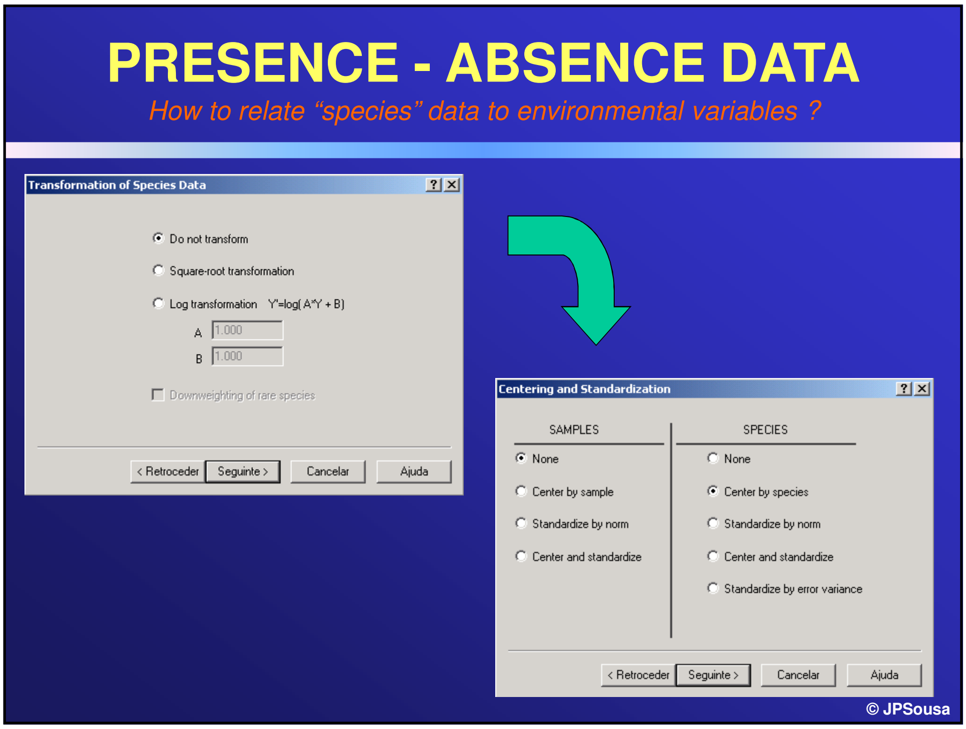
- Soil mesofauna data; soil pedological parameters (File Matrizes_CA_CCA.xls)
- 2 sites (Q e E) with four plots each (A, B, G, M) and each plot with 4 soil cores ;
- 32 samples in total with 45 collembola species identified;
- Objective: to evaluate the association between species and soil parameters (using a dbRDA)

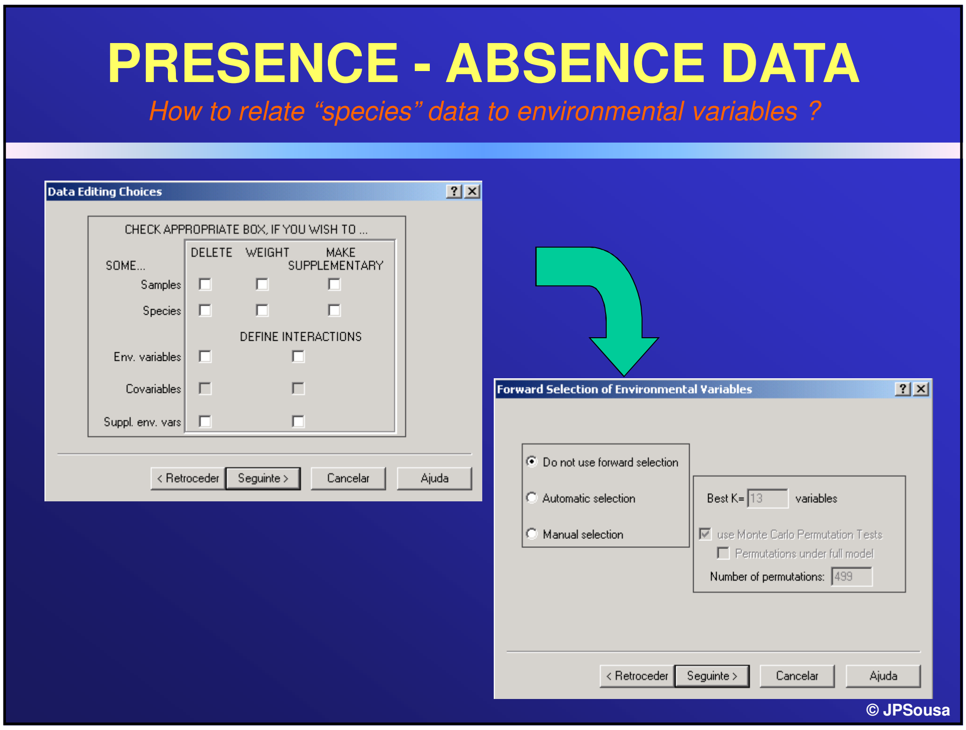

PRESENCE - ABSENCE DATA
How to relate "species" data to environmental variables ?
PrCoord
dbRDA
PrCoord
SOURCE DATA
Input datafile:
Browse ...
Calculate
é Paulo\Ambiente de trabalho\CURSO Amsterdão\PauloEuc\Pontos.dta
Exit
The above file contains symmetric matrix of distances, as ASCII with TABs
Help
Transformation: None Log(y+1) Square-root
Remove species with less than 1 occurrence(s)
DISTANCE MEASURE
Euclidean distance Chi-square distance Bray-Curtis distance SQRT(BC dist.)
Hellinger distance Chord distance SQRT(1 - Jaccard) SQRT(1 - Soerens)
RESULTS
Treatment of negative eigenvalues: Do not use Correct
Output file in CANOCO format:
C:\Documents and Settings\José Paulo\Ambiente de trabalho\CURSO
Browse ...
Input file with the response variables
Select the appropriate index
Output file with the PCoA coordinates
© JPSousa

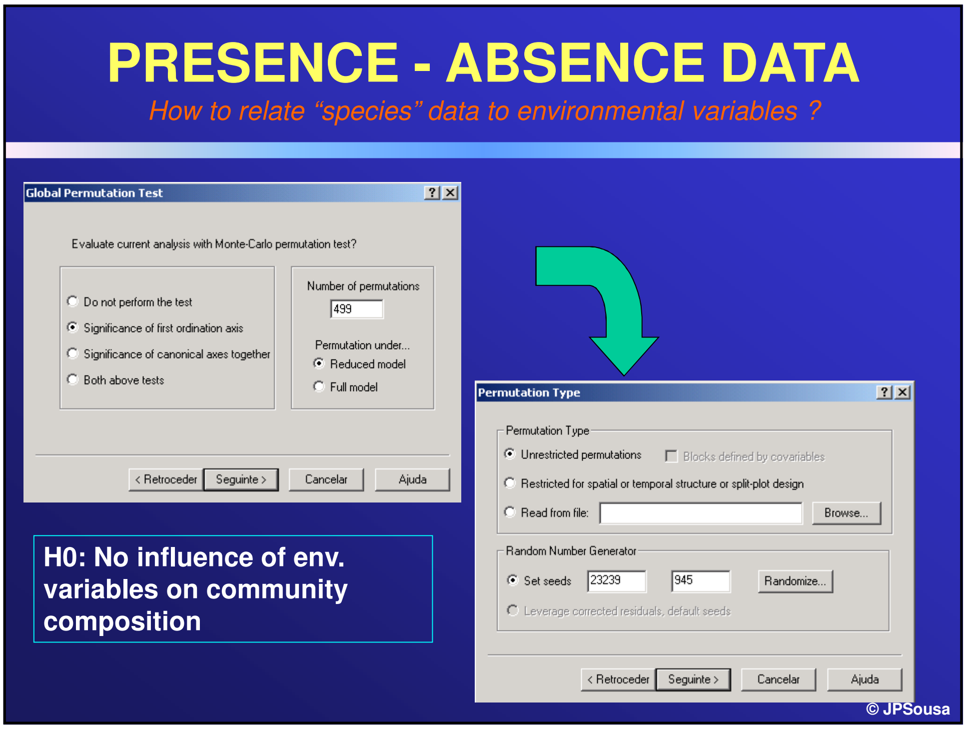

PRESENCE - ABSENCE DATA
How to relate "species" data to environmental variables ?
dbRDA
Available Data
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE USED TO:
extract patterns from the explained variation only (direct gradient analysis)
interpret patterns extracted from all variation (indirect gradient analysis)
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

