Consultations statistiques avec le logiciel 

Traits biologiques : variables ou *K*-tableaux ?

Résumé

La fiche donne des indications sur la nature des tableaux de traits biologiques. A partir de plusieurs exemples, on montre qu'il est possible de choisir entre deux types de stratégies, la première visant à transformer un ensemble de traits en paquets de variables quantitatives, la seconde gardant au tableau de traits le statut de *K*-tableaux et cherchant à le décomposer en éléments plus simples. La présence sous-jacente de la taxonomie pose des questions ouvertes.

.

Plan

1. Introduction

La question est posée dans la fiche *Traits biologiques des insectes* :

http://pbil.univ-lyon1.fr/R/pps/pps029.pdf

qui permet d'utiliser les données publiées par Statzner *et al.* (1997) . Sont réunies ici quelques éléments de réflexion sur ce type très particulier d'information numérique qu'on appelle *codage flou* (Chevenet et al. 1994) . Il s'agit de codage de l'information biologique conçue et réalisée par des biologistes (Bournaud *et al.* 1992) et non de codage numérique, procédure de transformation de variables quantitatives en variables qualitatives pour leur usage en analyse des correspondances (van Rijckevorsel 1987; van Rijckevorsel 1988; Cazes 1990).

Un trait biologique a m modalités. Pour chaque taxon constituant une ligne du tableau sont enregistrées $a_1, a_2, ..., a_m$ des notes d'association entre le taxon et chacune des modalités du trait. La donnée est en fait une distribution de fréquence ou une pondération des modalités associées au taxon. On calcule $a_\cdot = \sum_{j=1,m} a_j$ et $p_j = a_j / a_\cdot$.

Lorsqu'il y a plusieurs taxons et plusieurs traits, on notera :

- T, le nombre total de taxa ;
- t, le numéro d'ordre d'un taxon $1 \le t \le T$;
- V, le nombre total de traits biologiques ;
- v, le numéro d'ordre d'un trait biologique $1 \le v \le V$;
- M_v , le nombre total de modalités du trait v ;
- m_v , le numéro d'ordre d'une modalité du trait $1 \le m_v \le M_v$;
- a_{tm_v} la note d'association du taxon t à la modalité m_v
- $a_{t\bullet_v}$ la somme des note d'association du taxon t aux modalités du trait v ;
- f_{tm_v} le poids (fréquence) de la modalité m_v pour le taxon t ;

$$f_{tm_v} = a_{tm_v} / a_{t\bullet_v}$$

On a évidemment :

$$\sum_{m_v=1}^{m_v=M_v} f_{tm_v} = 1$$

- M le nombre total des modalités de tous les traits ($M = M_1 + M_2 + \ldots + M_v$) ;
- m le numéro d'ordre d'une modalité dans la série de toutes les modalités de tous les traits $1 \le m \le M$;

Les tableaux taxa-traits ont donc T lignes (taxa) et M colonnes (modalités) et se décomposent en V sous-tableaux (traits) contenant respectivement $m_1, m_2, \ldots, m_V$ colonnes (modalités des traits). On est directement, avec la notion de traits biologiques dans la partie multi-tableaux de la statistique multivariée. C'est pourquoi il a été fait peu d'usage de ce type d'informations et que de nombreuses décisions sont à prendre en ce qui concerne leur manipulation. Qu'est-ce que la variabilité d'un trait, la corrélation entre deux traits, la redondance d'une famille de traits sont des questions à résoudre.

2. Représenter un tableau de traits

Une contrainte est omniprésente : elle concerne mes données manquantes.

```
data(coleo)
coleo.fuzzy <- prep.fuzzy.var(coleo$tab, coleo$col.blocks)
```

```
2 missing data found in block 1
1 missing data found in block 3
2 missing data found in block 4
```

Ces données (Bournaud *et al.* 1992) regroupent 110 taxa et 32 modalités de 10 traits

```
coleo$col.blocks
```

```
    Vegetation       Sediment       Velocity      Pollution      Saprobity
             4              5              4              2              4
          Diet       Feeding    Temperature Temp_Amplitude
             3              5              3              2
```

```
coleo$tab[,1:4]
```

```
    1a 1b 1c 1d
1    5  6  2  2
2    6  3  2  0
...
27   4  4  2  2
28   0  0  0  0
29   4  4  2  2
...
105  4  0  0  0
106  0  0  1  0
107  0  0  0  0
108  2  2  1  2
109  2  1  1  0
110  6  1  0  0
```

Les données manquantes sont notées par une suite de zéros.

```
coleo.fuzzy[,1:4]
```

```
        1a     1b     1c     1d
1   0.3333 0.4000 0.1333 0.1333
2   0.5455 0.2727 0.1818 0.0000
...
27  0.3333 0.3333 0.1667 0.1667
28  0.4680 0.2606 0.1521 0.1193
29  0.3333 0.3333 0.1667 0.1667
...
105 1.0000 0.0000 0.0000 0.0000
106 0.0000 0.0000 1.0000 0.0000
107 0.4680 0.2606 0.1521 0.1193
108 0.2857 0.2857 0.1429 0.2857
109 0.5000 0.2500 0.2500 0.0000
110 0.8571 0.1429 0.0000 0.0000
```

Dans le tableau recodé, les données sont en fréquence et le profil moyen de tous les points renseignés est affecté au bloc manquant.

```
apply(coleo.fuzzy[,1:4],2,mean)
```

```
    1a     1b     1c     1d
0.4680 0.2606 0.1521 0.1193
```

De cette manière un recentrage remettra la point à l'origine. On suppose que chaque taxon a lui-même un poids qu'on notera π_t $(1 \leq t \leq T)$. Ce poids sera $1/T$ si on décide d'une pondération uniforme des taxa (ils ont tous la même importance *a priori*) ou ces poids peuvent provenir d'un tableau faunistique (ils auront un poids proportionnel à leur abondance).

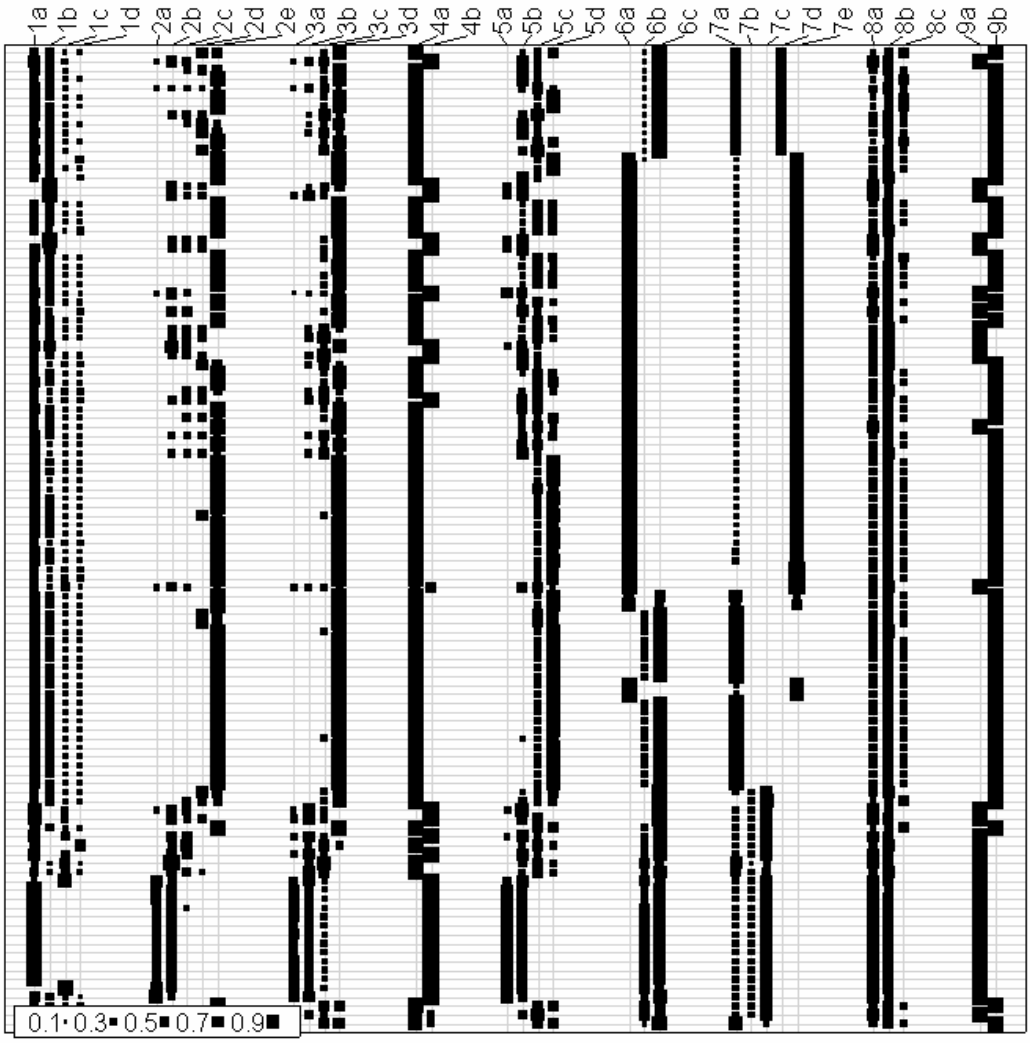
Pour le trait *t*, la moyenne du trait est défini par la pondération moyenne :

$$f*_{m_v} = \sum_{t=1}^{T} \pi_t f_{tm_v}$$

Il est aisé de visualiser entièrement les tableaux de traits.

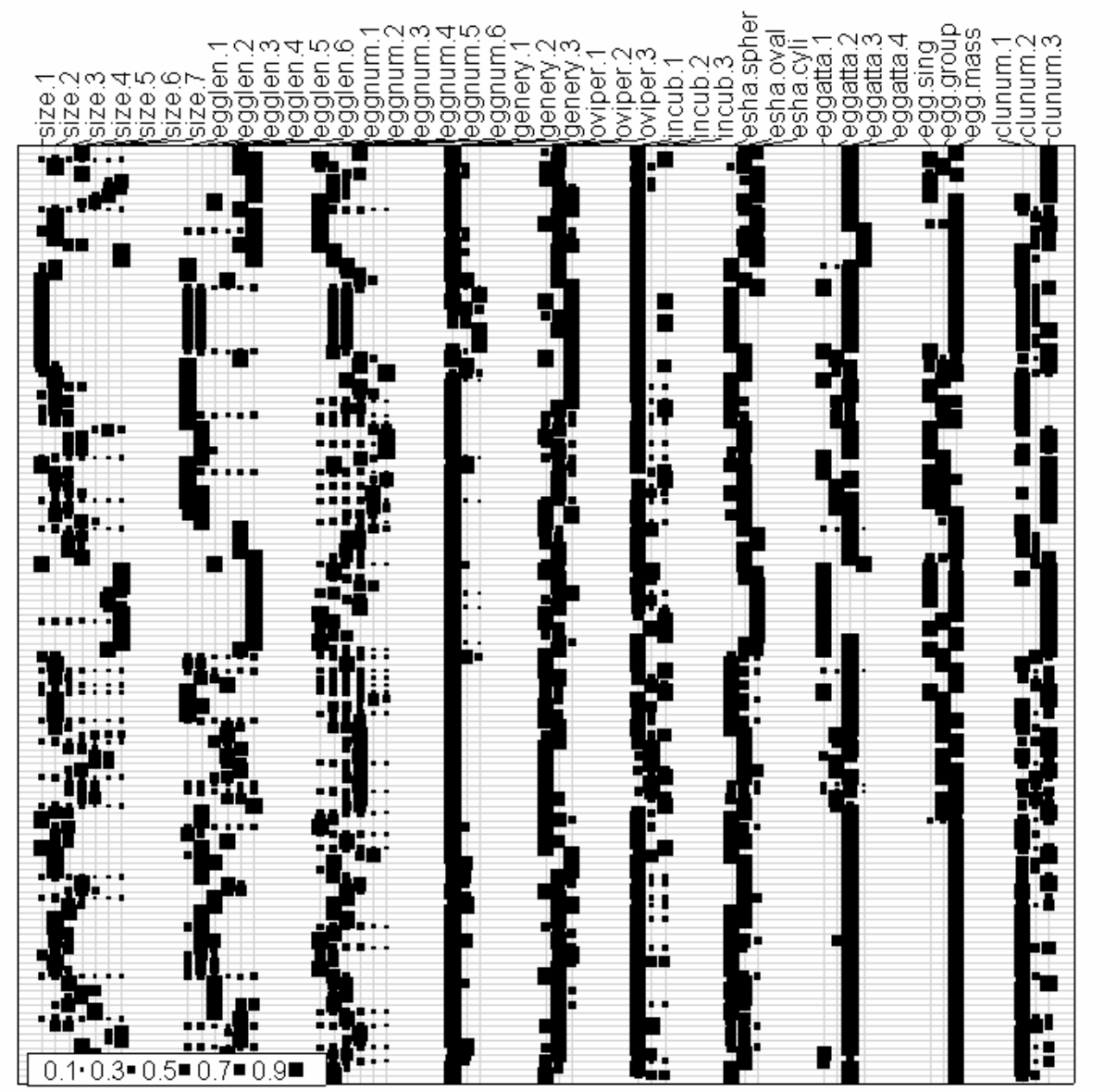
```
coleo$col.blocks
    Vegetation       Sediment       Velocity      Pollution      Saprobity
             4              5              4              2              4
          Diet       Feeding    Temperature Temp_Amplitude
             3              5              3              2
w = coleo$col.blocks
ww1=1:sum(w)
ww1
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32
ww0=seq(from=0,by=4,len=length(w))
ww0=rep(ww0,w)
ww0
 [1]  0  0  0  0  4  4  4  4  4  8  8  8  8 12 12 16 16 16 16 20 20 20 24 24 24
[26] 24 24 28 28 28 32 32

table.value(coleo.fuzzy,x=ww1,csi=0.25,clabel.row=0)
```

