

# Multivariate analysis of ecological data with ade4

Stéphane Dray

Univ. Lyon 1

CARME 2011, Rennes

## What?

ade4 is an  package for the exploratory analysis of ecological data

- Multivariate analysis
- Graphics

It contains

- 105 datasets
- 345 functions
  - 37 multivariate methods (16 developed in the lab)
  - 39 graphical functions

## What?

ade4 is an  package for the exploratory analysis of ecological data

- Multivariate analysis
- Graphics

It contains

- 105 datasets
- 345 functions
  - 37 multivariate methods (16 developed in the lab)
  - 39 graphical functions

## Why?

- to promote the methodological developments of the lab
- to facilitate the use by ecologists of these new statistical methods

## What?

ade4 is an  package for the exploratory analysis of ecological data

- Multivariate analysis
- Graphics

It contains

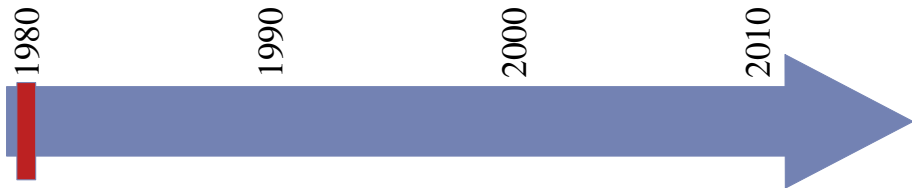
- 105 datasets
- 345 functions
  - 37 multivariate methods (16 developed in the lab)
  - 39 graphical functions

## Why?

- to promote the methodological developments of the lab
- to facilitate the use by ecologists of these new statistical methods

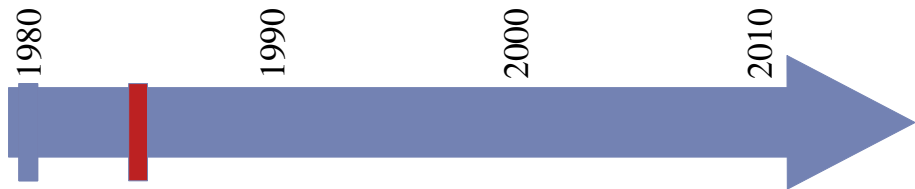
## How?

A long history



1980 : Set of programs written in BASIC  
on a Data General Nova 3

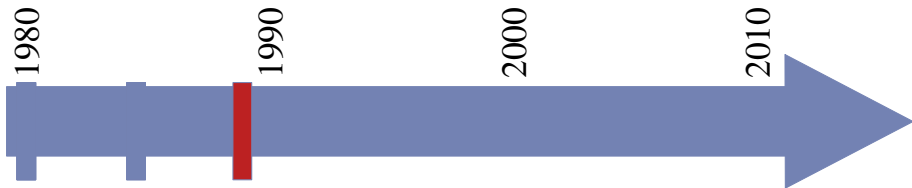




1985 : Diagonalization procedure (assembly language for the Eclipse S/140)

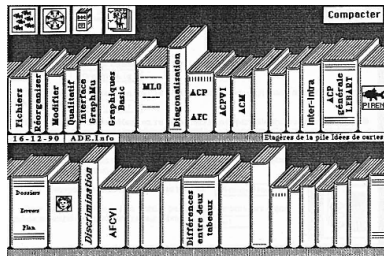
- ▶ Analysis of real ecological datasets in a "reasonable time"
- ▶ Use by the ecologists of the lab

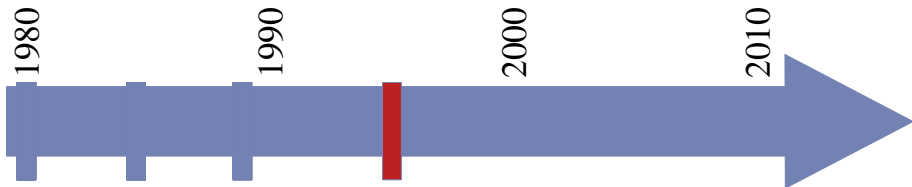




1989 : Distribution of ADECO on 

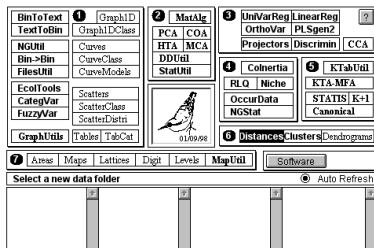
- modules in Microsoft QuickBasic
- Hypercard interface