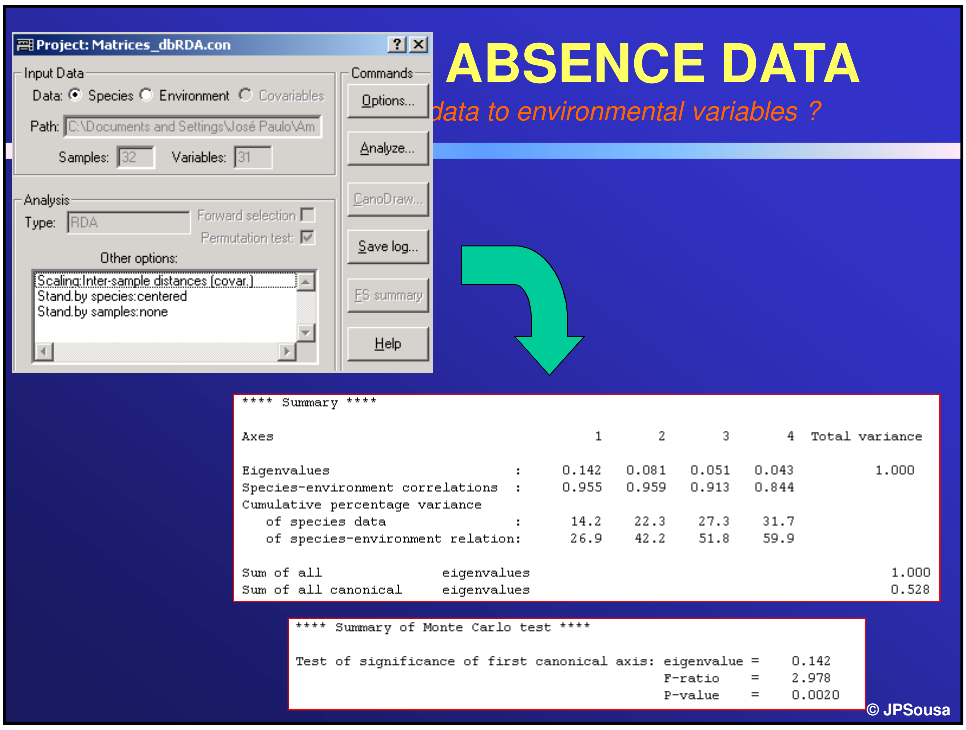

PRESENCE - ABSENCE DATA
How to relate “species” data to environmental variables ?
Data Files
Species data file name:
SO ESALQ\WinKyst\TABLE01-NMS_jaccard.dta
Environment data file name:
C:\Documents and Settings\José Paulo\Ambiente
Covariables data file name:
Supplementary environment data file name:
C:\Documents and Settings\José Paulo\Ambiente
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
< Retroceder
Seguinte >
Cancelar
Input file with the PCoA coordinates (output of PrCoord)
Environmental variables file
Response variables file
© JPSousa

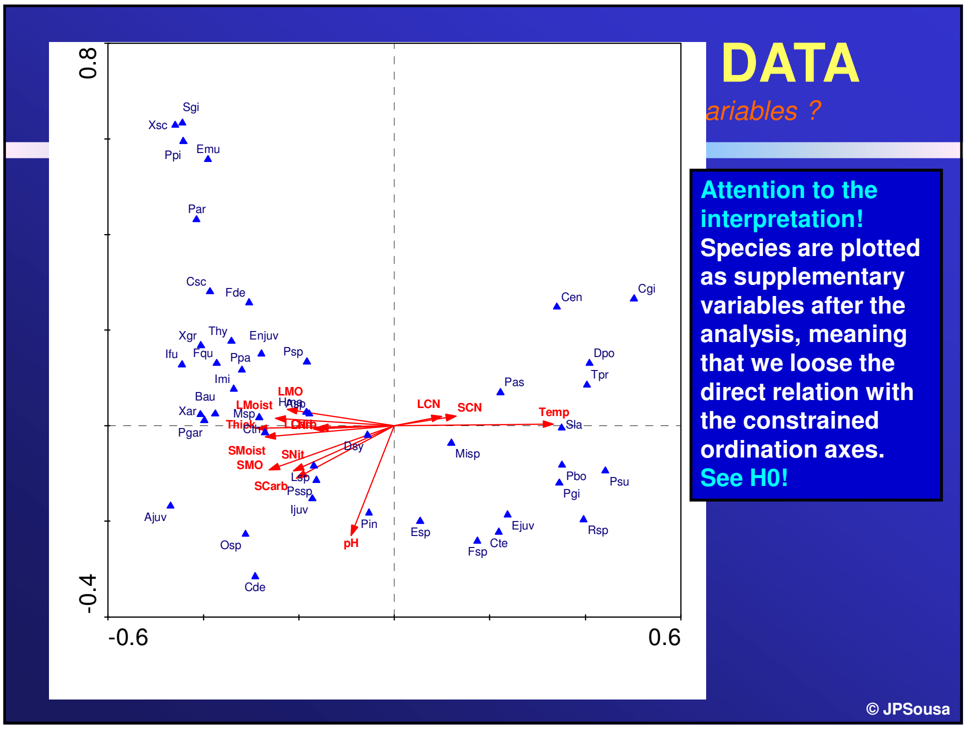

PRESENCE - ABSENCE DATA
How to relate “species” data to environmental variables ?
Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
< Retroceder
Seguinte >
Cancelar
Ajuda
Scaling: Linear Methods
Focus scaling on:
Inter-sample distances
Inter-species correlations
Symmetric
Species scores:
Divide by standard deviation
Do not post-transform
dbRDA
© JPSousa

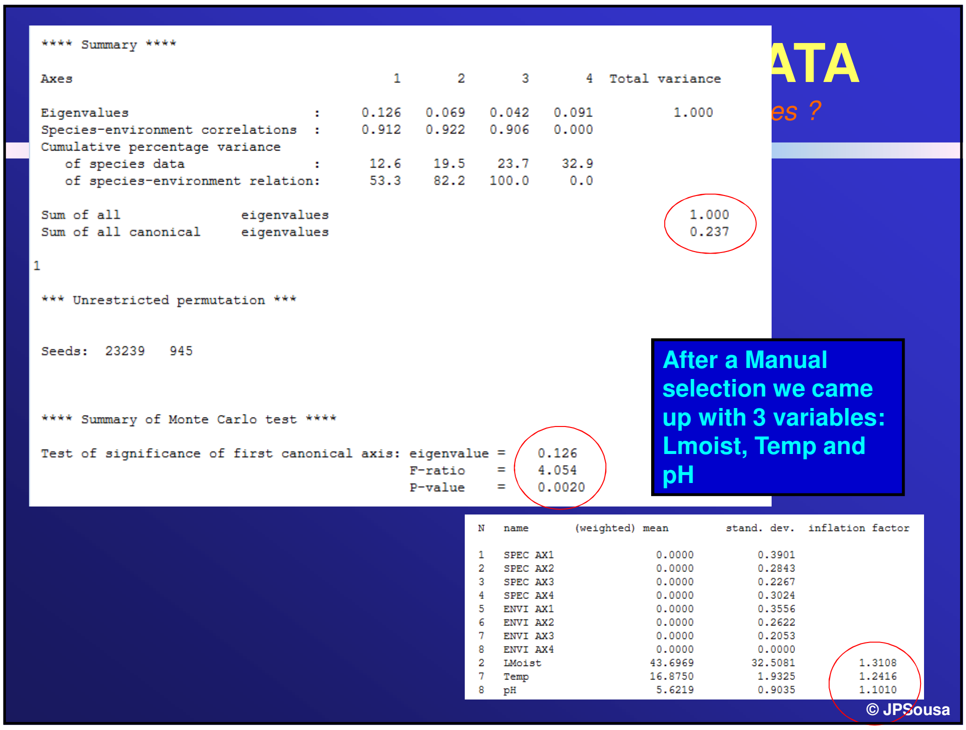

PRESENCE - ABSENCE DATA
How to relate "species" data to environmental variables ?
Transformation of Species Data
Do not transform
Square-root transformation
Log transformation Y'=log( A*Y + B)
A 1.000
B 1.000
Downweighting of rare species
< Retroceder
Seguinte >
Cancelar
Ajuda
Centering and Standardization
SAMPLES
SPECIES
None
Center by sample
Standardize by norm
Center and standardize
None
Center by species
Standardize by norm
Center and standardize
Standardize by error variance
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