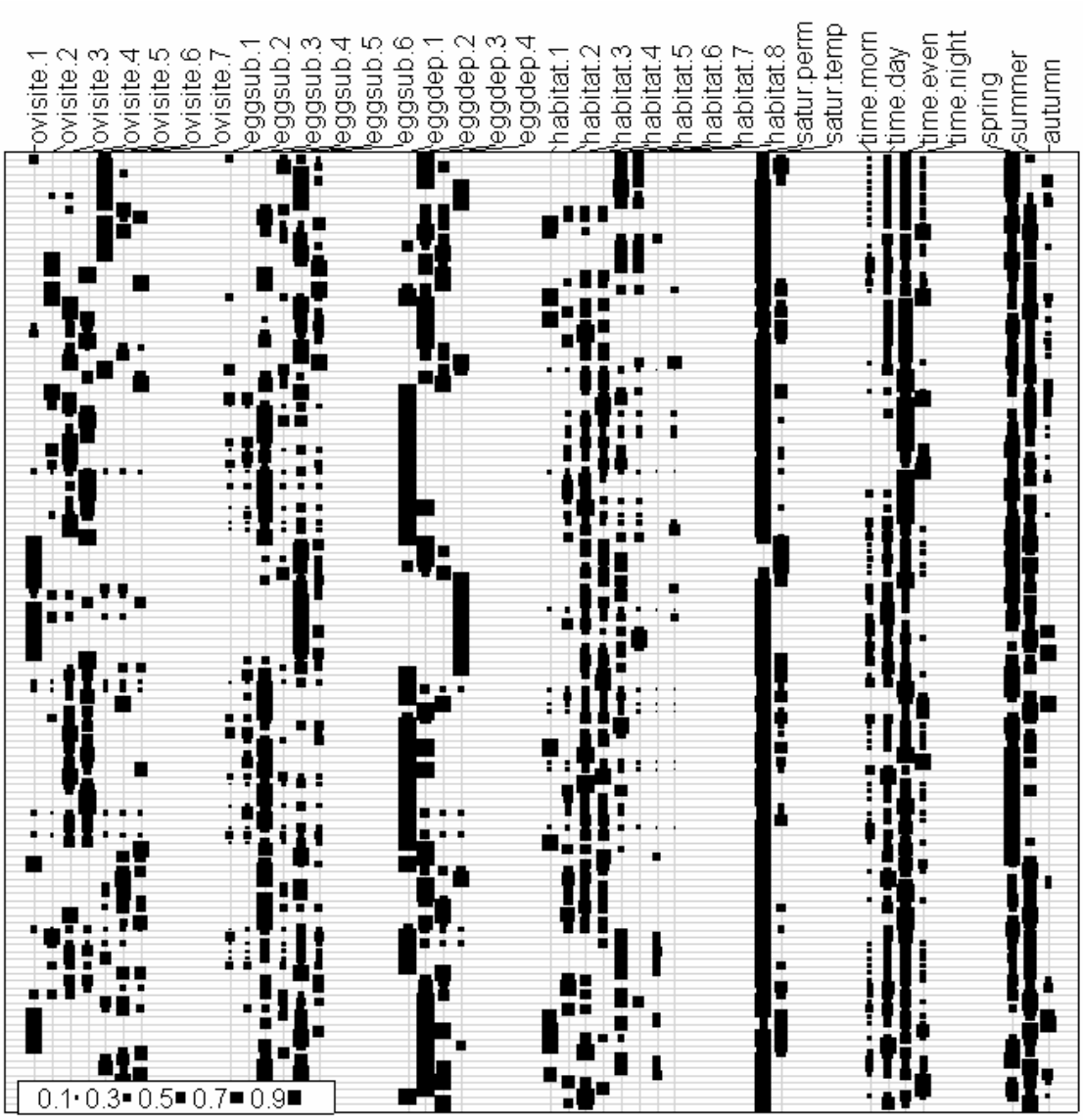


Représentation d'un tableau de traits biologiques. Chaque ligne est une espèce et chaque groupe de colonnes est une variable.

```
data(bsetal97)
w = bsetal97$biol.blo
ww1=1:sum(w)
ww0=seq(from=0,by=4,len=length(w))
ww0=rep(ww0,w)+ww1
biol.fuzzy = prep.fuzzy.var(bsetal97$biol,bsetal97$biol.blo)
table.value(biol.fuzzy,x=ww0,csi=0.25,clabel.row=0)
```




```
w = bsetal97$ecol.blo
ww1=1:sum(w)
ww0=seq(from=0,by=4,len=length(w))
ww0=rep(ww0,w)+ww1
ecol.fuzzy = prep.fuzzy.var(bsetal97$ecol,bsetal97$ecol.blo)
table.value(ecol.fuzzy,x=ww0,csi=0.25,clabel.row=0)
```



Les variables définissent des tableaux qui sont plus ou moins compliqués.

```
w = bsetal97$biol.blo[1:3]
w1 = biol.fuzzy[,1:19]
ww1 = 1:sum(w)
ww0 = seq(from=0,by=4,len=length(w))
ww0 = rep(ww0,w)+ww1
```

Préparer la taxonomie comme une phylogénie :

```
phy=taxo2phylog(as.taxo(bsetal97$taxo))
row.names(w1) = names(phy$leaves)

table.phylog(w1,phy,x=ww0,clabel.r=0,clabel.col=0.75,csi=0.5,clabel.n=0.75)
```

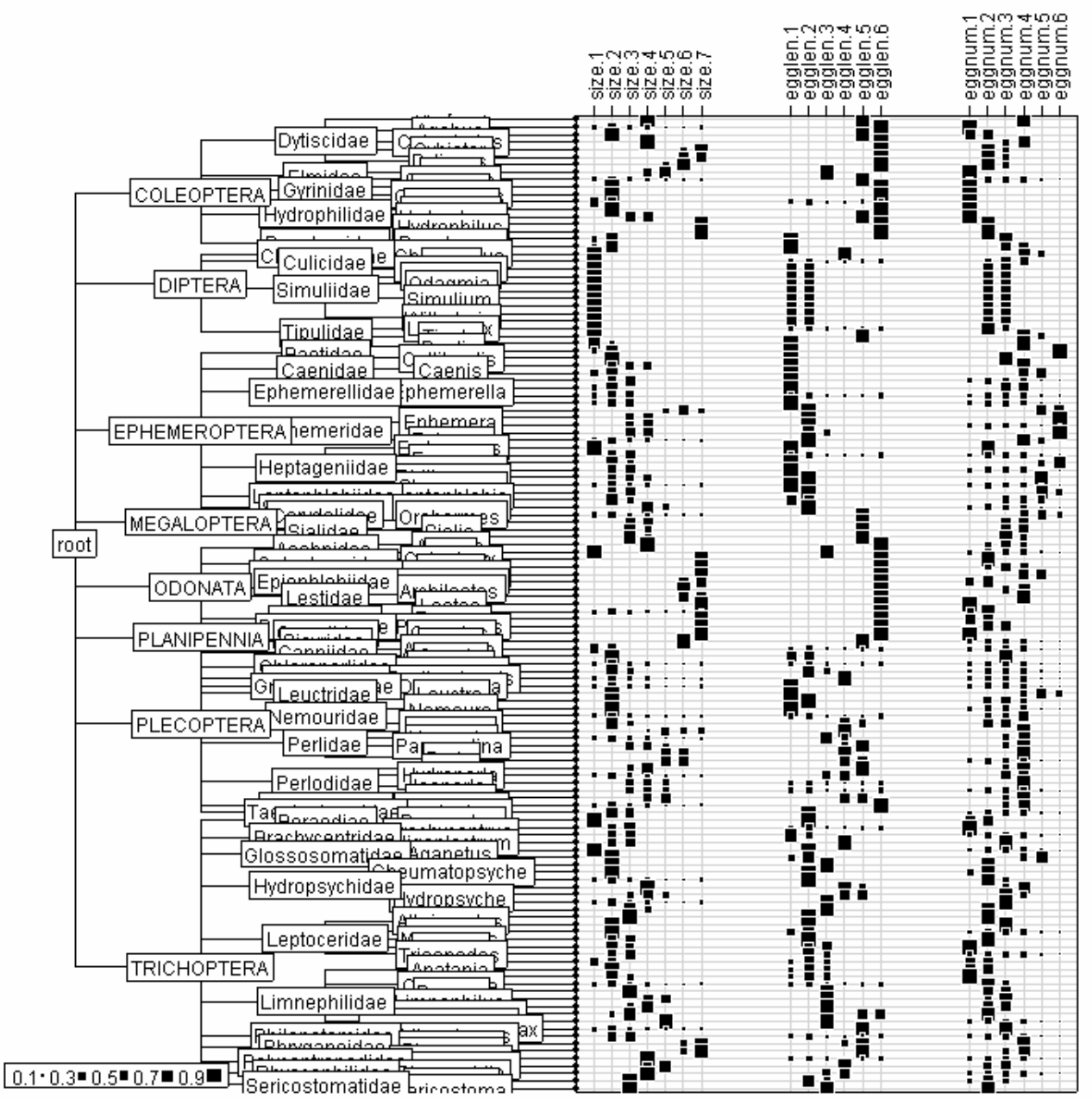


Tableau de trois traits biologiques en face de la taxonomie. On repère immédiatement le niveau ordre comme indicateur pour les odonates, les diptères, les coléoptères, ...

Ces représentations ont au moins la fonction de bien souligner la question qui se pose. Formellement, un trait est un tableau à plusieurs colonnes. Expérimentalement, c'est une variable qui donne une indication sur la stratégie d'une espèce. Statistiquement, un tableau de traits a de nombreuses colonnes réparties en blocs. Peut-on en faire une collection d'indicatrices, en vrac : il faudra nécessairement réduire leur nombre. La réduction doit-elle se faire par tableau, globalement ? Tout dépendra du type de redondance rencontrée dans ces enregistrements. Il faut savoir mesurer la corrélation entre deux traits et caractériser la part commune et la part spécifique dans la variabilité des traits.

