TD avec le logiciel  - Fiche 6.8

L'ordination simultanée de plusieurs tableaux

Plan

Sommaire

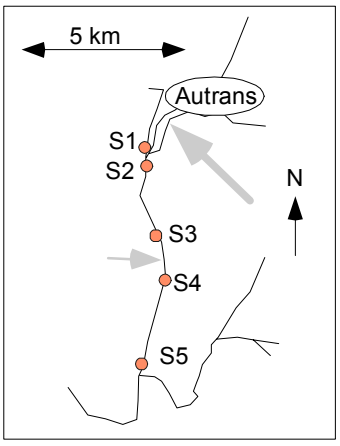
La fiche introduit à la pratique de quelques unes des méthode d'ordination *K*-tableaux. On présente quelques situations expérimentales caractéristiques et la nécessité de poser avec précision les objectifs poursuivis. On aborde les *K*-tableaux doublement appariés avec l'analyse triadique partielle et l'analyse des correspondances de Foucart. Les principes de l'analyse factorielle multiple et de STATIS sont explorés.

1. Des exemples

1.1. Stations-dates-faune-milieu : meaudret

La liste meaudret (Pegaz-Maucet 1980) contient trois data.frame. Le premier donne les variables environnementales mesurées sur une rivière pour 4 dates et 5 stations :

```
> data(meaudret)
> meaudret$mil
   Temp Debit  pH Condu Dbo5 Oxyd  Ammo Nitra Phos
1    10    41 8.5   295  2.3  1.4  0.12   3.4 0.11
2    11   158 8.3   315  7.6  3.3  2.85   2.7 1.50
3    11   198 8.5   290  3.3  1.5  0.40   4.0 0.10
4    12   280 8.6   290  3.5  1.5  0.45   4.0 0.73
5    13   322 8.5   285  3.6  1.6  0.48   4.6 0.84
...
17    3   252 8.3   360  9.5  2.9  2.52   4.6 1.60
18    3   315 8.3   370  8.7  2.8  2.80   4.8 2.85
19    3   498 8.3   330  4.8  1.6  1.04   4.4 0.82
20    2   390 8.2   330  1.7  1.2  0.56   5.0 0.60

Code des variables
1- Température (°C)
2- Débit (l/s)
3- pH
4- Conductivité (mmho/cm)
5- Oxygène (% saturation)
6- DBO5 (mg/l oxygène)
7- Oxydabilité (idem)
8- Ammoniaque (mg/l)
9- Nitrates (mg/l)
10- Orthophosphates (mg/l)
```



Le second donne le plan d'observations (4 périodes et 5 stations) :

```
> meaudret$plan

 dat sta
1  spring  S1
2  spring  S2
3  spring  S3
4  spring  S4
5  spring  S5
6  summer  S1
...
19 winter  S4
20 winter  S5
```

Sur la carte, les flèches grises indiquent les sources de pollution organique.

Le troisième donne l'abondance de 13 catégories d'Éphéméroptères (classe d'abondance en échelle logarithmique) :

```
> meaudret$fau
   Eda Bsp Brh Bni Bpu Cen Ecd Rhi Hla Hab Par Cae Eig
1    4   7  10   9   0   0   0   5   9   0   4   0   0
2    0   0   8   0   0   0   0   0   4   0   0   0   0
3    0   5   5   0   0   0   0   2   5   0   0   0   0
...
17   0   3   6   0   0   5   0   4   3   0   1   0   0
18   0   0   3   0   0   1   0   1   0   0   0   0   0
19   0   6  10   0   0   5   1   3   5   0   2   0   0
```

```
20   1   9  11   0   3   6   8   3   5   2   5   0   0
```

Cet exemple pédagogique permet de poser la question des *K*-tableaux, structure de données comportant plusieurs tableaux ayant en commun soit les lignes, soit les colonnes, soit les deux à la fois (cubes). Ici on a deux cubes de données.

1.2. Milieu et groupes faunistiques - friday87

La liste 'friday87' (Friday 1987) comprend 4 composantes :

```
> friday87$mil
  pond.area veg.area pH Conduc BOD hardness Alkali Phospha Nitra
Q      1380      187 38    215   8       21       2       1     1
P        87       35 40    125   7       14       3       1     1
R        63       63 42    110  21       15       5       1     1
J       472      407 47    150  14       24       2       7     1
E        23       23 50    150  15       32       3       6     2
C        67       67 55    145  20       26      13       4     4
D       100      100 55    150  17       26      13       5     4
K       292       38 60    180  35       34      16       3     1
B        80       80 65    255  29       50      42       4     2
A        72       72 65    295  31       90      42       7     5
G       323      227 68    310  34       64      53       8     6
M       449      134 68    500  17       84      39      12     1
L       438       16 69    460  84      110      86      80     1
F       515       25 71    305  26       86      48       8    20
H      2000      992 75    300  34       81      51       4     3
N       291      229 75    405  73       92      76      23    14
```