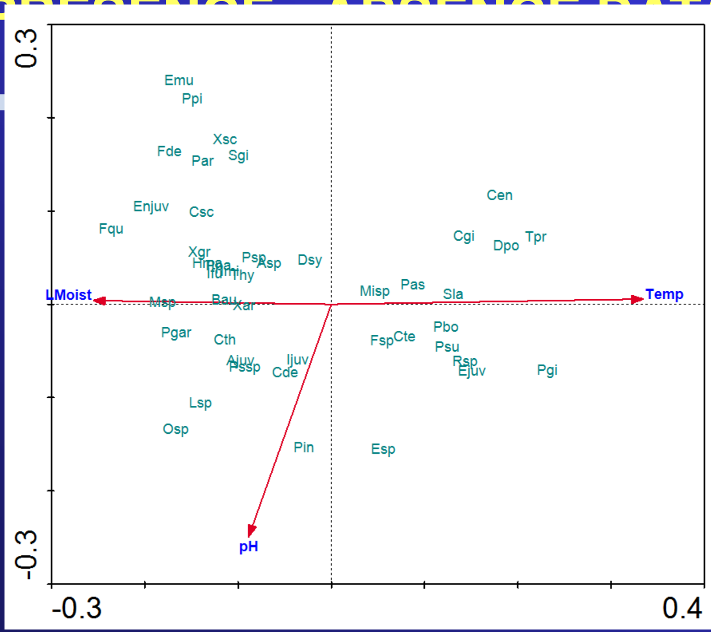

PRESENCE - ABSENCE DATA
How to relate "species" data to environmental variables ?
Data Editing Choices
CHECK APPROPRIATE BOX, IF YOU WISH TO ...
SOME...
DELETE
WEIGHT
MAKE SUPPLEMENTARY
Samples
Species
DEFINE INTERACTIONS
Env. variables
Covariables
Suppl. env. vars
< Retroceder
Seguinte >
Cancelar
Ajuda
Forward Selection of Environmental Variables
Do not use forward selection
Automatic selection
Manual selection
Best K= 13 variables
use Monte Carlo Permutation Tests
Permutations under full model
Number of permutations: 499
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa


PRESENCE - ABSENCE DATA
How to relate "species" data to environmental variables ?
Global Permutation Test
Evaluate current analysis with Monte-Carlo permutation test?
Do not perform the test
Significance of first ordination axis
Significance of canonical axes together
Both above tests
Number of permutations
499
Permutation under...
Reduced model
Full model
< Retroceder
Seguinte >
Cancelar
Ajuda
H0: No influence of env. variables on community composition
Permutation Type
Permutation Type
Unrestricted permutations
Blocks defined by covariables
Restricted for spatial or temporal structure or split-plot design
Read from file:
Browse...
Random Number Generator
Set seeds
23239
945
Randomize...
Leverage corrected residuals, default seeds
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa


Project: Matrices_dbRDA.con
Input Data
Data: Species Environment Covariables
Path: C:\Documents and Settings\José Paulo\Am
Samples: 32
Variables: 31
Analysis
Type: RDA
Forward selection
Permutation test:
Other options:
Scaling:Inter-sample distances (covar.)
Stand.by species:centered
Stand.by samples:none
Commands
Options...
Analyze...
CanoDraw...
Save log...
FS summary
Help
ABSENCE DATA
data to environmental variables ?
**** Summary ****
Axes 1 2 3 4 Total variance
Eigenvalues : 0.142 0.081 0.051 0.043 1.000
Species-environment correlations : 0.955 0.959 0.913 0.844
Cumulative percentage variance
of species data : 14.2 22.3 27.3 31.7
of species-environment relation: 26.9 42.2 51.8 59.9
Sum of all eigenvalues 1.000
Sum of all canonical eigenvalues 0.528
**** Summary of Monte Carlo test ****
Test of significance of first canonical axis: eigenvalue = 0.142
F-ratio = 2.978
P-value = 0.0020
© JPSousa


DATA
ariables ?
Attention to the interpretation! Species are plotted as supplementary variables after the analysis, meaning that we loose the direct relation with the constrained ordination axes. See H0!
0.8
-0.4
-0.6
0.6
Sgi
Xsc
Ppi
Emu
Par
Csc
Fde
Cen
Cgi
Xgr
Thy
Enjuv
Ifu
Fqu
Ppa
Psp
Dpo
Tpr
Imi
Pas
Bau
LMO
Xar
LMoist
Msp
LCN
SCN
Temp
Sla
Pgar
SMoist
SNit
Dsy
Misp
SMO
Pbo
Psu
Pgi
SCarb
Pssp
Ajuv
Ijuv
Pin
Esp
Ejuv
Rsp
Osp
pH
Cte
Fsp
Cde
© JPSousa


ATA
es ?
**** Summary ****
Axes 1 2 3 4 Total variance
Eigenvalues : 0.126 0.069 0.042 0.091 1.000
Species-environment correlations : 0.912 0.922 0.906 0.000
Cumulative percentage variance
of species data : 12.6 19.5 23.7 32.9
of species-environment relation: 53.3 82.2 100.0 0.0
Sum of all eigenvalues 1.000
Sum of all canonical eigenvalues 0.237
1
*** Unrestricted permutation ***
Seeds: 23239 945
**** Summary of Monte Carlo test ****
Test of significance of first canonical axis: eigenvalue = 0.126
F-ratio = 4.054
P-value = 0.0020
After a Manual selection we came up with 3 variables: Lmoist, Temp and pH
N name (weighted) mean stand. dev. inflation factor
1 SPEC AX1 0.0000 0.3901
2 SPEC AX2 0.0000 0.2843
3 SPEC AX3 0.0000 0.2267
4 SPEC AX4 0.0000 0.3024
5 ENVI AX1 0.0000 0.3556
6 ENVI AX2 0.0000 0.2622
7 ENVI AX3 0.0000 0.2053
8 ENVI AX4 0.0000 0.0000
2 LMoist 43.6969 32.5081 1.3108
7 Temp 16.8750 1.9325 1.2416
8 pH 5.6219 0.9035 1.1010
© JPSousa






Multivariate Statistical Tools in Ecology

ISCED, Lubango, March 2016

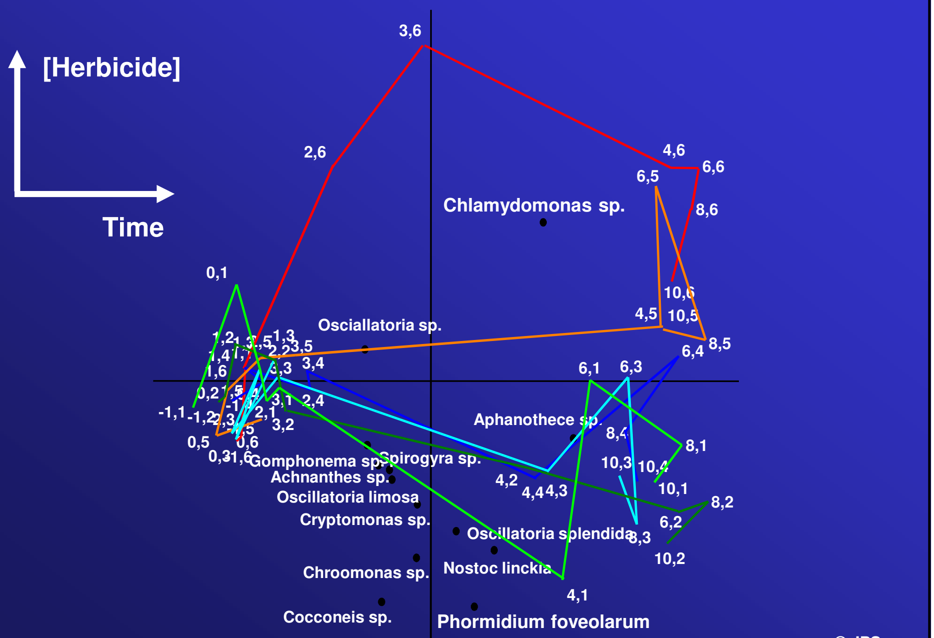
## Ordination Tools IV:
## Particular applications

# PRINCIPAL RESPONSE CURVES (PRC)

- Interesting method when having a REFERENCE tretament/site + several tretaments/sites & time sampling.
- We are interested NOT in the changes over time of each treatment BUT on the changes between treatments and control over time
- Time complicates the analysis!



## Ex: Phytoplankton vs. Herbicide - PCA



## Phytoplancton vs. Herbicide:
### PCA

- Results in the analysis (biplot) not easy to understand or communicate
- Results show treatment effects only in the two highest concentrations
- Results allow some interpretation back to species level



## Principal Response Curves

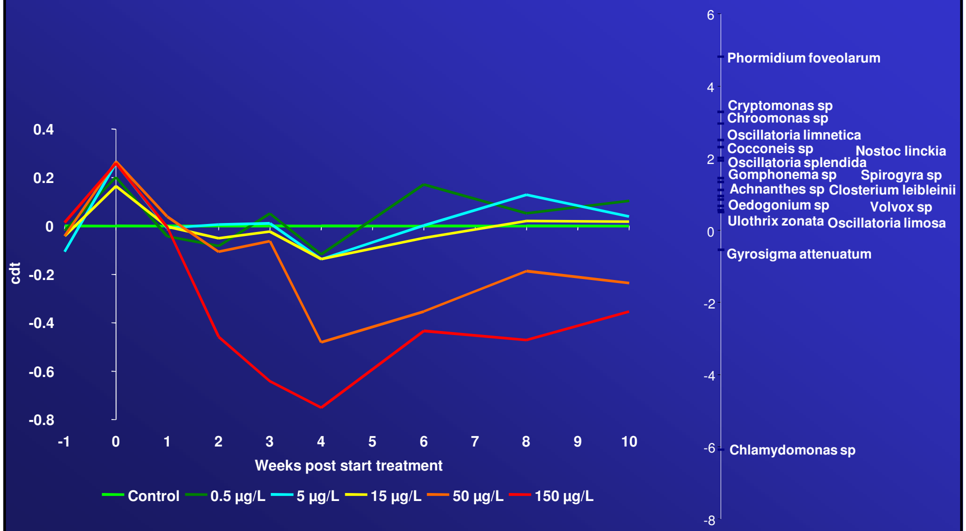
- Particular RDA technique
- PRC model:

yd(i)tk = y0tk + bk * cdt + ed(i)tk

Where:
yd(i)tk : log-abundance of species *k* in replicate *i* of treatment *d* at time *t*
y0tk : mean log-abundance of species *k* in the control (d=0) at time *t*
cdt : score of treatment *d* at time *t* (the response)
bk : weight of species *k*
ed(i)tk : error term