3. L'ACP floue

Dans cette stratégie, on considère que chaque trait définit une ACP centrée par modalité. Cette analyse utilise le triplet basé sur une pondération arbitraire des taxa :

$$\mathbf{D} = Diag\left(\pi_1, \pi_2, \ldots, \pi_T\right)$$

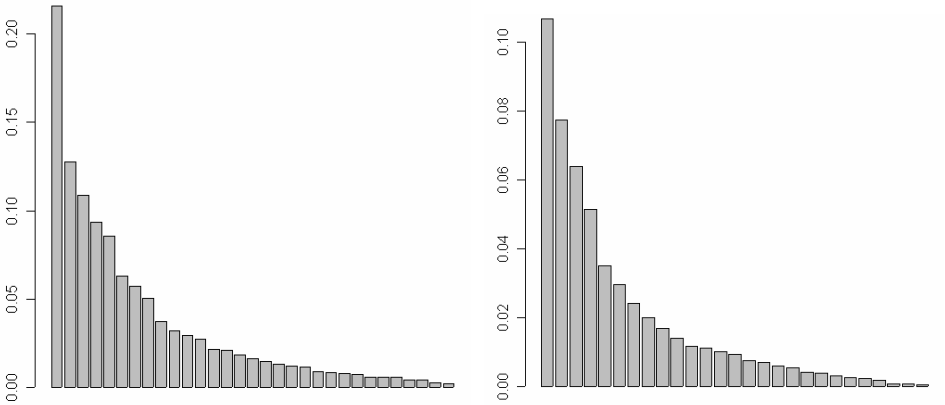
et le tableau centré :

$$\mathbf{X} = \left[f_{tm_v} - f_{*m_v}\right]_{\leq t \leq T, 1 \leq v \leq V, 1 \leq m_v \leq M_v}$$

On peut hésiter sur la pondération des modalités. Pour des raisons mathématiques, on prendra la pondération uniforme. L'option nouvelle **dudi.fpca** permet cette opération :

```
biol.fpca=dudi.fpca(biol.fuzzy)
Select the number of axes: 5

ecol.fpca=dudi.fpca(ecol.fuzzy)
Select the number of axes: 4
```



La première analyse ne mérite qu'un axe. On en prend ici 5 pour une illustration pédagogique. La seconde ne semble guère passionnante, on verra pourquoi.

Ces ACP, comme toutes les ACP centrées, donnent des combinaisons de variables de variance maximale. Les valeurs propres sont des inerties projetées mais la présence de blocs de variables modifient l'interprétation. On utilise le triplet très simple :

$$\left(\mathbf{X} = \left[f_{tm_v} - f_{*m_v}\right], Diag\left(\underbrace{\frac{1}{m_1}, \ldots, \frac{1}{m_1}}_{m_1}, \ldots, \underbrace{\frac{1}{m_V}, \ldots, \frac{1}{m_V}}_{m_V}\right), \frac{1}{T}\mathbf{I}_T\right)$$

```
dim(biol.fpca$tab)
[1] 131  41

biol.fpca$cw
     Fem.Size1      Fem.Size2      Fem.Size3      Fem.Size4      Fem.Size5
        0.1429         0.1429         0.1429         0.1429         0.1429
     Fem.Size6      Fem.Size7    ...
        0.1429         0.1429    ...
```

```
Clutch.struc1  Clutch.struc2  Clutch.struc3 Clutch.number1 Clutch.number2
       0.3333         0.3333         0.3333         0.3333         0.3333
Clutch.number3
       0.3333
```

```
unique(biol.fpca$lw) # 1/131
[1] 0.007634
```

```
biol.fpca
Duality diagramm
class: pca dudi
$call: dudi.pca(df = df, row.w = row.w, col.w = col.w, center = TRUE,
    scale = FALSE, scannf = scannf, nf = nf)

$nf: 5 axis-components saved
$rank: 31
eigen values: 0.2158 0.1275 0.1088 0.0935 0.08566 ...
  vector length mode    content
1 $cw    41     numeric column weights
2 $lw    131    numeric row weights
3 $eig   31     numeric eigen values

  data.frame nrow ncol content
1 $tab       131  41   modified array
2 $li        131  5    row coordinates
3 $l1        131  5    row normed scores
4 $co        41   5    column coordinates
5 $c1        41   5    column normed scores
other elements: cent norm blo indica FST inertia
```

L'objet contient, outre le standard d'une ACP, des informations spécifiques. La première de ces informations concerne la répartition de l'inertie totale (la variabilité biologique totale du groupe de taxa étudiés) entre les différentes descripteurs. La variance des colonnes est sommée par blocs (traits) et rapportée au total (édition en 1 pour 1000) dans `total` :

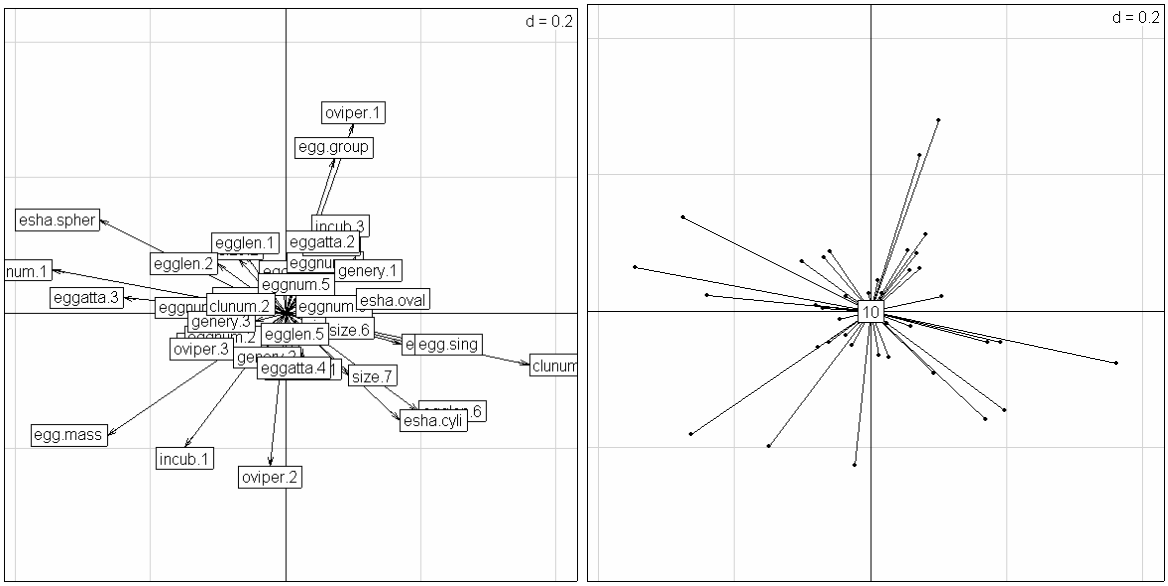
```
biol.fpca$inertia
              Ax1 Ax2 Ax3 Ax4 Ax5 total
Fem.Size       16  19  39  33  29    63
Egg.length     44  50  86  71   2    92
Egg.number      9  13  17  10   8    59
Generations    12  17  15 115 101    54
Oviposition    26 346 167 157 569   144
Incubation     52 155  11  74   1    91
Egg.shape     178 117 506 319  44   145
Egg.attach    105  29  39  25   8    94
Clutch.struc  173 227  94 175 179   139
Clutch.number 385  26  25  20  58   119
```

En terme de répartition de l'inertie entre les blocs de colonnes, la variation est de 1 à 3. les 4 variables (au sens biologique) les plus variables (au sens statistique) représente 60 % de la variabilité de la variabilité totale. En plaçant des poids de colonnes qui font une somme de 1 par blocs, on atténue le rôle du nombre de modalités.

```
round(1000*tapply(biol.fpca$co[,1]^2,biol.fpca$indica,mean)/biol.fpca$eig[1],0)
  1   2   3   4   5   6   7   8   9  10
 16  44   9  12  26  52 178 105 173 385
```

On a la même chose en terme de variance projetée sur chaque axe. La carte des colonnes regroupe toutes les modalités de toutes les variables :

```
s.arrow(biol.fpca$co)
```



Ceci n'a pas grand intérêt. Notons que, en calculant la covariance entre modalités, on fait simultanément le calcul des covariances entre modalités d'une même variable et entre modalités de variables différentes. La matrice de covariances **C** s'écrit ainsi :

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1V} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \mathbf{C}_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{V1} & \mathbf{C}_{V2} & \cdots & \mathbf{C}_{VV} \end{bmatrix}$$

$\mathbf{C}_{vv}$ est la matrice $m_v - m_v$ des covariances entre modalités de la même variable, alors que $\mathbf{C}_{vw}$ est la matrice $m_v - m_w$ des covariances entre les m_v modalités de la variable *v* et les m_w modalités de la variable *w*. Il est clair que les covariances entre modalités d'une même variable sont essentiellement artefactuelles : elles expriment simplement soit que deux modalités mériteraient de n'en faire qu'une (covariance positive entre modalités utilisées ensemble) soit que deux modalités ne sont pas utilisées simultanément (covariance négative inhérente à ce type de variables).

Par contre les covariances entre deux modalités de deux variables contiennent de l'information biologique. C'est l'association entre diverses modalités de différentes variables qui définit la stratégie globale d'une ou plusieurs espèces. Ce mélange intime de deux types de covariation, l'une artéfactuelle et l'autre signifiante laisse augurer de facteurs d'ACP consacrés soit à l'une, soit à l'autre, soit encore au mélange des deux. Il n'est donc pas possible de voir dans un tableau de variables floues un simple tableau de modalités juxtaposées.

Sur la carte des colonnes, la structure de ces données impose que les modalités ont des coordonnées centrées par bloc.

```
par(mfrow=c(4,3))
invisible(lapply(split(biol.fpca$co,biol.fpca$indica),
   s.arrow,xlim=c(-0.8,0.8)))
```

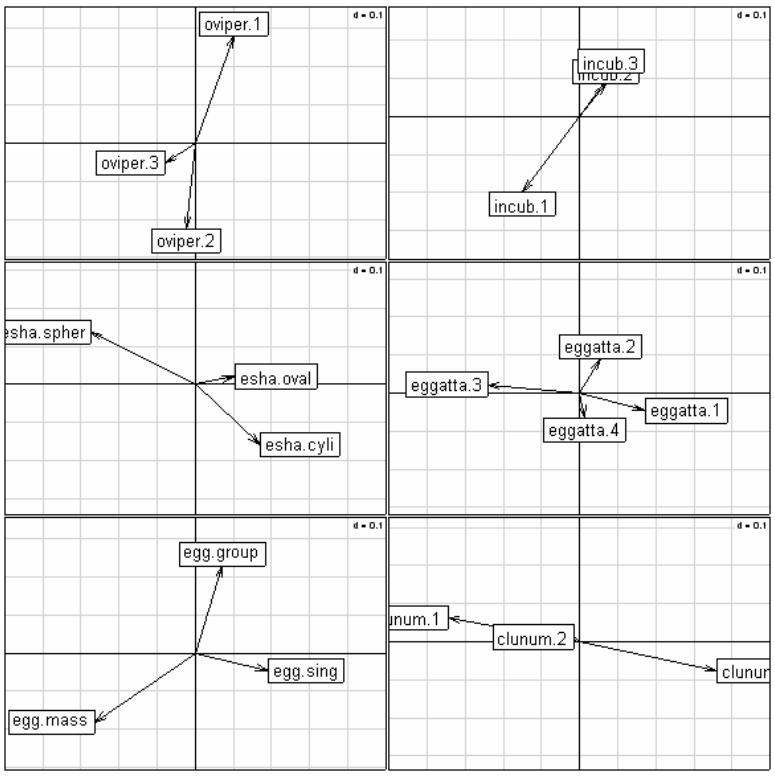
ou encore (ci-dessus, à droite) :

```
s.class(biol.fpca$co,biol.fpca$indica,cell=0)
```

ou encore :

```
par(mfrow=c(3,2))
```

```
for (k in 5:10) s.arrow(biol.fpca$co[ff==k,],xlim=c(-0.5,0.5),clab=2)
```