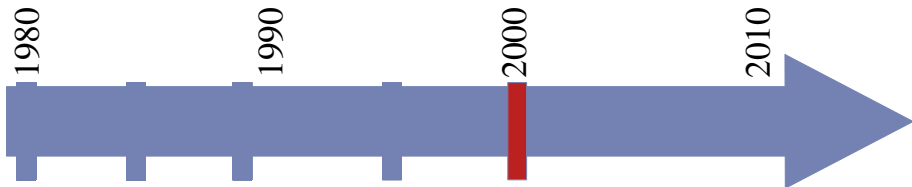




1995 : ADE-4

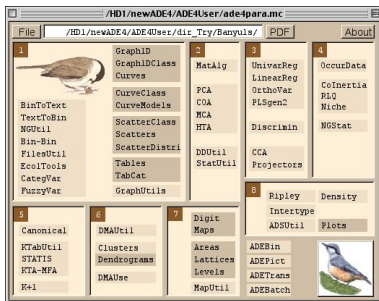
- modules in C
- Hypercard  and Winplus  interfaces

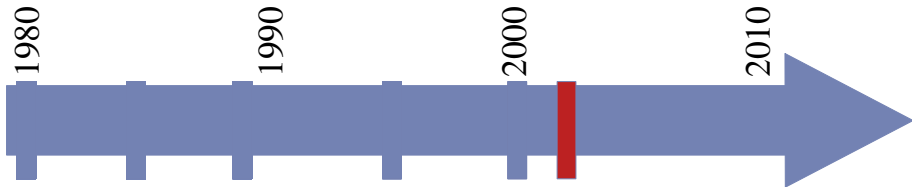





2000 : ADE-4

- Metacard interface
- batch mode





2002 : ade4 package for 

**DATA**

```
head(rpd15fau)
      C1  C2  C3  C4  C5  C6  C7  C8  C9  C10
1  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
2  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
3  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
4  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
5  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
```

**DOCUMENTATION**

```
## Canonical Correspondence Analysis
## Perform a Canonical Correspondence Analysis.
##
## Usage:
##   cca(site, sitenr, scenef = TRUE, nf = 2)
##
## Arguments:
##   site: a matrix of species data
##   sitenr: a vector of site numbers
##   scenef: a vector of site names
##   nf: the number of axes to retain
```

**GRAPHICAL AND STATISTICAL FUNCTIONS**

```
plot@ccall
```

**REFERENCES**

```
Ter Braak, C. J. F. (1986) Canonical correspondence analysis: a
new eigenvector technique for multivariate direct gradient
analysis. 'Ecology', 67, 1167-1179.

Ter Braak, C. J. F. (1987) The analysis of vegetation-environment
relationships by canonical correspondence analysis. 'Vegetatio',
69, 69-77.

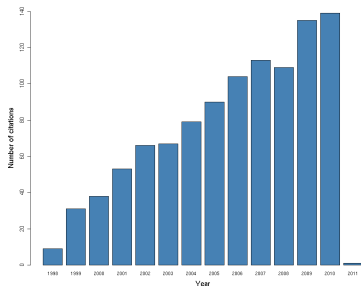
Chessel, D., Labretton, J. D. and Yoccoz, N. (1997) Proprietas de
l'analyse canonique des correspondances: une utilisation en
hydrobiologie. 'Revue de Statistique Appliquee', 35, 55-72.
```

## Who?

## The ade4 users (a bibliographic study)

An increasing community ...