Van den Brink & Ter Braak (1999) *ET&C, 18:138-148*

# Phytoplankton vs. Herbicide:
## Principal Response Curves

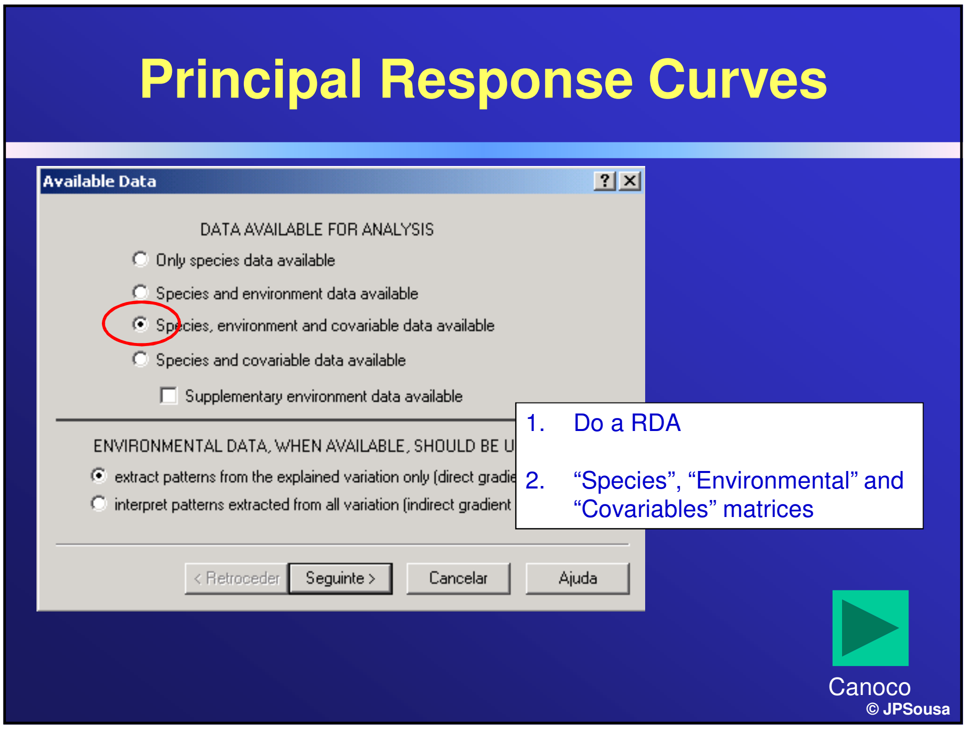


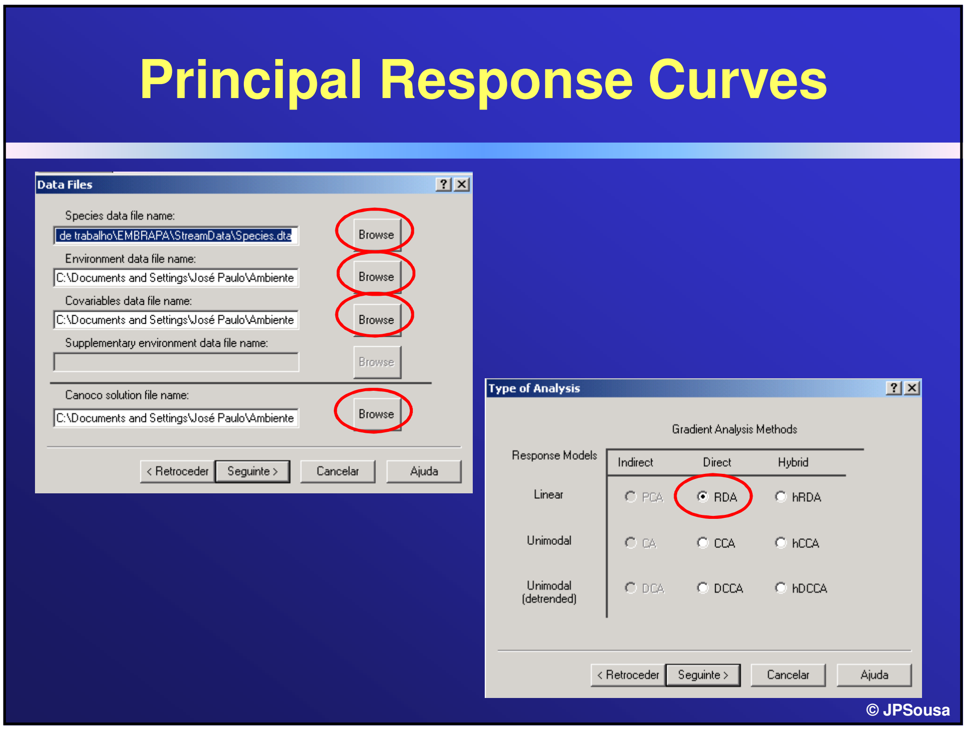


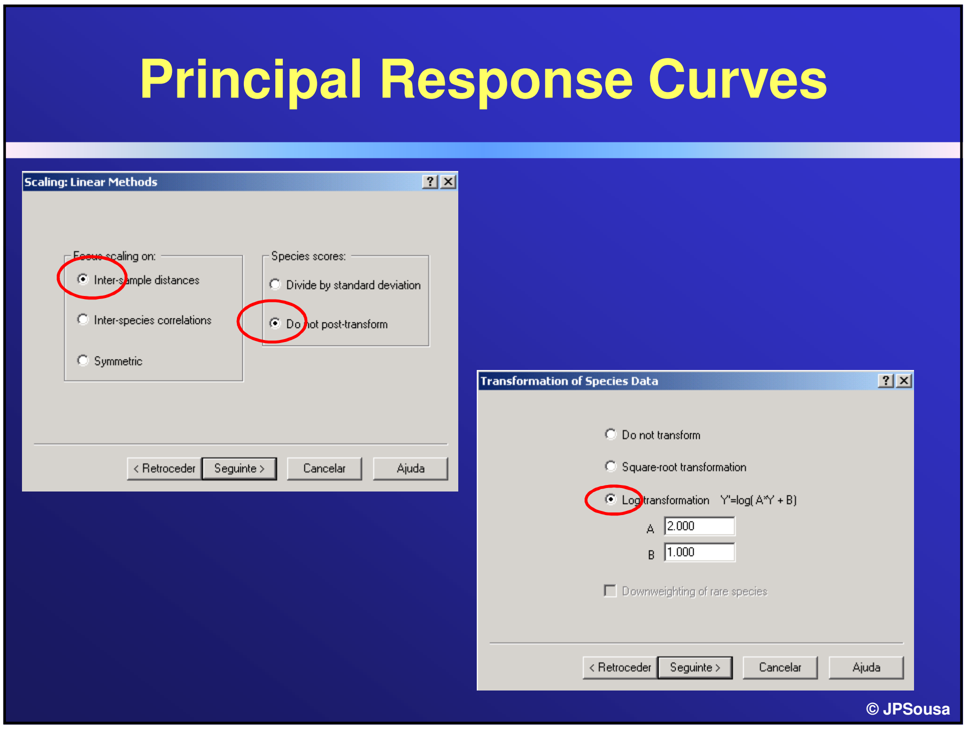

# Principal Response Curves

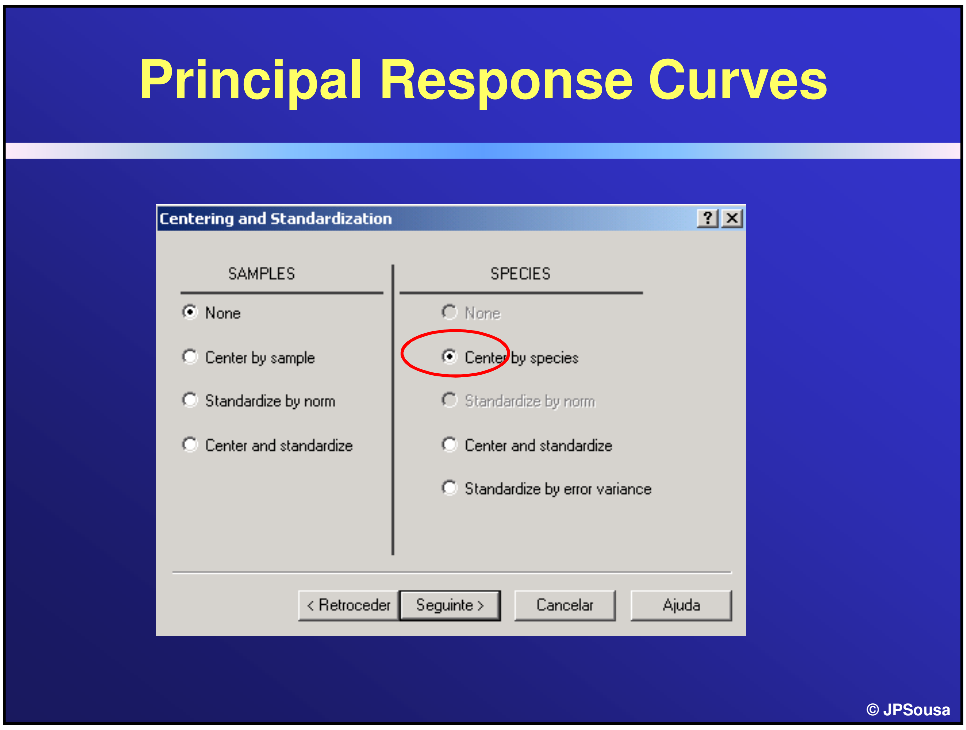
**Exemple:**

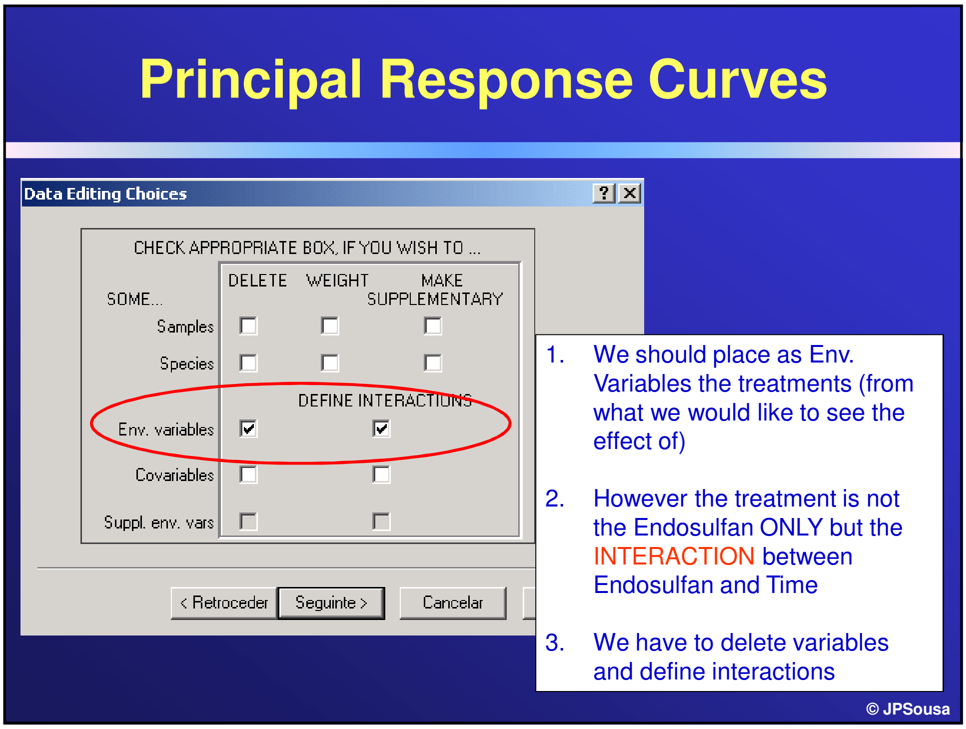
- Effect of contaminated run-off water from agricultural fields on freshwater invertebrates along time (File StreamData.xls)
- One contaminated stream (mainly with endosulfan) + one reference stream
- Sampling along four times

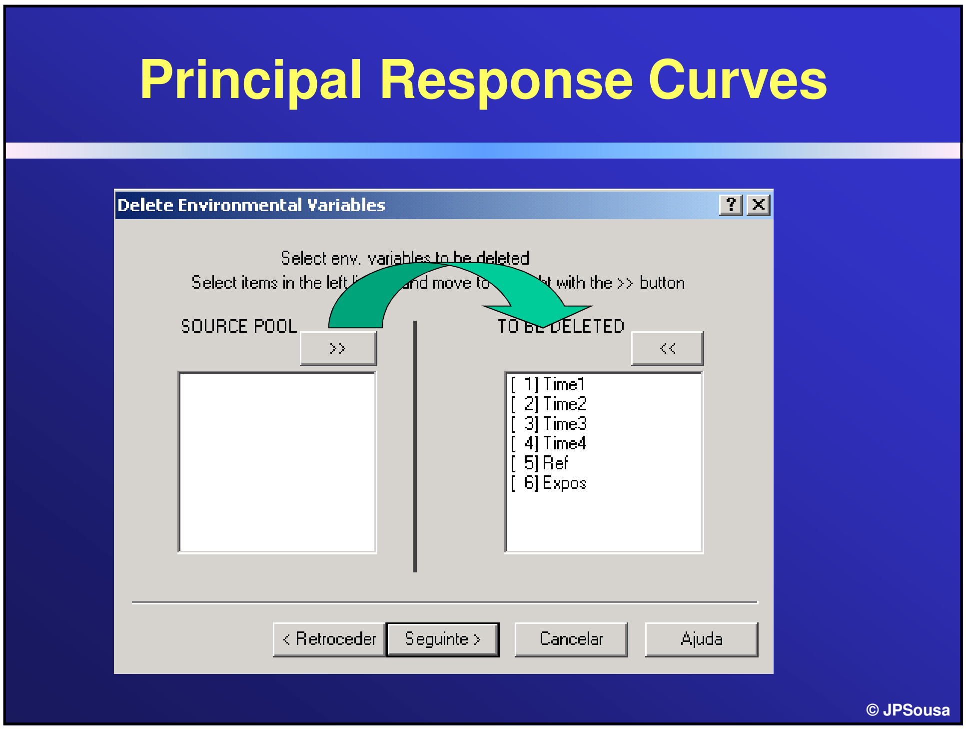

Principal Response Curves
Available Data
DATA AVAILABLE FOR ANALYSIS
Only species data available
Species and environment data available
Species, environment and covariable data available
Species and covariable data available
Supplementary environment data available
ENVIRONMENTAL DATA, WHEN AVAILABLE, SHOULD BE U