38% de la variance projetée sur l'axe 1 appartient à la dernière variable et l'ambiguïté est grande. De plus les coordonnées sur deux axes des modalités d'une même variable peuvent être covariantes, comme on le voit bien sur cette figure. Considérer les modalité comme variables n'est donc pas sans effet. Pour mesurer la redondance réelle ou l'indépendance réelle des traits biologiques, il est indispensable de considérer ces données dans le point de vue K-tableaux. Un tableau de variables floues doit être considéré comme un *K*-tableaux qu'on notera :

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_V]$$

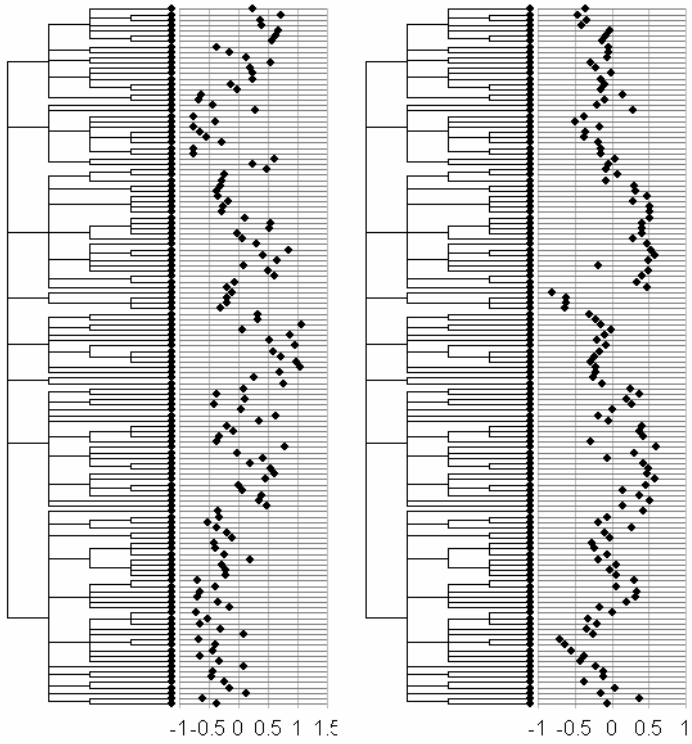
Mais on utilisera d'abord **dudi.fpca** qui prépare ce passage et donne en plus :

```
biol.fpca$FST
```

	inertia	max	FST
Fem.Size	0.07051	0.8214	0.08583
Egg.length	0.10262	0.8182	0.12542
Egg.number	0.06582	0.7878	0.08355
Generations	0.06061	0.2648	0.22887
Oviposition	0.16021	0.6355	0.25210
Incubation	0.10138	0.4624	0.21927
Egg.shape	0.16212	0.6336	0.25585
Egg.attach	0.10488	0.4794	0.21877
Clutch.struc	0.15503	0.5175	0.29956
Clutch.number	0.13219	0.5836	0.22650

On utilise ici le langage des tableaux de fréquences alléliques qui ont le même structure. L'inertie d'un trait est la somme des variances des modalités. Le maximum est l'inertie de la configuration dans laquelle chaque espèce a un profil disjonctif complet (il n'y a pas de variabilité intra spécifique). Le FST est le rapport des deux. La variabilité de 7 des 11 traits est alors très comparable. Les coordonnées des lignes ont de bonnes propriétés de résumés réduisant l'information bien que faiblement interprétable. On peut s'en servir de *marqueur* :

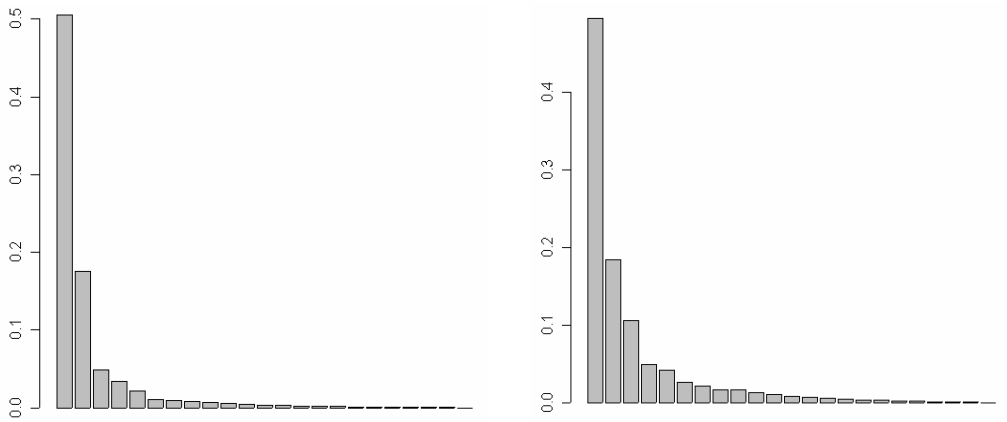
```
par(mfrow=c(1,2))
dotchart.phylog(phy,biol.fpca$li[,1])
dotchart.phylog(phy,ecol.fpca$li[,1])
```



Une bonne part de la variabilité organisée est d'origine taxonomique.

4. L'analyse des correspondances floues

Elle a été proposée pour traiter les tableaux de traits (Chevenet *et al.* 1994). Elle s'exécute par **`dudi.fca`**.

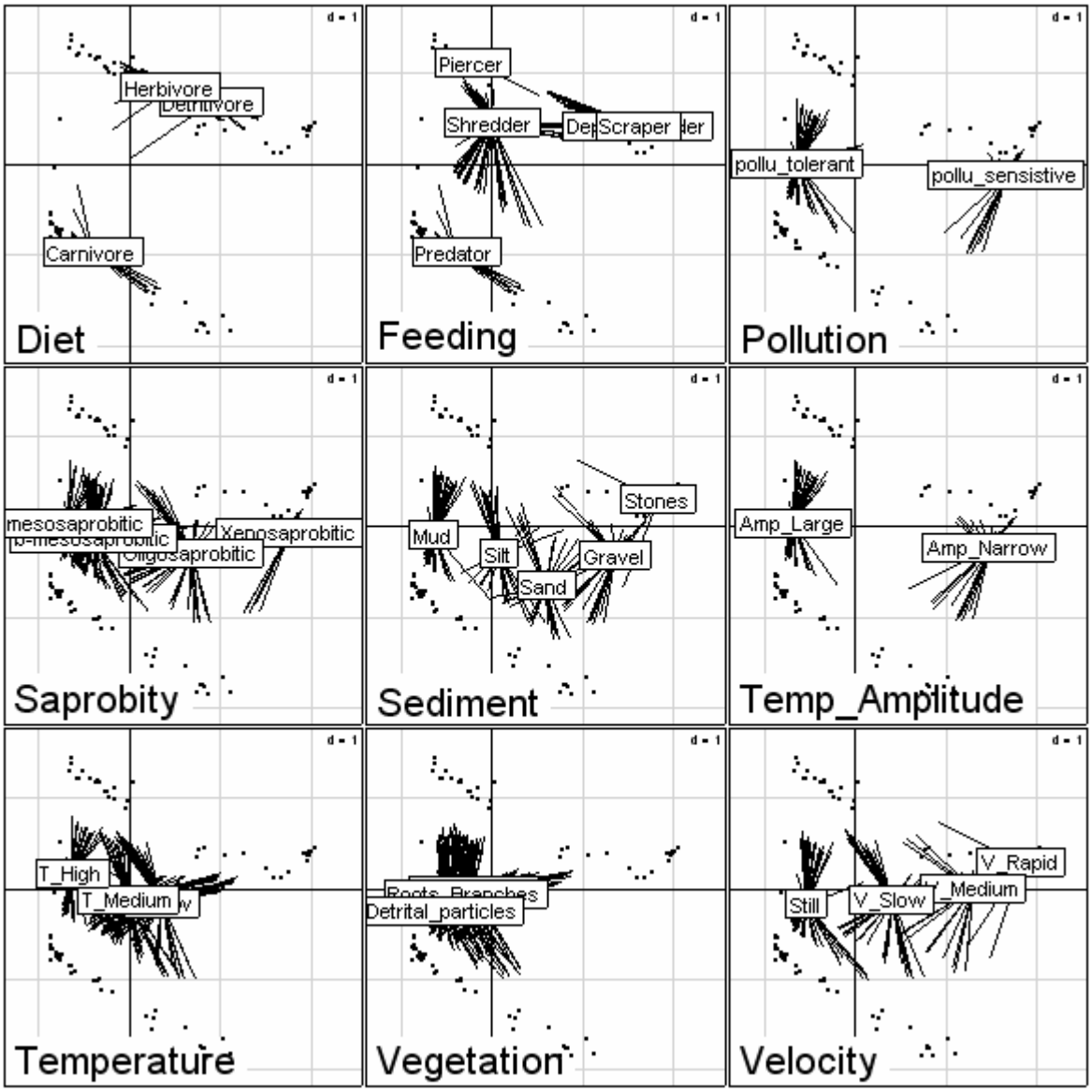



```
coleo.fpca=dudi.fpca(coleo.fuzzy)
Select the number of axes: 2
coleo.fcoa=dudi.fca(coleo.fuzzy)
Select the number of axes: 3

cor(coleo.fpca$li,coleo.fcoa$li)
         Axis1    Axis2    Axis3
Axis1  0.98530 -0.09682 -0.04738
Axis2 -0.09238 -0.98988  0.06781
```

Les coordonnées des lignes entre les deux analyses sont souvent très proches. C'est moins vrai pour les modalités qui sont soumises à des contraintes de centrage avec des pondérations très différentes.

```
par(mfrow = c(3,3))
indica <- factor(rep(names(coleo$col), coleo$col))
for (j in levels(indica))
   s.distri (coleo.fcoa$l1,
         coleo$tab[,which(indica==j)], clab = 1.5, sub = as.character(j),
         cell = 0, csta = 0.5, csub = 3,
         label = coleo$moda.names[which(indica == j)])
```



La logique de cette analyse est plus parlante dans le sens espèces -> modalités. On place sur chaque axe les espèces avec des scores centrés et réduits pour la pondération uniforme :

```
apply(coleo.fcoa$l1,2,mean)
       RS1        RS2        RS3
-1.520e-17  1.113e-16  1.352e-16

apply(coleo.fcoa$l1,2,function(x) mean(x*x))
RS1 RS2 RS3
  1   1   1
```

On observe le profil d'utilisation de chaque modalité entre les espèces, ce qui permet de positionner les modalité à la moyenne pondérée des espèces qui les utilisent. La variance de ces positions

moyennes est un rapport de corrélation et c'est la moyenne de ces rapports qui est optimisé (valeur propre). Ceci est explicite dans le tableau des rapports de corrélation voisin de celui d'une ACM :

```
coleo.fcoa$cr
                    RS1      RS2      RS3
Vegetation      0.07153 0.015964 0.020374
Sediment        0.81371 0.050609 0.385186
Velocity        0.66557 0.007663 0.165392
Pollution       0.78878 0.027167 0.011901
Saprobity       0.56916 0.041047 0.147240
Diet            0.17002 0.804106 0.022380
Feeding         0.46371 0.643580 0.192859
Temperature     0.10350 0.014883 0.000599
Temp_Amplitude 0.81314 0.051987 0.001737
```