... of ecologists

Thioulouse et al. (1997) *Statistics and Computing*.Chessel et al. (2004) *R News*.Dray et al. (2007) *R News*.Dray and Dufour (2007) *JSS*.

Subject Area	(%)
ECOLOGY	30.27 %
MARINE & FRESHWATER BIOLOGY	18.95 %
ENVIRONMENTAL SCIENCES	12.18 %
MICROBIOLOGY	8.51 %
PLANT SCIENCES	8.31 %
SOIL SCIENCE	6.67 %
GENETICS & HEREDITY	6.18 %
EVOLUTIONARY BIOLOGY	6.09 %
BIODIVERSITY CONSERVATION	5.60 %
BIOCHEMISTRY & MOLECULAR BIOLOGY	5.41 %
FORESTRY	5.41 %
LIMNOLOGY	5.31 %
...	...

# Ecology, a fertile ground for methodological developments

Great diversity

- 1 Biological questions/models

# Ecology, a fertile ground for methodological developments

Great diversity

- 1 Biological questions/models
- 2 Sampling methods/tools

# Ecology, a fertile ground for methodological developments

## Great diversity

- 1 Biological questions/models
- 2 Sampling methods/tools
- 3 Data structures
  - variables (quantitative, qualitative, ordinal, fuzzy, etc)
  - constraints on individuals or variables (weights, spatial, phylogenetic, hierarchical, etc)

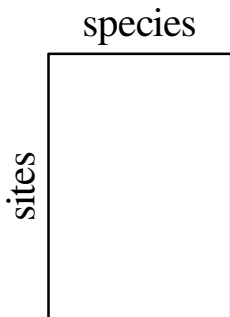
# Ecology, a fertile ground for methodological developments

## Great diversity

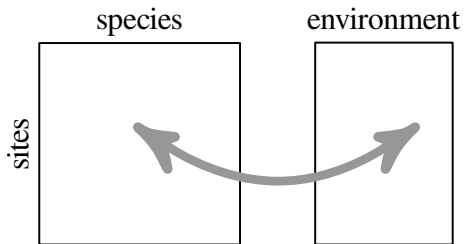
- 1 Biological questions/models
- 2 Sampling methods/tools
- 3 Data structures
  - variables (quantitative, qualitative, ordinal, fuzzy, etc)
  - constraints on individuals or variables (weights, spatial, phylogenetic, hierarchical, etc)

Usually, multivariate data

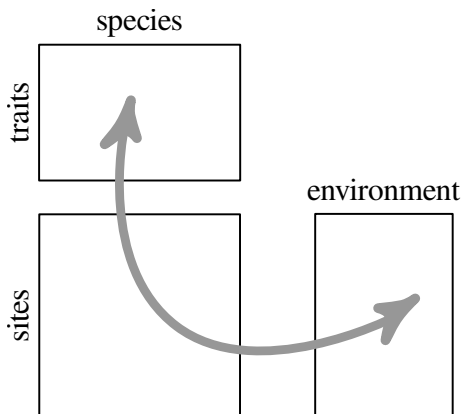
# One table : summarizing community data



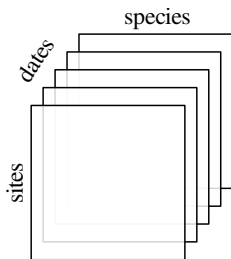
# Two tables : linking species to environment



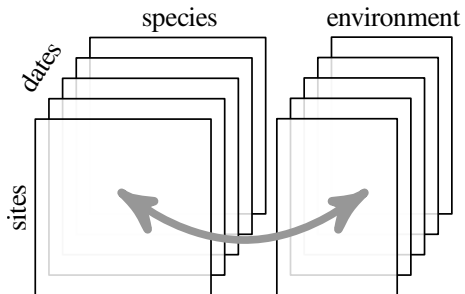
# Three tables : linking species traits to environment