extract patterns from the explained variation only (direct gradie
interpret patterns extracted from all variation (indirect gradient
< Retroceder
Seguinte >
Cancelar
Ajuda
1. Do a RDA
2. "Species", "Environmental" and "Covariables" matrices
Canoco
© JPSousa

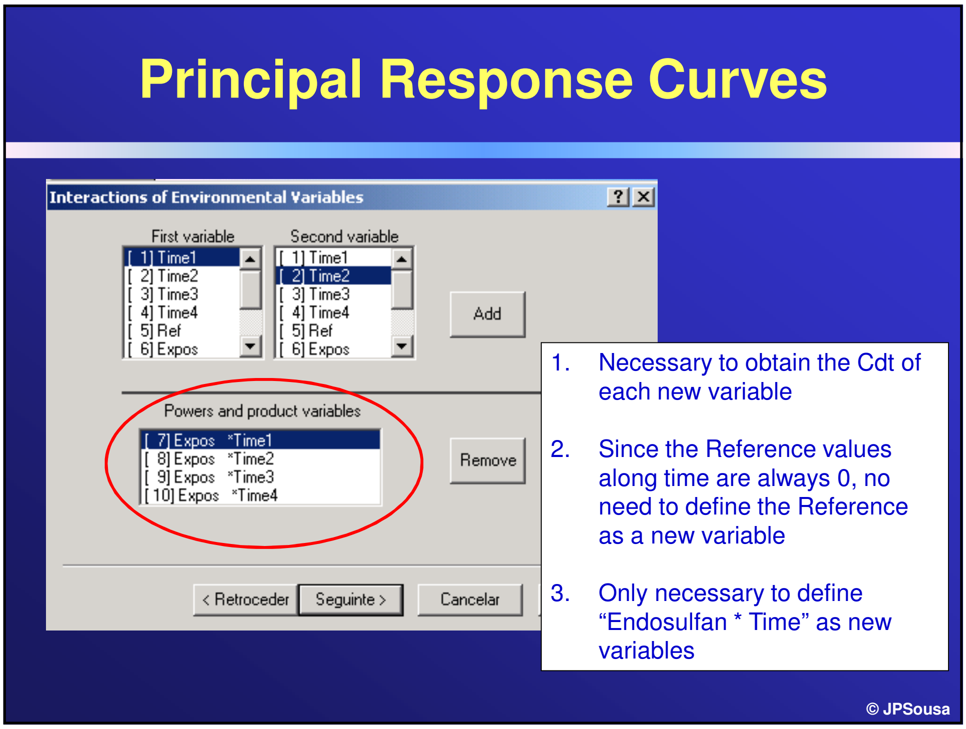

Principal Response Curves
Data Files
Species data file name:
de trabalho\EMBRAPA\StreamData\Species.dta
Browse
Environment data file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
Covariables data file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
Supplementary environment data file name:
Browse
Canoco solution file name:
C:\Documents and Settings\José Paulo\Ambiente
Browse
< Retroceder
Seguinte >
Cancelar
Ajuda
Type of Analysis
Gradient Analysis Methods
Response Models
Indirect
Direct
Hybrid
Linear
PCA
RDA
hRDA
Unimodal
CA
CCA
hCCA
Unimodal (detrended)
DCA
DCCA
hDCCA
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