Le facteur 1 est consacré à la position dans le gradient amont-aval, le second à l'alimentation, les deux faits biologiques n'étant pas complètement indépendant. La forme très particulière du tableau fait que l'AFC floue est une AFC (avec un complément pour l'interprétation) mais aussi une Aanalyse des correspondances internes (Cazes *et al.* 1988) :

```
coa1=dudi.coa(coleo.fuzzy)
Select the number of axes: 3

ww1=witwit.coa(coa1,110,coleo.fcoa$blo)
Select the number of axes: 3

round(coleo.fcoa$eig,4)
 [1] 0.4955 0.1841 0.1053 0.0485 0.0413 0.0261 0.0213 0.0161 0.0158 0.0130
[11] 0.0104 0.0081 0.0066 0.0054 0.0049 0.0035 0.0029 0.0023 0.0018 0.0012
[21] 0.0008 0.0003 0.0001

round(coa1$eig,4)
 [1] 0.4955 0.1841 0.1053 0.0485 0.0413 0.0261 0.0213 0.0161 0.0158 0.0130
[11] 0.0104 0.0081 0.0066 0.0054 0.0049 0.0035 0.0029 0.0023 0.0018 0.0012
[21] 0.0008 0.0003 0.0001

round(ww1$eig,4)
 [1] 0.4955 0.1841 0.1053 0.0485 0.0413 0.0261 0.0213 0.0161 0.0158 0.0130
[11] 0.0104 0.0081 0.0066 0.0054 0.0049 0.0035 0.0029 0.0023 0.0018 0.0012
[21] 0.0008 0.0003 0.0001 0.0000
```

Les trois analyses ont exactement les mêmes valeurs propres, les mêmes coordonnées des lignes et les mêmes coordonnées des colonnes. Ces trois analyses sont strictement identiques. Ceci vient simplement du fait que le tableau flou contient des distributions de fréquences par bloc et que le recentrage par bloc explicite dans l'analyse des l'analyse des correspondances internes est implicite dans l'analyse des correspondances floues. Ce point est sans grande importance. Il permet simplement de refaire l'opération de décomposition des valeurs propres par sous-tableaux. On peut donc interpréter l'analyse comme une AFC intra-bloc de colonnes :

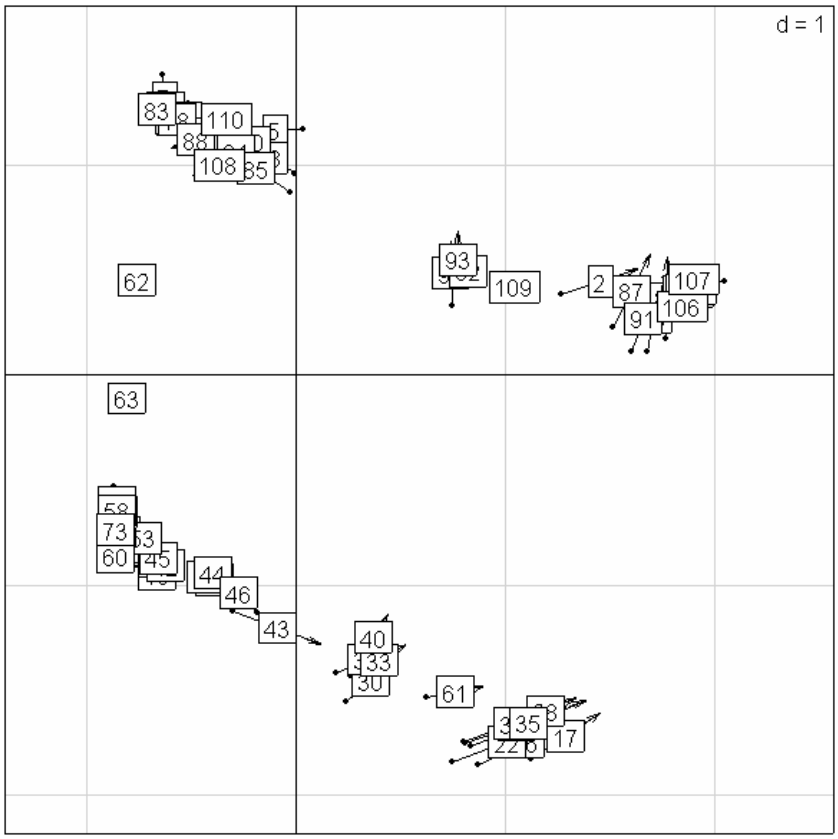
```
summary(ww1)
Eigen value decomposition among column blocks
               Comp1  Comp2  Comp3  weights
Vegetation     0.0715 0.016  0.0204 0.1111
Sediment       0.8137 0.0506 0.3852 0.1111
Velocity       0.6656 0.0077 0.1654 0.1111
Pollution      0.7888 0.0272 0.0119 0.1111
Saprobity      0.5692 0.041  0.1472 0.1111
Diet           0.17   0.8041 0.0224 0.1111
Feeding        0.4637 0.6436 0.1929 0.1111
Temperature    0.1035 0.0149 6e-04  0.1111
Temp_Amplitude 0.8131 0.052  0.0017 0.1111
mean           0.4955 0.1841 0.1053
```

Ce résultat est plus nuancé que celui de l'ACP qui donne :

```
coleo.fpca$inertia
                Ax1 Ax2 total
Vegetation        8   9    30
Sediment         89   3    82
Velocity         81   2    76
Pollution       312   8   218
Saprobity        60   8    67
Diet             25 681   163
Feeding          24 263    90
Temperature      17   3    16
Temp_Amplitude  384  24   257
```

Il n'y a pas de contradiction mais des modes de fonctionnement très différents. Rappelons que les deux analyse font la même typologie des espèces. Pour s'en convaincre :

```
wa=coleo.fcoa$l1[,1:2]
wb=cbind.data.frame(coleo.fpca$l1[,1],-coleo.fpca$l1[,2])
names(wb)=names(wa)
s.match(wa,wb)
```



Carte d'ACP et carte d'AFC superposées !

L'AFC utilise l'averaging espèces -> modalités alors que l'ACP fait des combinaisons linéaires qui ne sont pas des moyennes et enregistre de ce fait les différences de variance qui perturbe l'expression numérique des résultats. Il faudra la couche multi-tableaux pour éliminer cette perturbation.

Les principales propriétés de l'AFC floue sont (i) la symétrie lignes-colonnes, (ii) la métrique du Khi2 et (iii) la faible réduction de dimension.

(i) La symétrie lignes-colonnes est sensible sur la figure ci-dessus où l'analyse apparaît comme une AFC simultanée de chacun des tableaux-variables utilisant les mêmes scores-lignes.

Une espèce est vue comme une distribution de fréquence dans son usage d'une variable biologique et chaque modalité de chaque variable est vue comme une distribution de fréquence de son utilisation par les espèces. Le double averaging implicite conduit à la représentation simultanée des espèces et des modalités. Pour une description détaillée de la typologie des stratégies biologiques, cette analyse est indiscutablement utile.

(ii) Elle utilise cependant implicitement la métrique du Khi2 qui est contestable. Les carrés des différences entre deux profils sont pondérés par l'inverse de l'utilisation moyenne d'une modalité. La distance entre les taxons t_1 et t_2 s'écrit :

$$d^2_{t_1,t_2} = \sum_{v=1}^{V} \sum_{m_v=1}^{M_v} \frac{\left(f_{t_1 m_v} - f_{t_2 m_v}\right)^2}{f_{* m_v}}$$

Une différence d'usage sur une modalité rare est ainsi fortement amplifiée, ce qui vue le type de mesure est loin de s'imposer. On préférerait simplement :

$$d^2_{t_1,t_2} = \sum_{v=1}^{V} \sum_{m_v=1}^{M_v} \left(f_{t_1 m_v} - f_{t_2 m_v}\right)^2$$

Mais l'analyse simple (ACP) associée à cette métrique est plutôt défavorable.

(iii) Le nombre de valeurs propres à prendre en compte est enfin élevé. Il est difficile de dire si cela vient vraiment des données et que les combinaisons stratégiques sont effectivement variées et complexes ou si cela vient du fait que les AFC multiples sont de mauvaises analyses d'inertie dans l'espace des modalités. L'AFC multiple est d'abord une analyse canonique (on y reviendra en détail) mais cette propriété est perdue dans la variante sur variables floues.

Ceci pose la question des méthodes *K*-tableaux dans l'usage des variables floues. L'analyse qui suit vise à savoir si cette introduction qui est relativement difficile au plan théorique en vaut la peine.

5. Traits biologiques et méthodes K-tableaux

Pour utiliser ces méthodes, on a besoin d'objets de la classe **ktab**. Les manipulations préliminaires comme le centrage et la normalisation, le double centrage, ... sont directement assurées par l'utilisateur et non incorporé aux fonctions de cette classe (il faudrait une quantité vertigineuse de cas particuliers). Ici nous dirons simplement : chaque trait est un tableau d'ACP, il suffit de partir du tableau des fréquences centrées. On pourrait avoir les versions AFC, mais cela obscurcit le débat. Pour faire un K-tableaux :

```
biol.ktab <- ktab.data.frame (biol.fpca$tab,biol.fpca$blo,w.col=biol.fpca$cw)
ecol.ktab <- ktab.data.frame (ecol.fpca$tab,ecol.fpca$blo,w.col=ecol.fpca$cw)
coleo.ktab <- ktab.data.frame (coleo.fpca$tab,coleo.fpca$blo,w.col=coleo.fpca$cw)

biol.ktab
class: ktab

tab number:   10
```

Il y a 10 tableaux dans ce K–tableaux. Leur nom et leurs dimensions sont :

```
   data.frame    nrow ncol
1  Fem.Size      131  7
```

```
2  Egg.length    131  6
3  Egg.number    131  6
4  Generations   131  3
5  Oviposition   131  3
6  Incubation    131  3
7  Egg.shape     131  3
8  Egg.attach    131  4
9  Clutch.struc  131  3
10 Clutch.number 131  3
```

Les tableaux sont appariés par les lignes et on a un seul vecteur de poids des lignes (**lw**) et un seul vecteur des poids de colonnes (**cw**) qui est, de fait, la collection des poids des colonnes de plusieurs analyses séparées. On n'a pas précisé de poids des tableaux qui pourrait jouer un rôle dans d'autres méthodes :

```
   vector length mode    content
11 $lw    131    numeric row weigths
12 $cw    41     numeric column weights
13 $blo   10     numeric column numbers
14 $tabw  0      NULL    array weights
```

On a préparé des tableaux de facteurs pour ranger des coordonnées de lignes par tableaux (131*10 lignes prévues), des coordonnées des colonnes par tableaux (41 colonnes prévues) et des valeurs multiples par tableaux (typiquement les 4 premiers axes principaux, 10*4 lignes prévues) :

```
   data.frame nrow ncol content
15 $TL        1310 2    Factors Table number Line number
16 $TC        41   2    Factors Table number Col number
17 $T4        40   2    Factors Table number 1234
```

L'ordre d'appel de la création du K-tableaux est conservé :

```
18 $call: ktab.data.frame(df = biol.fpca$tab, blocks = biol.fpca$blo, w.col =
biol.fpca$cw)
```

Les noms des tableaux sont les **tab.names** du K-tableaux et les noms des variables des tableaux sont les **col.names** du K-tableaux :

```
names :
Fem.Size : size.1 size.2 size.3 size.4 size.5 size.6 size.7
Egg.length : egglen.1 egglen.2 egglen.3 egglen.4 egglen.5 egglen.6
...
Clutch.number : clunum.1 clunum.2 clunum.3
```

Les poids des colonnes par tableau ont été importés de l'analyse simple (**cw**) :