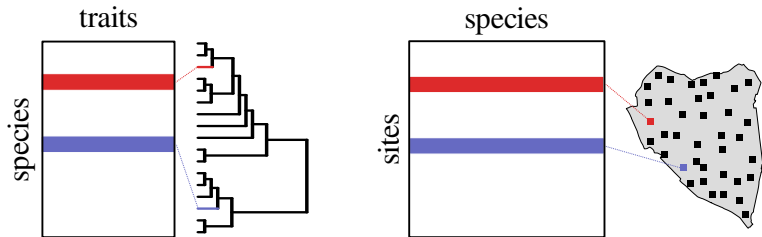
# K-tables : temporal evolution of structures



# K-tables : temporal evolution of co-structures



# Some constraints : space, phylogeny



# "French way" of multivariate analysis



*Journal of Statistical Software*

September 2007, Volume 22, Issue 4.

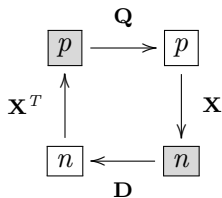
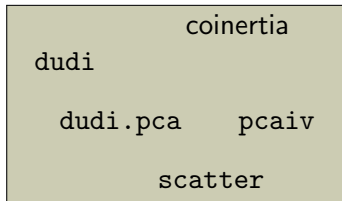
<http://www.jstatsoft.org/>

The ade4 Package: Implementing the Duality  
Diagram for Ecologists

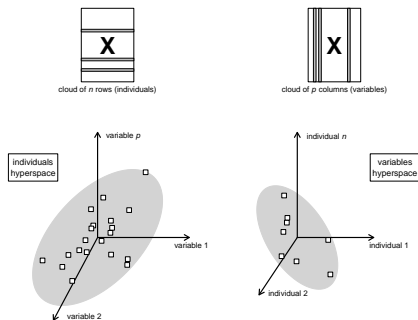
**ade4**

=

**theory**



# One table, two viewpoints



- what are the relationships between the variables ?
- what are the resemblances/differences between the individuals ?

## Statistical triplet

Multivariate methods aim to answer these two questions and seek for small dimension hyperspaces (few axes) where the representations of individuals and variables are as close as possible to the original ones.

To answer the two previous questions, we define

- **Q**, a  $p \times p$  positive symmetric matrix, used as an inner product in  $\mathbb{R}^p$  and thus allows to measure distances between the  $n$  individuals
- **D**, a  $n \times n$  positive symmetric matrix, used as an inner product in  $\mathbb{R}^n$  and thus allows to measure relationships between the  $p$  variables.

$$(\mathbf{X}, \mathbf{Q}, \mathbf{D})$$

$$\mathbf{XQX}^T \mathbf{DK} = \mathbf{K}\Lambda$$

$$\mathbf{X}^T \mathbf{DXQA} = \mathbf{A}\Lambda$$

- $\mathbf{K}$  contains the principal components ( $\mathbf{K}^T \mathbf{DK} = \mathbf{I}_r$ ).
- $\mathbf{A}$  contains the principal axis ( $\mathbf{A}^T \mathbf{QA} = \mathbf{I}_r$ ).
- $\mathbf{L} = \mathbf{XQA}$  contains the row scores (projection of the rows of  $\mathbf{X}$  onto the principal axes)
- $\mathbf{C} = \mathbf{X}^T \mathbf{DK}$  contains the column scores (projection of the columns of  $\mathbf{X}$  onto the principal components)

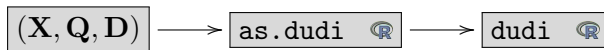
Maximization of :

$$Q(\mathbf{a}) = \mathbf{a}^T \mathbf{Q}^T \mathbf{X}^T \mathbf{DXQA} = \lambda \text{ and } S(\mathbf{k}) = \mathbf{k}^T \mathbf{D}^T \mathbf{XQX}^T \mathbf{DK} = \lambda$$

$$\langle \mathbf{XQA} | \mathbf{k} \rangle_{\mathbf{D}} = \langle \mathbf{X}^t \mathbf{DK} | \mathbf{a} \rangle_{\mathbf{Q}} = \sqrt{\lambda}$$