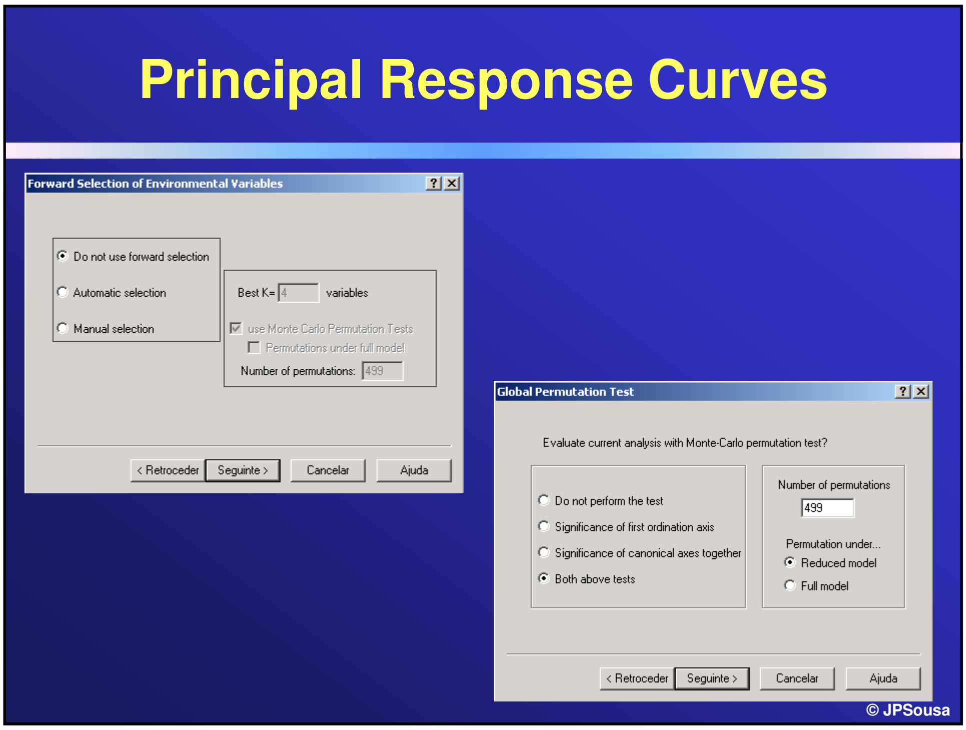

Principal Response Curves
Scaling: Linear Methods
Focus scaling on:
Inter-sample distances
Inter-species correlations
Symmetric
Species scores:
Divide by standard deviation
Do not post-transform
< Retroceder
Seguinte >
Cancelar
Ajuda
Transformation of Species Data
Do not transform
Square-root transformation
Log transformation Y'=log( A*Y + B)
A 2.000
B 1.000
Downweighting of rare species
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

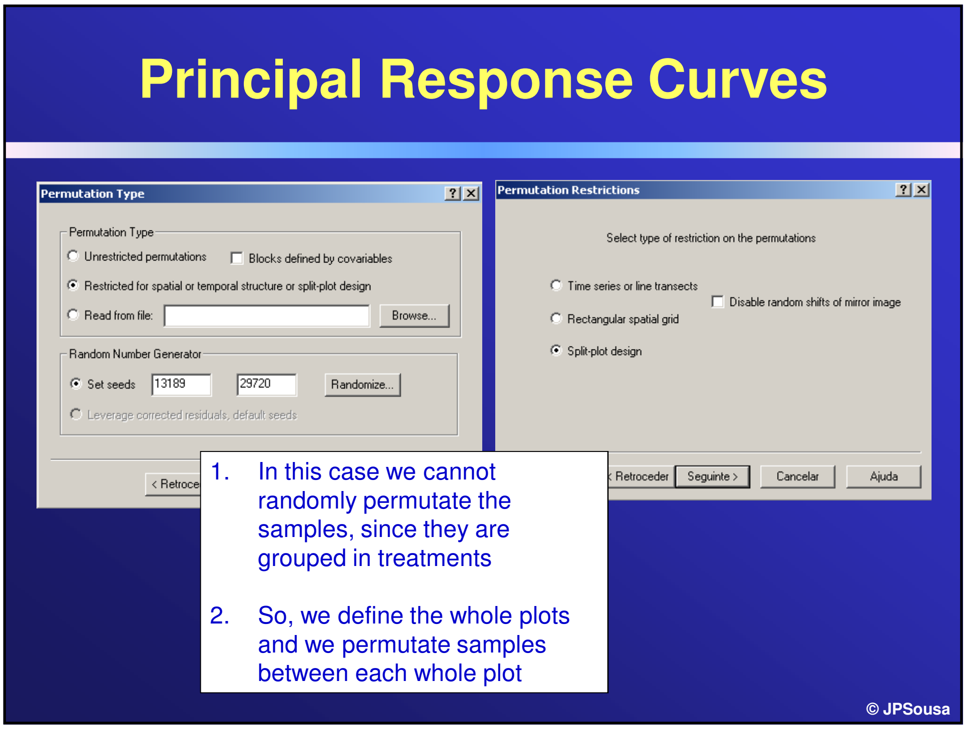

Principal Response Curves
Centering and Standardization
SAMPLES
None
Center by sample
Standardize by norm
Center and standardize
SPECIES
None
Center by species
Standardize by norm
Center and standardize
Standardize by error variance
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

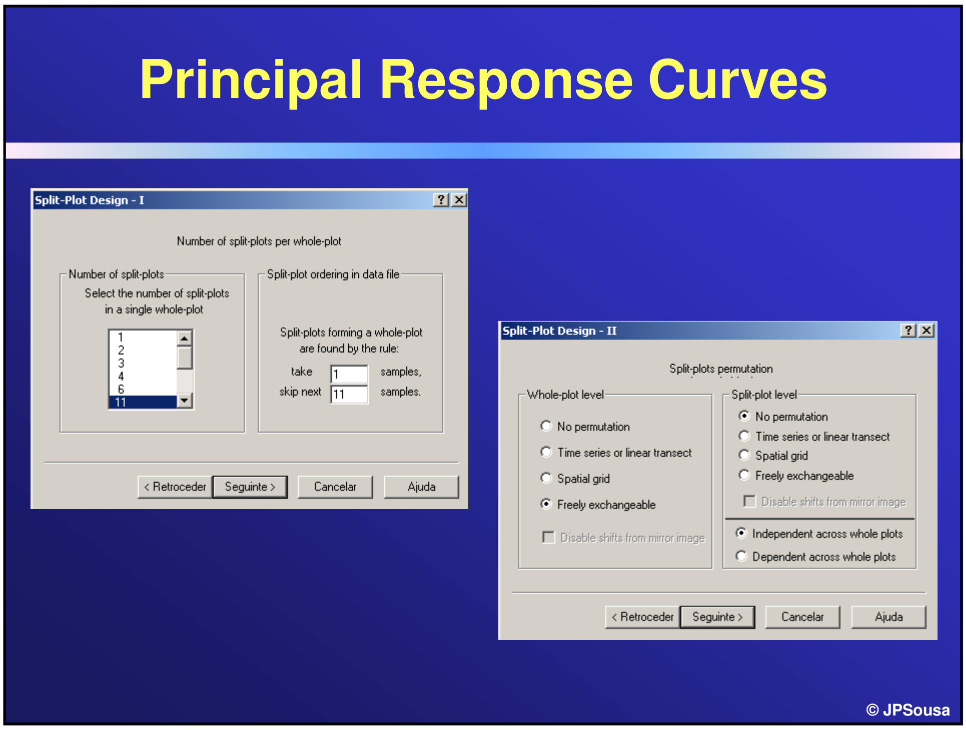

Principal Response Curves
Data Editing Choices
CHECK APPROPRIATE BOX, IF YOU WISH TO ...
SOME...
DELETE
WEIGHT
MAKE SUPPLEMENTARY
Samples
Species
DEFINE INTERACTIONS
Env. variables
Covariables
Suppl. env. vars
< Retroceder
Seguinte >
Cancelar
1. We should place as Env. Variables the treatments (from what we would like to see the effect of)
2. However the treatment is not the Endosulfan ONLY but the INTERACTION between Endosulfan and Time
3. We have to delete variables and define interactions
© JPSousa