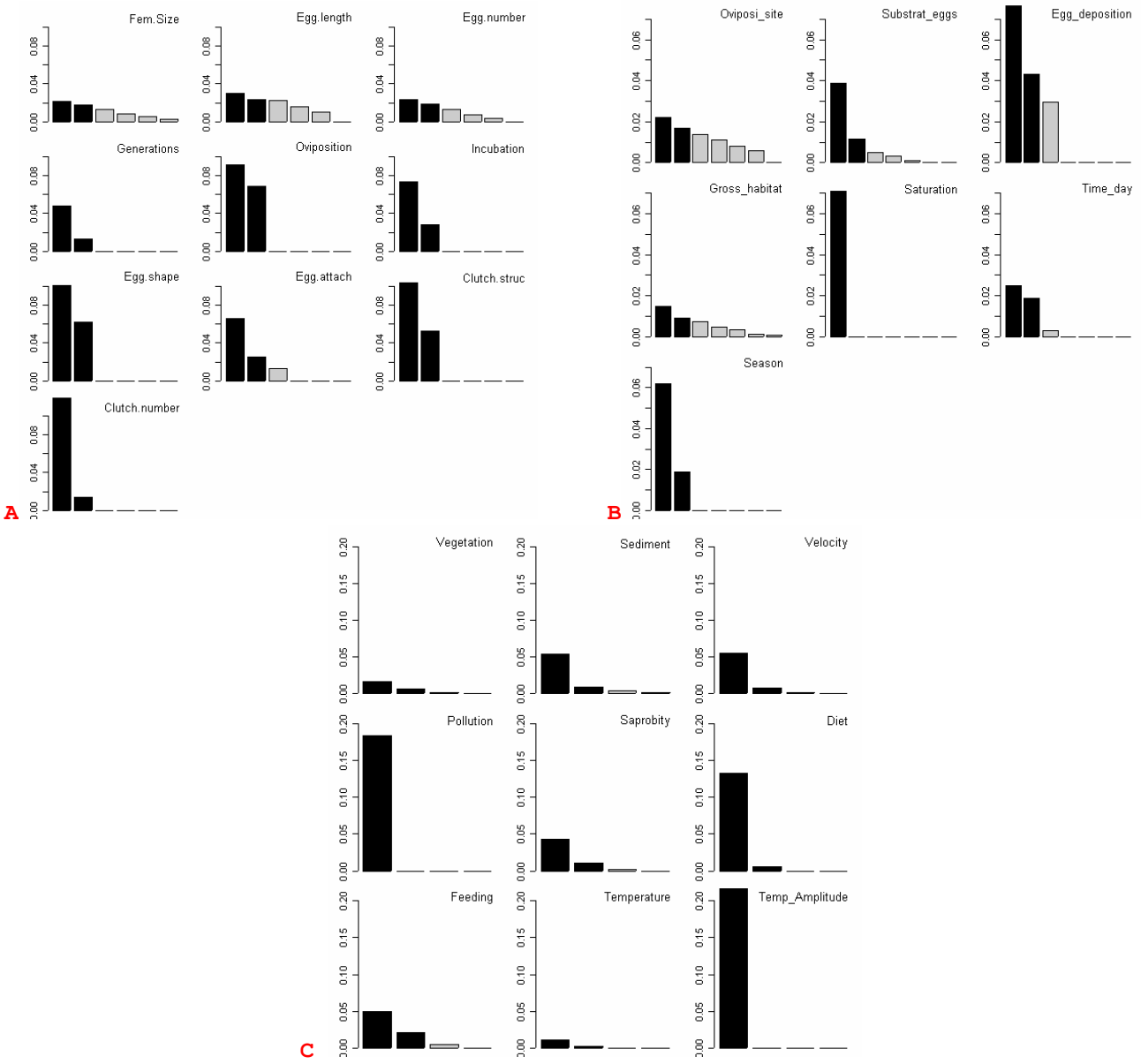
```
Col weigths :
Fem.Size : 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571
Egg.length : 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
...
Clutch.number : 0.3333333 0.3333333 0.3333333
```

Les poids des lignes communs aux différents tableaux ont été importés de l'analyse simple (**lw**) :

```
Row weigths :
0.007633588 0.007633588 0.007633588 0.007633588 0.007633588 0.007633588 0.007633588
...
0.007633588 0.007633588 0.007633588 0.007633588 0.007633588
```

On a donc tous les éléments pour faire 10 analyses séparées :

```
plot(sepan(biol.ktab))
plot(sepan(ecol.ktab))
plot(sepan(coleo.ktab))
```



Graphes des valeurs propres des analyses séparées de 3 K-tableaux.

On voit immédiatement que les traits biologiques ont des identités très variables. Certains sont des mélanges sans structures de 5 modalités, d'autres des variables binaires, d'autres définissent des typologies franches à 2 ou 3 dimensions. L'analyse simple du troisième fonctionne bien parce que, grossièrement, chaque tableau apporte une seule dimension. Par contre, dans les deux premiers le mélange de types très différents est patent. Cette opération est essentielle. Elle correspond à la description univariée qui doit précéder toute analyse multivariée, les variables sont devenues des tableaux et le multivarié est devenu du multi-tableaux. Comme il se doit, les articles publiés sur le sujet ont fait l'impasse sur ces éléments de base, tout investissement méthodologique étant généralement considéré comme du temps perdu.

On notera que **`biol.fuzzy`** est un data.frame :

```
names(biol.fuzzy)
 [1] "size.1"     "size.2"     "size.3"     "size.4"     "size.5"
 [6] "size.6"     "size.7"     "egglen.1"   "egglen.2"   "egglen.3"
[11] "egglen.4"   "egglen.5"   "egglen.6"   "eggnum.1"   "eggnum.2"
[16] "eggnum.3"   "eggnum.4"   "eggnum.5"   "eggnum.6"   "genery.1"
[21] "genery.2"   "genery.3"   "oviper.1"   "oviper.2"   "oviper.3"
[26] "incub.1"    "incub.2"    "incub.3"    "esha.spher" "esha.oval"
```

```
[31] "esha.cyli"  "eggatta.1"  "eggatta.2"  "eggatta.3"  "eggatta.4"
[36] "egg.sing"   "egg.group"  "egg.mass"   "clunum.1"   "clunum.2"
[41] "clunum.3"
```

qui a des attributs supplémentaires :

```
attributes(biol.fuzzy)
$names
 [1] "size.1"     "size.2"     "size.3"     "size.4"     "size.5"     ...
$row.names
  [1] "E1"   "E2"   "E3"   "E4"   "E5"   "E6"   "E7"   "E8"   "E9"   "E10" ...
$class
[1] "data.frame"

$col.blocks
     Fem.Size    Egg.length    Egg.number   Generations   Oviposition
            7             6             6             3             3
...

$row.w
  [1] 0.007633588 0.007633588 0.007633588 0.007633588 0.007633588 0.007633588 ...
$col.freq
 [1] 0.17529240 0.28055556 0.17309942 0.13706140 0.05899123 0.05921053 ...
$col.num
 [1] 1  1  1  1  1  1  1  2  2  2  2  2  2  3  3  3  3  3  3  4  4  4  5  5  5
[26] 6  6  6  7  7  7  8  8  8  8  9  9  9  10 10 10
Levels: 1 2 3 4 5 6 7 8 9 10
```

biol.fcpa est un **dudi** qui a des composantes supplémentaires :

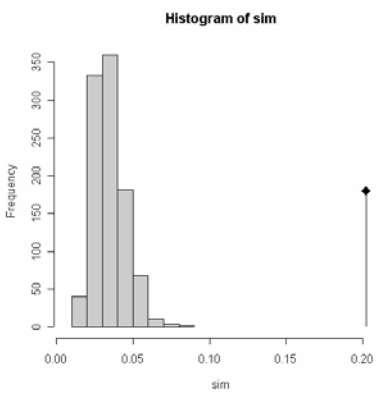
```
names(biol.fpca)
 [1] "tab"      "cw"       "lw"       "eig"      "rank"     "nf"       "c1"
 [8] "l1"       "co"       "li"       "call"     "cent"     "norm"     "blo"
[15] "indica"   "FST"      "inertia"
```

Enfin **biol.ktab** est un **ktab**, donc une liste de **data.frame** avec des composantes supplémentaires :

```
names(biol.ktab)
 [1] "Fem.Size"      "Egg.length"    "Egg.number"    "Generations"
 [5] "Oviposition"   "Incubation"    "Egg.shape"     "Egg.attach"
 [9] "Clutch.struc"  "Clutch.number" "lw"            "cw"
[13] "blo"           "TL"            "TC"            "T4"
[17] "call"
```

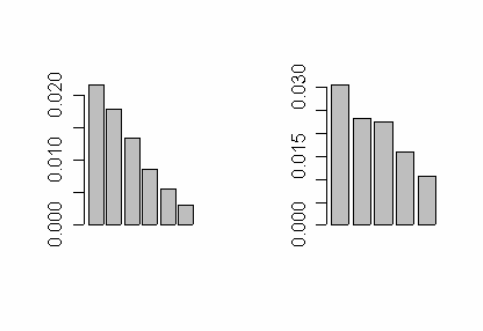
On peut ainsi retrouver toute l'information utile. Faisons, par exemple, à titre pédagogique l'analyse de co-inertie des deux premiers tableaux de **biol.ktab** :

```
dd1 <- as.dudi(df = biol.ktab[[1]], col.w = biol.ktab$cw[1:7],
   row.w = biol.ktab$lw, scannf=FALSE, nf=2, call = match.call(), type = "")
dd2 <- as.dudi(df = biol.ktab[[2]], col.w = biol.ktab$cw[8:13],
   row.w = biol.ktab$lw, scannf=FALSE, nf=2, call = match.call(), type = "")
coi1 <- coinertia(dd1,dd2,scannf=FALSE)
plot(randtest(coi1))
```



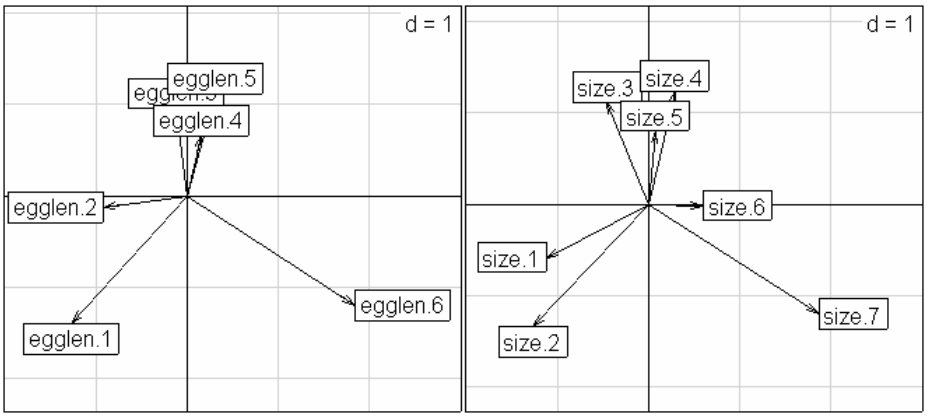
On voit alors clairement que chacun des traits n'autorisent pas de réduction de dimension :

```
par(mfrow=c(1,2))
barplot(dd1$eig)
barplot(dd2$eig)
```



mais ont une co-inertie très significative statistiquement et sa signification biologique est très simple :

```
s.arrow(coi1$l1,xlim=c(-2,3))
s.arrow(coi1$c1,xlim=c(-2,3))
```




```
coi1$RV
[1] 0.2021185
```

L'association entre les deux variables est mesurée par le coefficient RV (Escoufier 1973, 1977, 1980, 1985, 1987), rapport de la co-structure (mesurée par la covariance vectorielle ou co-inertie totale) à la structure (mesurée par le produit des variances vectorielles expression de la structure totale) :

$$RV(\mathbf{X},\mathbf{Y}) = \frac{CVV(\mathbf{X},\mathbf{Y})}{VV(\mathbf{X})VV(\mathbf{Y})}$$

```
sum(coi1$eig)/sqrt(sum(dd1$eig^2))/sqrt(sum(dd2$eig^2))
[1] 0.2021185
```

La méthode STATIS calcule les RV de chacun des couples de tableaux, donne une représentation euclidienne des tableaux équivalente d'un cercle des corrélations d'ACP normée, en un mot fait de l'ACP de tableaux, ce qui est exactement le problème.

```
biol.statis <- statis(biol.ktab, scannf=F)
ecol.statis <- statis(ecol.ktab, scannf=F)
coleo.statis <- statis(coleo.ktab, scannf=F)
```