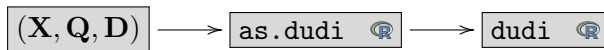
## Implementation in ade4

Computation are performed by the `as.dudi` (diagonalization in the smaller dimension) function :



## Implementation in ade4

Computation are performed by the `as.dudi` (diagonalization in the smaller dimension) function :



## "Generic" function of the dudi class

- `print.dudi` : display a dudi object
- `is.dudi` : test if an object is of the class dudi
- `redo.dudi` : recomputes an analysis with new number of axes
- `t.dudi` : transpose a dudi  $(\mathbf{X}, \mathbf{Q}, \mathbf{D}) \rightarrow (\mathbf{X}^T, \mathbf{D}, \mathbf{Q})$
- `scatter.dudi` / `biplot.dudi` : biplot
- `screeplot.dudi` : barplot of eigenvalues
- `summary.dudi` : main information related to an analysis
- `inertia.dudi` : inertia statistics (absolute, relative =  $\cos^2$ )
- `dist.dudi` : dudi-based distance among rows/columns
- `reconst` : data approximation
- `suprow` / `supcol` : projection of supplementary rows/columns

## User-level functions

- The `as.dudi` function is an internal function.
- Is called by user-friendly functions corresponding to different analyses.
- It can be used by experimented users to define their own analysis.

```
apropos("dudi.")
```

```
[1] "dudi.acm"      "dudi.coa"      "dudi.dec"
[4] "dudi.fca"      "dudi.fpca"     "dudi.hillsmith"
[7] "dudi.mix"      "dudi.nsc"      "dudi.pca"
[10] "dudi.pco"
```

# Available methods



Function name	Analysis name
dudi.pca	Principal component analysis
dudi.pco	Principal coordinate analysis
dudi.coa	Correspondence analysis
dudi.acm	Multiple correspondence analysis
dudi.dec	Decentered correspondence analysis
dudi.fca	Fuzzy correspondence analysis
dudi.fpca	Fuzzy PCA
dudi.mix	Mixed type analysis
dudi.hillsmith	Hill & Smith type analysis
dudi.nsc	Non-symmetric correspondence analysis

Principal Component Analysis : `dudi.pca(df)`

- $\mathbf{X} = [x_{ij} - (\bar{x}^j)]$
- $\mathbf{Q} = \mathbf{I}_p$
- $\mathbf{D} = \frac{1}{n} \mathbf{I}_n$

Maximization of :

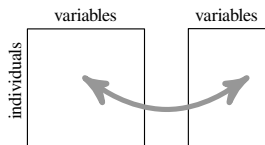
$$Q(\mathbf{a}) = \mathbf{a}^\top \mathbf{Q}^\top \mathbf{X}^\top \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{a} = \| \mathbf{X} \mathbf{Q} \mathbf{a} \|_{\mathbf{D}}^2 = \text{var}(\mathbf{X} \mathbf{Q} \mathbf{a})$$

$$S(\mathbf{k}) = \mathbf{k}^\top \mathbf{D}^\top \mathbf{X} \mathbf{Q} \mathbf{X}^\top \mathbf{D} \mathbf{k} = \| \mathbf{X}^\top \mathbf{D} \mathbf{k} \|_{\mathbf{Q}}^2 = \sum_{j=1}^p \text{cov}^2(\mathbf{k}, \mathbf{x}^j)$$

# Graphic function

Function name	Type of graph
s.arrow	Factor map with arrows (projection of a vector basis)
s.chull	Factor map with convex hulls
s.class	Factor map with classes of points
s.corcircle	Factor map with correlation circle
s.distri	Factor map with frequency distribution
s.hist	Factor map with marginal histograms
s.image	Factor map with background image and contour curves
s.kde2d	Factor map with kernel density estimation
s.label	Factor map with labels
s.logo	Factor map with logos (pictures)
s.match	Factor map with paired coordinates
s.match.class	Factor map with paired coordinates and classes of points
s.multinom	Factor map with frequency profiles (genetics)
s.traject	Factor map with trajectories
s.value	Factor map with symbols proportional to some values
sco.boxplot	Boxplots on a score for a set of factors
sco.class	Labels on a score grouped by a factor
sco.distri	Mean & Std Dev for a weighted score
sco.gauss	Gauss curves on a score and a set of factors
sco.label	Labels on a score
sco.match	labels on two scores
sco.quant	Relations between a score and quantitative variables