Principal Response Curves
Delete Environmental Variables
Select env. variables to be deleted
Select items in the left list and move to the right with the >> button
SOURCE POOL
TO BE DELETED
>>
<<
[ 1] Time1
[ 2] Time2
[ 3] Time3
[ 4] Time4
[ 5] Ref
[ 6] Expos
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

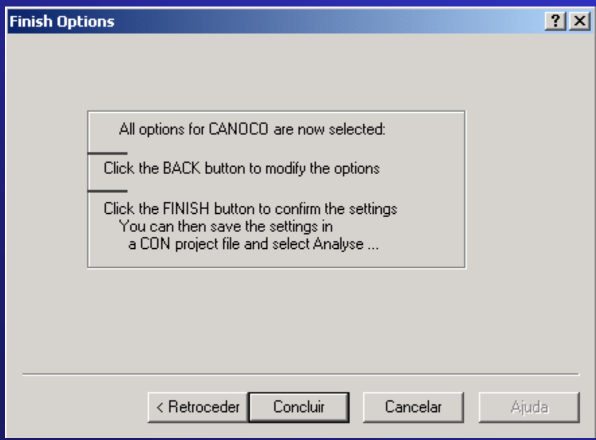

Principal Response Curves
Interactions of Environmental Variables
First variable
Second variable
[ 1] Time1
[ 2] Time2
[ 3] Time3
[ 4] Time4
[ 5] Ref
[ 6] Expos
[ 1] Time1
[ 2] Time2
[ 3] Time3
[ 4] Time4
[ 5] Ref
[ 6] Expos
Add
Powers and product variables
[ 7] Expos *Time1
[ 8] Expos *Time2
[ 9] Expos *Time3
[ 10] Expos *Time4
Remove
< Retroceder
Seguinte >
Cancelar
1. Necessary to obtain the Cdt of each new variable
2. Since the Reference values along time are always 0, no need to define the Reference as a new variable
3. Only necessary to define "Endosulfan * Time" as new variables
© JPSousa


Principal Response Curves
Forward Selection of Environmental Variables
Do not use forward selection
Automatic selection
Manual selection
Best K= 4 variables
use Monte Carlo Permutation Tests
Permutations under full model
Number of permutations: 499
< Retroceder
Seguinte >
Cancelar
Ajuda
Global Permutation Test
Evaluate current analysis with Monte-Carlo permutation test?
Do not perform the test
Significance of first ordination axis
Significance of canonical axes together
Both above tests
Number of permutations
499
Permutation under...
Reduced model
Full model
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa


Principal Response Curves
Permutation Type
Permutation Type
Unrestricted permutations
Blocks defined by covariables
Restricted for spatial or temporal structure or split-plot design
Read from file:
Browse...
Random Number Generator
Set seeds
13189
29720
Randomize...
Leverage corrected residuals, default seeds
< Retroce
Permutation Restrictions
Select type of restriction on the permutations
Time series or line transects
Disable random shifts of mirror image
Rectangular spatial grid
Split-plot design
< Retroceder
Seguinte >
Cancelar
Ajuda
1. In this case we cannot randomly permutate the samples, since they are grouped in treatments
2. So, we define the whole plots and we permutate samples between each whole plot
© JPSousa


Principal Response Curves
Split-Plot Design - I
Number of split-plots per whole-plot
Number of split-plots
Select the number of split-plots in a single whole-plot
1
2
3
4
6
11
Split-plot ordering in data file
Split-plots forming a whole-plot are found by the rule:
take 1 samples,
skip next 11 samples.
< Retroceder
Seguinte >
Cancelar
Ajuda
Split-Plot Design - II
Split-plots permutation
Whole-plot level
No permutation
Time series or linear transect
Spatial grid
Freely exchangeable
Disable shifts from mirror image
Split-plot level
No permutation
Time series or linear transect
Spatial grid
Freely exchangeable
Disable shifts from mirror image
Independent across whole plots
Dependent across whole plots
< Retroceder
Seguinte >
Cancelar
Ajuda
© JPSousa

## Principal Response Curves




… ect…



## Principal Response Curves

To calculate Cdt (the treatment PRC) we need:

- Canonical Coefficients (SOL file)
- Treatment SD (LOG file)
- TAU – total species SD (LOG file)
- Cdt = (CanCoef * TAU) / SD
- Control Cdt along time is always “0”

To construct the species interpretation aid we need the Bk’s of the species (Species scores of SOL file)

# Principal Response Curves

| | LOG FILE | LOG FILE | SOL FILE | |
|---|---|---|---|---|
| | SD | TAU | RegCoef | **Cdt** |
| Exp*Time1 | 0,2836 | 1,33447 | 0,0465 | 0,218804 |
| Exp*Time2 | 0,2836 | 1,33447 | 0,1304 | 0,613593 |
| Exp*Time3 | 0,2836 | 1,33447 | 0,2478 | 1,166014 |
| Exp*Time4 | 0,2836 | 1,33447 | 0,4078 | 1,918889 |

**Cdt = RegCoef*TAU/SD**

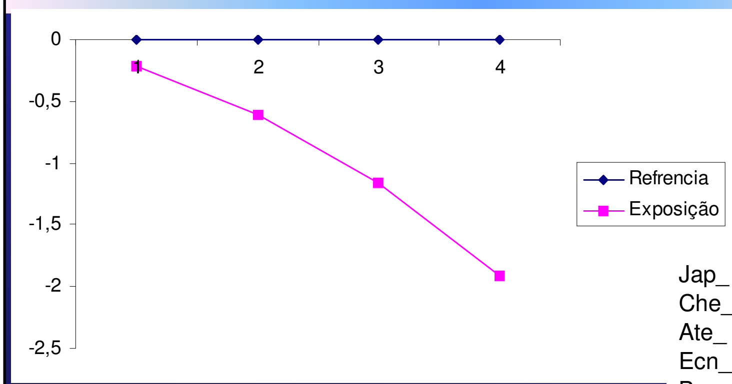
| **Partição Variabilidade** | **LOG FILE** |
|---|---|
| % Tempo | 24,7 |
| **% Tratamento** | **22,6** |
| % Residual | 52,7 |

**** Summary of Monte Carlo test **** **LOG FILE**

Test of significance of first canonical axis: eigenvalue = 0.175
F-ratio = 18.189
P-value = 0.0020



# Principal Response Curves




| | SOL FILE |
|---|---|
| Jap_ Kut | 1,7904 |
| Che_ sp | 1,3256 |
| Ate_ aus | 1,2392 |
| Ecn_ sp | 1,1366 |
| Bae_ sp | 1,1169 |
| Tas_ sp | 0,9959 |
| Moll usc | 0,5436 |
| Olig och | 0,5257 |
| Othe r | 0,407 |
| Prat ya | 0,3541 |
| Chi_ spp | 0,3282 |

# CANOCO for Windows

Hands on !
Part 6



# Partition of Variation

- Interest in evaluating the importance of different explanatory variables in explaining the response variables:
  - Ex: verify the importance of forest type or vegetation cover in explaning the species composition of soil fauna communities
  - Ex: verify the influence of space and metal concentration in influencing the allele frequency in *Orchesella cincta*
- % of variation of the response variables explained by each environmental variable (or groups of environmental variables)
- Done via several CCA/RDA and playing with co-variables

## Hands on ! Allele frequency

### Allele frequency in *Orchesella cincta*

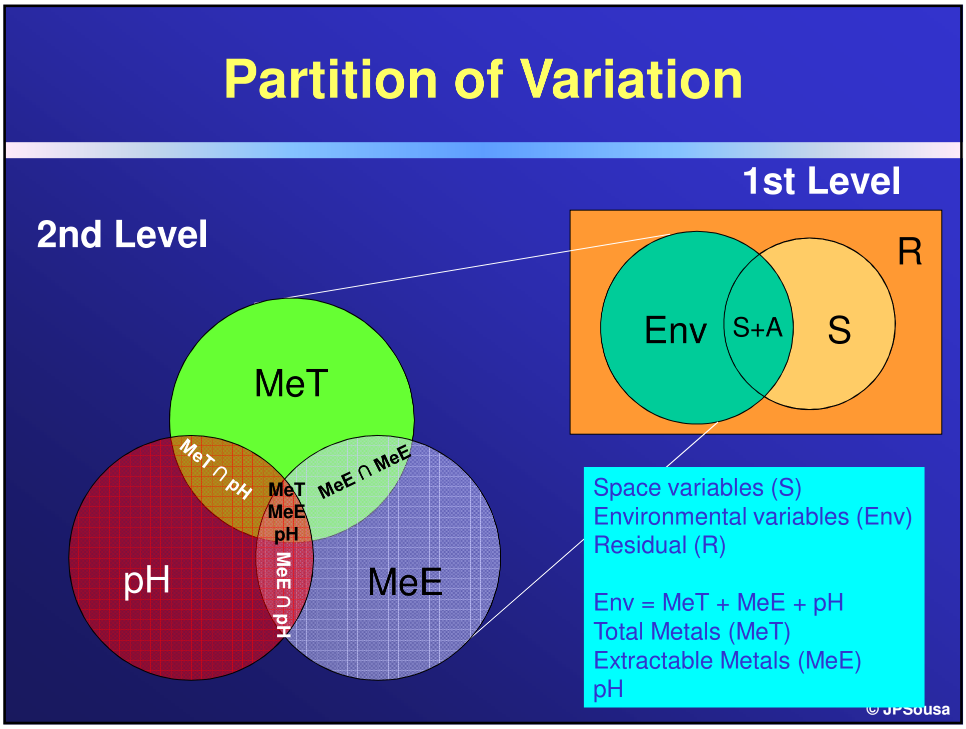
- Data on MT allele promoter frequencies (File: Thierry_data.xls);
- Several sites (divided in to 4 site types) in Belgium;
- Allele frequencies & Explanatory variables (pH, Metals & Spatial variables).