Pour récupérer la matrice des RV :

```
round(biol.statis$RV,4)
```

	Fem.Size	Egg.length	Egg.number	Generations	Oviposition	Incubation
Fem.Size	1.0000	**0.2021**	0.0708	0.1473	0.0527	0.0250
Egg.length	**0.2021**	1.0000	0.1015	0.0451	0.0432	0.0311
Egg.number	0.0708	0.1015	1.0000	0.0355	0.0242	0.0283
Generations	0.1473	0.0451	0.0355	1.0000	0.0831	0.0449
Oviposition	0.0527	0.0432	0.0242	0.0831	1.0000	0.0920
Incubation	0.0250	0.0311	0.0283	0.0449	0.0920	1.0000
Egg.shape	0.1330	0.2637	0.0523	0.0360	0.0149	0.0059
Egg.attach	0.0897	0.1035	0.0427	0.0091	0.0070	0.0671
Clutch.struc	0.0594	0.0623	0.0335	0.0250	0.0329	0.0291
Clutch.number	0.0351	0.0972	0.0388	0.0044	0.0053	0.0434

	Egg.shape	Egg.attach	Clutch.struc	Clutch.number
Fem.Size	0.1330	0.0897	0.0594	0.0351
Egg.length	0.2637	0.1035	0.0623	0.0972
Egg.number	0.0523	0.0427	0.0335	0.0388
Generations	0.0360	0.0091	0.0250	0.0044
Oviposition	0.0149	0.0070	0.0329	0.0053
Incubation	0.0059	0.0671	0.0291	0.0434
Egg.shape	1.0000	0.1060	0.0258	0.1612
Egg.attach	0.1060	1.0000	0.0905	0.1044
Clutch.struc	0.0258	0.0905	1.0000	0.1460
Clutch.number	0.1612	0.1044	0.1460	1.0000

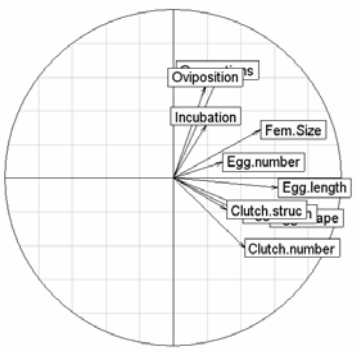
```
round(coleo.statis$RV,4)
```

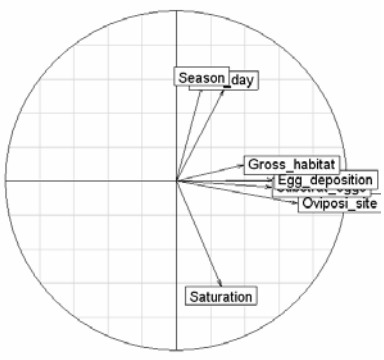
	Vegetation	Sediment	Velocity	Pollution	Saprobity	Diet	Feeding
Vegetation	1.0000	0.1594	0.2411	0.1994	0.1284	0.1526	0.2000
Sediment	0.1594	1.0000	0.7553	0.6042	0.6529	0.0600	0.1699
Velocity	0.2411	0.7553	1.0000	0.5142	0.5825	0.1076	0.2263
Pollution	0.1994	0.6042	0.5142	1.0000	0.6038	0.0384	0.1524
Saprobity	0.1284	0.6529	0.5825	0.6038	1.0000	0.0204	0.1545
Diet	0.1526	0.0600	0.1076	0.0384	0.0204	1.0000	**0.8938**
Feeding	0.2000	0.1699	0.2263	0.1524	0.1545	0.8938	1.0000
Temperature	0.1084	0.6116	0.4976	0.6115	0.5546	0.0171	0.1280
Temp_Amplitude	0.1265	0.7254	0.5791	0.6589	0.5340	0.0308	0.1567

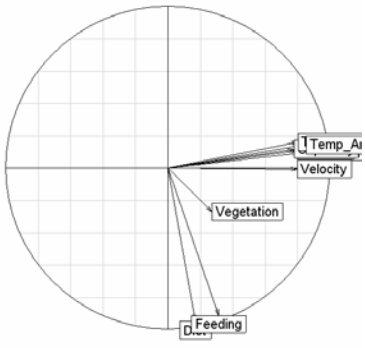
	Temperature	Temp_Amplitude
Vegetation	0.1084	0.1265
Sediment	0.6116	0.7254
Velocity	0.4976	0.5791
Pollution	0.6115	0.6589
Saprobity	0.5546	0.5340
Diet	0.0171	0.0308
Feeding	0.1280	0.1567
Temperature	1.0000	0.7752
Temp_Amplitude	0.7752	1.0000

Ls deux cas sont très différents. Dans le premier la redondance est faible, dans le second cas elle est omniprésente. Des images de synthèse de cette redondance sont très simple à obtenir :

```
s.corcircle(biol.statis$RV.coo,clab=1.25)
s.corcircle(ecol.statis$RV.coo,clab=1.25)
s.corcircle(coleo.statis$RV.coo,clab=1.25)
```







A droite, il y a deux paquets de traits redondants. Le premier donne la position du taxon dans le gradient amont-aval de différents point de vue, tous équivalents. Le second caractérise la stratégie trophique. Cette situation génère deux axes marqués dans les analyses simples du tableau complet.

A gauche, l'image euclidienne d'un ensemble de traits biologiques largement dissociés les uns des autres qui renvoie à une analyse simple peu interprétable. Au centre, un cas intermédiaire. Ces figures ne sont pas des cercles de corrélation mais s'en rapprochent. La différence essentielle est qu'un RV est toujours positif, on ne sait pas ce qu'est une corrélation négative entre typologies. Les RV sont des cosinus entre opérateurs mais les angles entre ces objets sont toujours aigus (on dit le cône des opérateurs). La ressemblance porte sur la longueur des vecteurs. Les vecteurs courts indique des directions orthogonales au plan (à droite, `vegetation`, à gauche `incubation`), donc des tableaux peu redondants avec les autres.

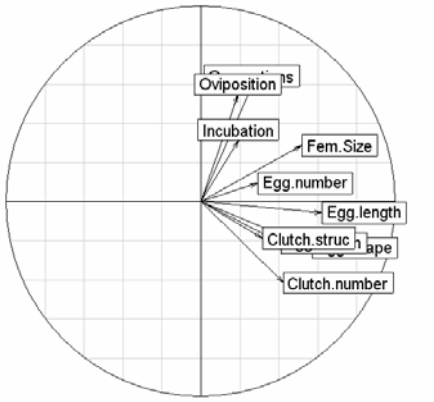
Ce qui est frappant dans les données de B. Statzner et ses collègues est la complexité de la redondance des traits biologiques. C'est un résultat inattendu, qui vient d'une recherche précise.

```
f1 <- function(x,y) {
    a <- biol.ktab[[x]]
    b <- biol.ktab[[y]]
    w <- procuste.randtest(a,b)$pvalue
    1000*w
}

res <- matrix(NA,10,10)
dimnames(res) <- list(names(biol.ktab)[1:10],1:10)
for (k in 1:10) {
    res[k,-k] <- unlist(lapply((1:10)[-k], f1, y=k))
}
```

```
res
```

		1	2	3	4	5	6	7	8	9	10
1	Fem.Size	NA	1	3	1	24	287	1	4	3	63
2	Egg.length	1	NA	4	15	48	116	1	1	1	1
3	Egg.number	2	1	NA	53	261	272	9	11	84	66
4	Generations	1	9	41	NA	1	23	16	660	148	732
5	Oviposition	24	43	256	1	NA	4	591	741	206	636
6	Incubation	286	96	300	32	2	NA	717	10	139	29
7	Egg.shape	1	1	7	14	563	703	NA	1	77	1
8	Egg.attach	5	1	8	620	766	12	1	NA	1	1
9	Clutch.struc	2	3	76	133	168	152	75	1	NA	1
10	Clutch.number	71	1	66	733	645	28	1	1	1	NA

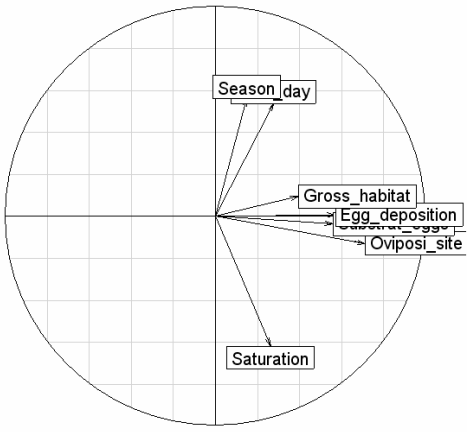


On a appliqué ici le test de Jackson (Jackson 1995) lié au couplage par rotation procuste. Les relations entre test RV (Heo & Gabriel 1998) et PROTEST, entre co-inertie et rotation procuste sont étroites et explicitées dans Dray *et al.* (2003). Sur 45 tests indépendants, la moitié peuvent être considérés comme non significatifs, l'autre moitié comme très significatifs. La représentation de `statis` rend globalement compte de cette complexité.

```
f1 <- function(x,y) {
    a <- ecol.ktab[[x]] ; b <- ecol.ktab[[y]]
    w <- 1000*procuste.randtest(a,b)$pvalue
}
res <- matrix(NA,7,7)
dimnames(res) <- list(names(ecol.ktab)[1:7],1:7)
for (k in 1:7) res[k,-k] <- unlist(
     lapply((1:7)[-k], f1, y=k))
```

```
res
```

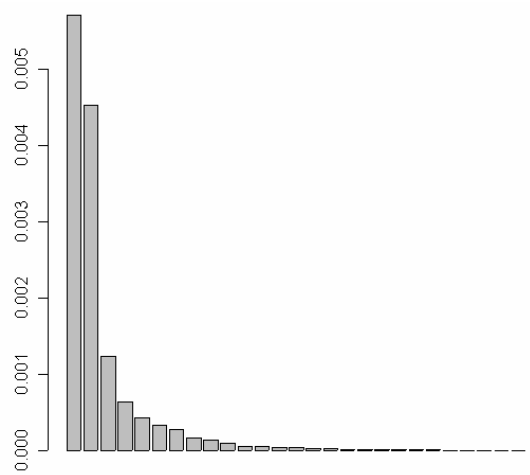
	1	2	3	4	5	6	7
Oviposi_site	NA	1	1	1	1	11	65
Substrat_eggs	1	NA	1	1	182	16	872
Egg_deposition	1	1	NA	250	8	8	37



```
Gross_habitat   1   1 260  NA 502  80 617
Saturation      1 207   9 517  NA 909 839
Time_day        8  22   5  71 902  NA 114
Season         52 863  34 651 836 154  NA
```

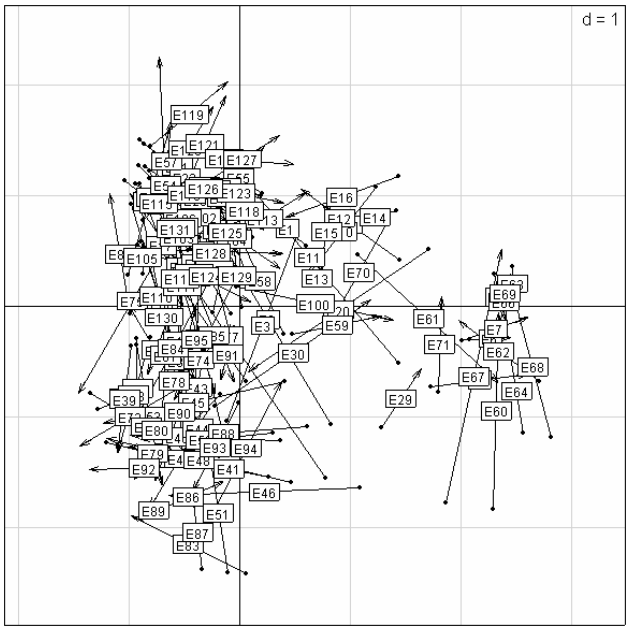
On retrouve des variables très liées définissant un compromis et d'autres quasiment indépendantes. Ces dernières seront effacées des bilans (elles ne sont pas redondantes) sans qu'on ait prouvé pour autant qu'elles sont sans signification biologique. La question posée par ces traits est donc beaucoup plus compliquée qu'on a voulu le laisser croire. Or, ce qui pousse à faire l'impasse sur ce type d'examen, c'est le fait que la co-inertie des deux blocs à plat fonctionne bien.

```
coitot=coinertia (biol.fpca, ecol.fpca)
Select the number of axes: 3
```