# Available methods



Function name	Analysis name
between	Between-class analysis
within	Within-class analysis
discrimin	Discriminant analysis
coinertia	Coinertia analysis
cca	Canonical correspondence analysis
pcaiv	PCA on Instrumental Variables
pcaivortho	Orthogonal PCAIV
procuste	Procustes analysis
niche	Niche (OMI) analysis

# Principal component analysis on instrumental variables : `pcaiv(dudi,df)`

- $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$
- $\mathbf{Z}$  a  $n \times q$  matrix of explanatory variables

$$(\hat{\mathbf{X}}, \mathbf{Q}, \mathbf{D})$$

where :

$$\hat{\mathbf{X}} = \mathbf{P}_Z \mathbf{X} = \mathbf{Z}(\mathbf{Z}^T \mathbf{D} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{D} \mathbf{X}$$

Maximization of :

$$Q(\mathbf{a}) = \mathbf{a}^T \mathbf{Q}^T \hat{\mathbf{X}}^T \mathbf{D} \hat{\mathbf{X}} \mathbf{Q} \mathbf{a} = \|\hat{\mathbf{X}} \mathbf{Q} \mathbf{a}\|_{\mathbf{D}}^2 = \text{var}(\hat{\mathbf{X}} \mathbf{Q} \mathbf{a})$$

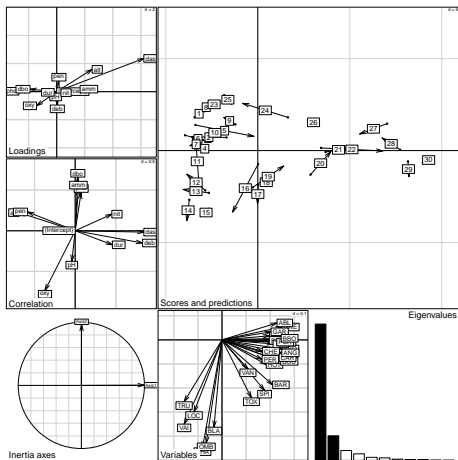
$$S(\mathbf{k}) = \mathbf{k}^T \mathbf{D}^T \hat{\mathbf{X}} \mathbf{Q} \hat{\mathbf{X}}^T \mathbf{D} \mathbf{k} = \|\hat{\mathbf{X}}^T \mathbf{D} \mathbf{k}\|_{\mathbf{Q}}^2 = \sum_{j=1}^p \text{cov}^2(\mathbf{k}, \hat{\mathbf{x}}^j)$$

## Specific summary and plot function :

```

data(doubs)
acp1 <- dudi.pca(doubs$poi, scannf = FALSE)
pcaiv1 <- pcaiv(acp1, doubs$mil, scannf = FALSE)
plot(pcaiv1)

```



Co-inertia analysis :  $\text{coinertia}(\text{dudiX}, \text{dudiY})$ 

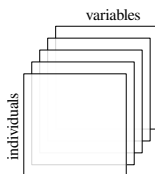
- $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$
- $(\mathbf{Y}, \mathbf{R}, \mathbf{D})$

$$(\mathbf{X}^T \mathbf{D} \mathbf{Y}, \mathbf{R}, \mathbf{Q})$$

Maximization of :

$$\langle \mathbf{X}^T \mathbf{D} \mathbf{Y} \mathbf{R} \mathbf{a} | \mathbf{k} \rangle_{\mathbf{Q}} = \mathbf{k}^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{Y} \mathbf{R} \mathbf{a} = \text{cov}(\mathbf{Y} \mathbf{R} \mathbf{a}, \mathbf{X} \mathbf{Q} \mathbf{k})$$