## Hands on ! Allele frequencies

Aims of the Exercise:

1. Evaluate the importance of spatial variables and environmental variables in explaining allele frequency data
2. Evaluate the influence of the different environmental variables (pH, total metals & extractable metals) in explaining allele frequency




## CANOCO for Windows - RDA

### Step 1 – Select variables from each group

- Perform a CCA/RDA using "forward selection" for each group of variables, i.e., spatial coordinates, total metals, available metals and pH– selected variables
- Take care not to have unbalanced groups – over weighting !
- Confirm the results using Monte-Carlo permutations

# CANOCO for Windows - RDA

**Step 2 – Perform variance partition on 1st level**
**Use only selected variables; Species matrix is always the same**

- Perform a CCA/RDA with all selected environmental and spatial variables (no co-variables) – % total var. explained (Env + E&S + S)
- Perform a CCA/RDA with all selected environmental variables (spatial variables as co-variables) – Env
- Perform a CCA/RDA with all selected spatial variables (environmental variables as co-variables) – S
- Calculate shared variance (E&S) by the difference



# CANOCO for Windows - RDA

**Step 3 – Perform Level 2 decomposition of variance**
**Use only selected variables; Species matrix is always the same**

- Use the same principle to calculate each partition of the variation
- Do not forget to use ALWAYS space variables as co-variables in the analysis
- Variables entering as co-variables are those we which to rule out their influence

# CANOCO for Windows - RDA

**Example**
Calculate %var. explained by MeT entering into account with the interaction with other variables (green circle)

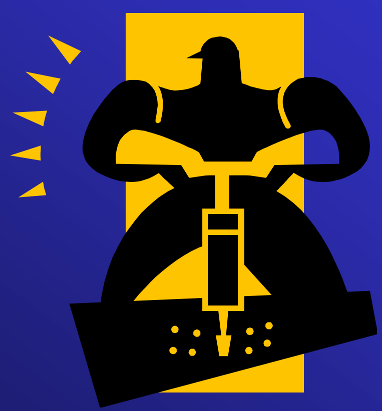
Environmental matrix: Selected MeT;
Co-variable matriz: space variables

**Exemple**
Calculate %var. explained **ONLY** by MeT, not entering into account with the interaction with other variables (part of the green circle)

Environmental matrix: Selected MeT;
Co-variable matriz: space variables + selected MeE + pH



# CANOCO for Windows



Hands on !
Part 7

Multivariate Statistical Tools in Ecology

ISCED, Lubango, March 2016

## Summarizing the Week



## Summarizing the Week


YOU

## Summarizing the Week

- Methods allowing the interpretation of the underlying structure of the data (PCA, CA, NMDS)

- Allows to visualize how samples are related based on the response variables and which response variables are more related to the groups of samples

- Important aspects to take care of:
  - Data transformations (e.g., Log for response variables with large variance)
  - Scaling (species vs. samples)
  - Centering and standardizing in PCA (mandatory when variables are in different units)

- Don't forget - length of the gradient with a DCA



## Summarizing the Week

- Methods allowing to discriminate groups of samples ('*a priori*' or '*a posteriori*' defined groups) (ANOSIM, DA)

- ANOSIM works based on all response variables (advantage to select the distance metric to use)

- DA works based on the discriminant variables; the aim is to keep it simple but having the highest discriminating power as possible; pay attention to:
  - Significance of discriminant variable (selection criteria is important)
  - Significance of discriminating functions (axis)
  - Std canonical coefficients of the variables for each axis
  - Pay attention to colinearity!

## Summarizing the Week

❖ Methods allowing to relate response variables to explanatory variables by "direct gradient analysis" (*RDA; CCA*)

❖ The key point is the variable selection;
- Check for significant variables
- Pay attention to colinearity!
- Check for the significance of the model
- Check for the % variance explained by the model and the % explained in the first axis
- Pay attention to the permutation scheme!

❖ Final model may be a "compromise solution" regarding explaining a bit less but having a good % explained in axis 1!



## Summarizing the Week

❖ Methods allowing to evaluate time changes in relative responses of treatments when compared to a control situation (Principal Response Curves - *PRC*)

❖ The key point is the permutation model
❖ You have to have all samples in every sampling times

❖ Pay attention to the numerical output (Cdt and Bk values) and decide if you have to multiply my -1.

## Summarizing the Week

- Methods allowing to the contribution of different groups of explanatory variables in explaining the response variables (*Decomposition of variance*)
- STEP 1 is variable selection (pay attention to the unbalanced number of variables in each group). Selection criteria is the same as in any RDA/CCA
- STEP 2 is to verify the influence of Space and Environment (level 1 of decomposition)
- STEP 3 is to verify the influence of the different groups of Env variables (level 2 of decomposition).



In the SOLUTION file

Spec: Species scores – Species scores

Samp: Sample scores – Samples scores derived from species scores. Point "ORIGIN" represents the origin in the original species space before centering has been applied

CFit: Cumulative fit per species as fraction of variance of species – Relative contribution of each axis to the variance of that species. A "%EXPL" represents the Cfit considering all axes together.

SqRL: Squared residual length per sample with s axes - distance between sample point and its location in the s-dimentional plane (the lower the better). The "%FIT" represents how well the samples fit into the s-dimenttional plane.

Regr: Regression/canonical coefficients for standardized variables – Are the coefficients derived from multiple regression of the species-derived sample scores (Samp) on the standardized environmental variables. Unstable when environmental variables are correlated to each other

tVal: t-values of regression coefficients – t-values for the Regr. When lower than 2.1 implies that the variable does not contribute much to the fit of the species data. This is important when selecting a sub-set of variables explaining the species data (another way as doing a forward selection of environmental variables). FR EXPLAINED – the fraction of the variance explained by the axis (= to variance expl in the summary in SOL file)

StBi: Species coordinates for t-value biplot

EtBi: Environmental coordinates for t-value biplot

CorE: Inter-set correlations of environmental variables with axes – Correlation between the environmental variables with the samples scores derived from species data (Samp). FR EXTRACTED – the fraction of the variance of Env.Variables extracted by each axis

BipE: Biplot scores of environmental variables –

CenE: Centroids of environmental variables (mean.gt.0) in ordination diagram -

SamE: Sample scores which are linear combinations of environmental variables - Samples scores derived from environmental variables. The "%FIT" represents how well the samples fit into the s-dimenttional plane.

In the LOG file

SPEC AX1 – Axis (representing the sample scores) derived from species data

ENV AX1 – Axis (representing the sample scores) derived from environmental data

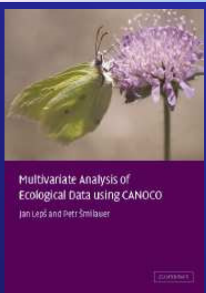
Corr “Env.Var” vs. “SPEC AX1” – Inter-set correlations (=CorE on SOL file)

Corr “Env.Var” vs. “ENV AX1” – Intra-set correlations

Corr “SPEC AX1” vs. “ENV AX1” – Species-environmental correlation



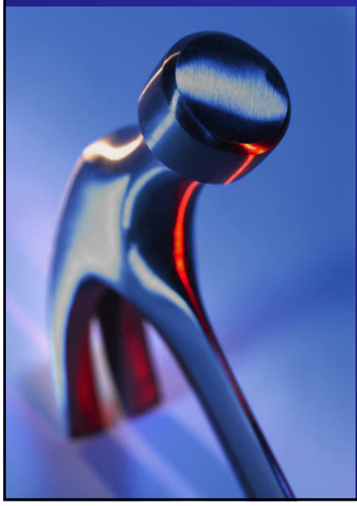
# Some important literature




Multivariate Analysis of Ecological Data using CANOCO
Jan Lepš & Petr Šmilauer
Cambridge University Press


Data Analysis in Community and Landscape Ecology
R. H. G. Jongman, C. J. F. Ter Braak, O. F. R. van Tongeren "

## Multivariate Tools



*... and a long life to the Tools!*



## *And…That is all….Folks!*


**Thanks !**