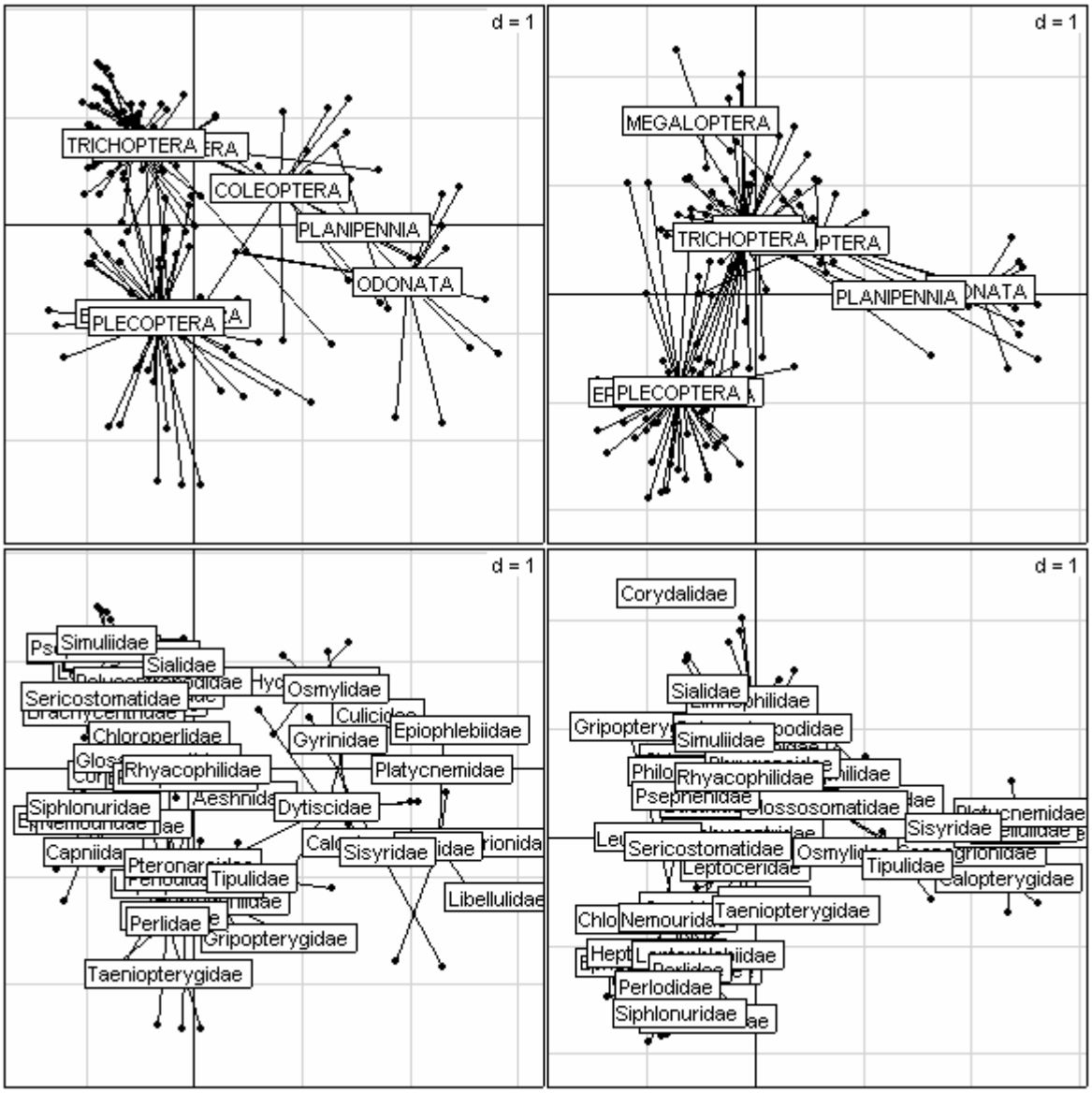
Toutes les questions semblent avoir disparu.

```
s.match(coitot$mX,coitot$mY,clab=0.75)
```



Oui, mais :

```
s.class(coitot$mX,bsetal97$taxo$ord,cell=0)
s.class(coitot$mY,bsetal97$taxo$ord,cell=0)
s.class(coitot$mX,bsetal97$taxo$fam,cell=0)
s.class(coitot$mY,bsetal97$taxo$fam,cell=0)
```



Plan de co-inertie vu par le biais de la taxonomie (à gauche traits biologiques, à droite traits écologiques)

Des analyses de variance montre que 80 % de la variabilité des codes de synthèse est inter familles. Par exemple :

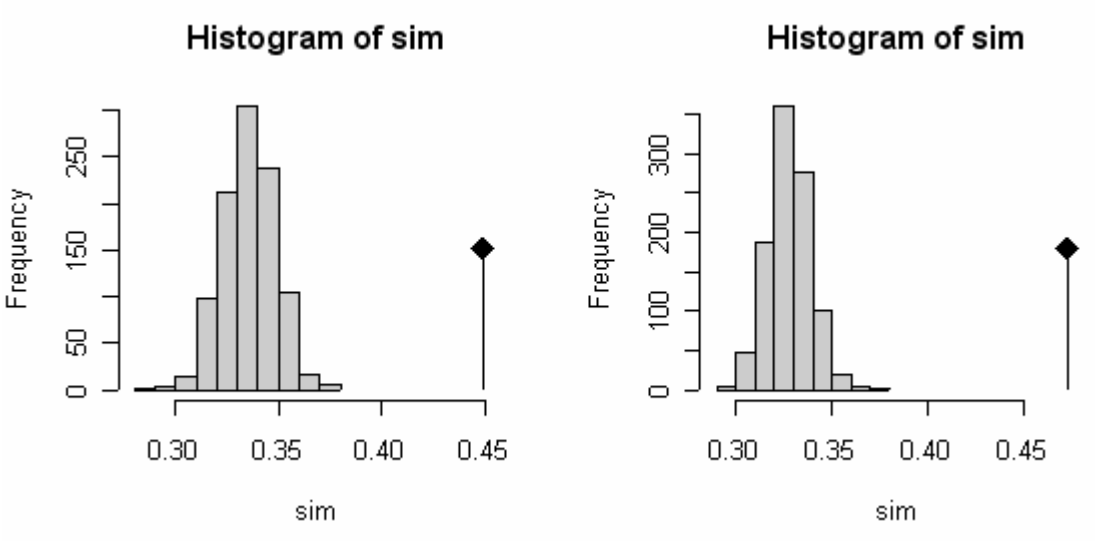
```
anova(lm(coitot$mX[,1]~ord+fam+gen,data=bsetal97$taxo))
Analysis of Variance Table

Response: coitot$mX[, 1]
          Df Sum Sq Mean Sq F value Pr(>F)
ord        7   80.9    11.6   95.72 <2e-16
fam       40   35.7     0.9    7.38  1e-08
gen       47   10.1     0.2    1.79  0.036
Residuals 36    4.3     0.1
```

La coinertie des deux tableaux a donc retrouvé la structure taxonomique qui sous-tend chacun d'eux (au moins comme première approximation de la phylogénie de ces groupes). La question de Felsenstein (1985) à l'origine des méthodes phylogénétiques est donc ici en jeu. Mais y a-t-il encore un lien derrière cette super contrainte ? La question n'est déjà pas simple avec deux variables, elle ne le sera pas avec deux K-tableaux !

On fait un essai très simple. La coinertie prend les tableaux avec un centrage global toutes espèces confondues. On peut déjà les recentrer par familles pour être sûr d'enlever des effets importants :

```
w1=within (biol.fpca,bsetal97$taxo$fam,scan=F)$tab
w2=within (ecol.fpca,bsetal97$taxo$fam,scan=F)$tab
plot(procuste.randtest(w1,w2))
plot(procuste.randtest(biol.fpca$tab,ecol.fpca$tab))
```



On a presque autant de signification avec le centrage par famille que sans, mais les permutations devrait se faire à l'intérieur des blocs pour être sûr de l'effet local. Une autre fonction ...

6. Conclusion

On notera pour conclure cette première approche les points suivants.

1) L'AFC floue est une méthode de synthèse des tableaux de traits biologiques qui a l'avantage de la simplicité. C'est un premier pas dont la propriété principale d'AFC internes implicite dans la direction de la statistique multi-tableau.

2) La métrique du Khi2 qu'elle sous-tend est peu en accord avec la nature des mesures et l'ACP centrée simple des tableaux de variables floues est plus justifiée de ce point de vue.

3) La réduction d'un tableau de traits par variables floues peut ou non être une bonne chose. Elle est judicieuse en cas de redondance entre traits, elle ne l'est pas dans le cas contraire. Les tableaux analysés ici ont des structures caractérisée par une bonne part d'indépendance des traits biologiques entre eux. Cela veut dire que sur plusieurs critères les taxons ont utilisé des combinaisons de modalités très diversifiées, chaque variable induisant une typologie d'espèce originale par rapport aux autres. Pour avoir un avis rapide sur ce point crucial, **`statis`** (opérateurs), dans sa partie inter-structure est efficace.

4) Indépendance ou redondance des traits influent fortement sur le lien avec un tableau écologique (sites-espèces) ou un tableau de traits écologiques. Si la réduction du tableau de traits est complexe, le lien avec un autre tableau le sera plus encore.

5) On peut alors prévoir qu'une bonne stratégie consiste à décomposer la partie biologique en morceaux indépendants contenant un seul trait ou plusieurs traits covariants. La nécessité de méthodes de typologie de tableaux est ici très forte. C'est un problème ouvert.

7. Références

Bournaud M., Richoux P. & Usseglio-Polatera P. (1992) An approach to the synthesis of qualitative ecological information from aquatic coleoptera communities. *Regulated rivers: Research and Management*, 7, 165-180

Cazes P. (1990) Codage d'une variable continue en vue de l'analyse des correspondances. *Revue de Statistique Appliquée*, 38, 35-51

Cazes P., Chessel D. & Dolédec S. (1988) L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. *Revue de Statistique Appliquée*, 36, 39-54

Chevenet F., Dolédec S. & Chessel D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, 31, 295-309

Dray S., Chessel D. & Thioulouse J. (2003) Procustean co-inertia analysis for the linking of multivariate datasets. *Ecoscience*, 10, 110-119

Escoufier Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29, 750-760

Escoufier Y. (1977) Operators related to a data matrix. In: *Recent developments in Statistics* (eds. Barra JR & Coll.), pp. 125-131. North-Holland

Escoufier Y. (1980) L'analyse conjointe de plusieurs matrices de données. In: *Biométrie et Temps* (ed. Jolivet M), pp. 59-76. Société Française de Biométrie, Paris

Escoufier Y. (1985) Objectifs et procédures de l'analyse conjointe de plusieurs tableaux de données. *Statistique et Analyse des Données*, 10, 1-10

Escoufier Y. (1987) Three-mode data analysis: the STATIS method. In: *Methods for multidimensional data analysis*, pp. 325-338. ECAS

Felsenstein J. (1985) Phylogenies and the comparative method. *The American Naturalist*, 125, 1-15

Heo M. & Gabriel K.R. (1998) A permutation test of association between configurations by means of the RV coefficient. *Communications in Statistics - Simulation and Computation*, 27, 843-856

Jackson D.A. (1995) PROTEST: a PROcustean randomization TEST of community environment concordance. *Ecosciences*, 2, 297-303

Statzner B., Hoppenhaus K., Arens M.-F. & Richoux P. (1997) Reproductive traits, habitat use and templet theory: a synthesis of world-wide data on aquatic insects. *Freshwater Biology*, 38, 109-135

van Rijckevorsel J. (1987) *The application of fuzzy coding and hoerseshoes in multiple correspondence analysis*. DSWO Press, Leiden.

van Rijckevorsel J.L.A. (1988) Fuzzy coding and B-splines. In: *Component and correspondences analysis* (eds. van Rijckevorsel JLA & de Leeuw J), pp. 33-54. John Wiley & Sons Ltd, New York