## K-table



Function name	Analysis name
sepan	K-table separate analyses
pta	Partial triadic analysis
foucart	Foucart analysis
statis	STATIS analysis
mfa	Multiple factor analysis
mcoa	Multiple coinertia analysis
statico	2 K-table analysis

## K-table : class ktab

Series of tables are stored in object of class `ktab` which can be created using :

- `ktab.within`
- `ktab.list.df`
- `ktab.list.dudi`
- `ktab.data.frame`

## Other methods

Function name	Analysis name
witwit.coa	Internal correspondence analysis
betweencoinertia	Between-class coinertia analysis
withincoinertia	Within-class coinertia analysis
rlq	RLQ analysis
dpcoa	Double principal coordinate analysis
multispati	Spatial data analysis

- `adehabitat` : analysis of habitat selection by animals
- `adegetet` : classes and methods for the multivariate analysis of genetic markers
- `adephylo` : exploratory analyses for the phylogenetic comparative method

# ade4TkGUI : ade4 Tcl/Tk Graphical User Interface

The screenshot displays the ade4TkGUI interface with three main windows:

- dudi.pca (Principal components analysis):** Shows input data frame 'doubtsmil' (30 x 11), output name 'acpmil', and options for centering and scaling. It includes a 'Duality diagram' section with a plot of Eigenvalues (6.322, 2.232, 1.004, 0.50) and lists vectors and dataframes.
- R Graphics: Device 2 (ACTIVE):** Displays a scatter plot of the first two principal components with points labeled 'dud' and 'pho'.
- ade4TkGUI (Main Panel):** Features a menu bar (File, Windows, 1table, 1table-groups, 2tables, Graphics) and a toolbar with buttons for 'Read a data file', 'Load a data set', 'PCA', 'COA', 'MCA', 'PCO', 'BGGA', 'WGA', 'DA', 'Coinertia', 'CCA', 'PCAIV', 'Labels', 'Classes', 'Values', 'dudi display', 'MCTests', 'Explore', 'ordiCust', 'Quit R (save)', 'Dismiss', and 'Quit R (don't save)'.

The console window at the bottom left shows the following R commands:

```
> ade4TkGUI(T,T)
> acpmil <- dudi.pca(doubtsmil)
> scatter(acpmil,1,2)
> 
```

- mailing list <http://listes.univ-lyon1.fr/wws/info/adelist>
- web sites
  - Development on R-Forge :  
[https://r-forge.r-project.org/R/?group\\_id=199](https://r-forge.r-project.org/R/?group_id=199)
  - Software : <http://pbil.univ-lyon1.fr/ADE-4/>
  - Courses : <http://pbil.univ-lyon1.fr/R/enseignement.html>

<b>Accueil</b>
Page d'entrée
Page de liens
English Section
Espace invités
Maintenance
<b>Cours</b>
Introductions
Biologie et modélisation (L)
Tests d'hypothèse
Analyse des données
Fiches de stage
Probabilité & Statistique
Écologie et Statistique
Évolution moléculaire
Modélisation
Divers
<b>Annales</b>
Biostatistiques (L)
Biologie et modélisation (L)
Analyse des données (M)
aMIG (M)
Algèbre & Statistique
Statistiques & logiciels
Autres disciplines et VRlogo
<b>Fiches de TD</b>
Le logiciel R
Biologie et modélisation (L)
Statistique descriptive

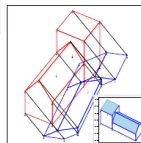
## Enseignements de Statistique en Biologie

Recherche dans ce site



Melange de lois normales

Recherche Google



[Fiches de TD / Le logiciel R / TOR17](#)

A.B. Dufour D. Chessel J.R. Lobry  
Contributeurs  
S. Mousset S. Dray  
Maintenance système S. Penel

**Notes de cours, illustrations, exercices,  
problèmes, fiches de Travaux Dirigés  
Jeux de données pour la pratique de la  
statistique**



> 300 documents > 4000 pages