Code des variables : mesures sur 16 étangs

```
1- pond.area surface de l'étang (ha x 1000)
2- veg.area surface en végétation (ha x 1000)
3- pH x 10
4- Conduc Conductivité (mmho/cm)
6- BOD DBO5 (mg/l oxygène x 10)
7- hardness Dureté totale (mg/l)
8- Alkali Dureté calcique (mg/l)
9- Phospha Phosphates (mg/l x 100)
10- Nitra Nitrates (mg/l x 10)
```

```
> friday87$fau
  A1 A2 A3 A4 A5 A6 A7 A8 A9 Aa Ab B1 B2 B3 B4 B5 B6 B7 C1 C2 C3 C4 C5 C6 C7 C8
Q  3  2  2  1  1  0  0  0  0  0  0  0  0  0  2  0  0  0  2  1  0  0  0  0  0  0
P  0  0  3  0  0  1  0  0  0  0  0  2  2  3  2  2  0  0  3  0  1  0  0  0  0  0
...
H  0  0  0  0  1  2  0  0  0  0  0  0  0  0  1  1  2  3  0  2  1  0  0  3  2  1
N  0  1  0  1  1  3  1  1  3  1  2  0  0  1  3  3  3  3  0  3  0  2  0  2  2  0
...
  I2 I3 I4 I5 I6 I7 I8 J1 J2 J3 J4 J5 J6
Q  0  0  0  0  0  0  0  4  2  2  0  0  0
P  0  0  0  0  0  0  0  3  0  0  0  0  0
...
H  2  0  1  0  1  2  2  2  2  3  0  3  3
N  3  0  3  3  0  0  3  3  3  2  0  0  3
```

Abondances dans 16 étangs de 91 espèces d'invertébrés benthiques. Mesures par classes d'abondances (0 / absent, 1 / présent, 2 / 1à10 individus par unités d'échantillonnage, 3 / 11 à 100 individus par unités d'échantillonnage, 4 / plus de 100 individus par unités d'échantillonnage). Ces taxons appartiennent à 10 groupes (ordre) :

```
> friday87$tab.names
 [1] "Hemiptera"     "Odonata"       "Trichoptera"   "Ephemeroptera"
 [5] "Coleoptera"    "Diptera"       "Hydracarina"   "Malacostraca"
```

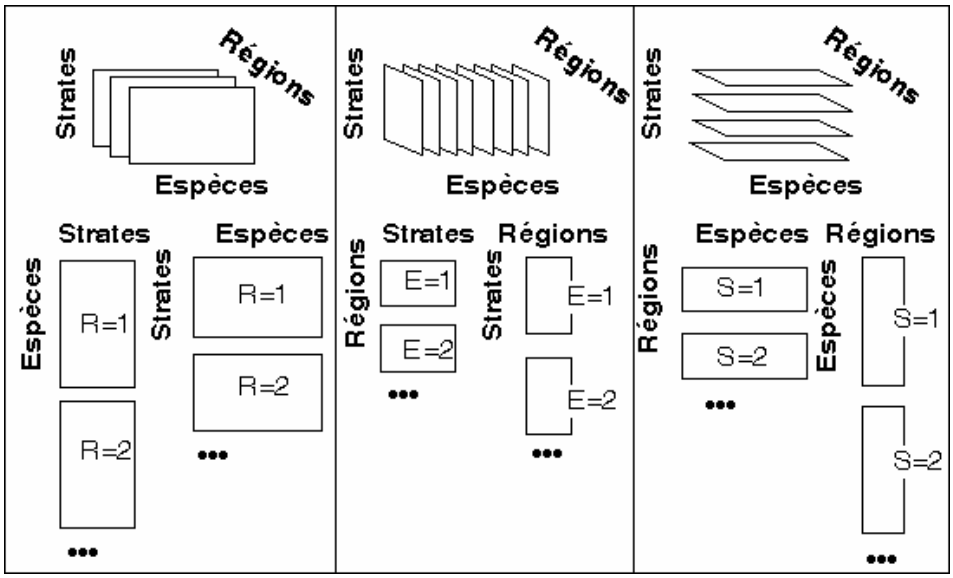
```
 [9] "Mollusca"      "Oligochaeta"

> friday87$fau.blo
    Hemiptera       Odonata    Trichoptera Ephemeroptera     Coleoptera
           11             7             13             4             13
      Diptera   Hydracarina  Malacostraca      Mollusca   Oligochaeta
           22             4             3              8             6
```

Ici, on a implicitement 10 tableaux faunistiques et 1 tableau de milieu.

1.3. Cube de données faunistique : bf88

La liste 'bf88' (Blondel and Farre 1988) contient 6 composantes :

```
> data(bf88)
> names(bf88)
[1] "S1" "S2" "S3" "S4" "S5" "S6"
> names(bf88$S1)
[1] "Pol" "Bur" "Pro" "Cor"
```



Les tableaux forment un cube avec :

- 6 strates qui définissent le type de végétation depuis la plus ouverte (1- végétation buissonnante basse de hauteur inférieure à 1 m) jusqu'à la plus fermée (6 - forêts de plus de 20 m de hauteur)
- 4 régions, Pologne, Bourgogne , Provence et Corse
- 79 colonnes espèces d'oiseaux.

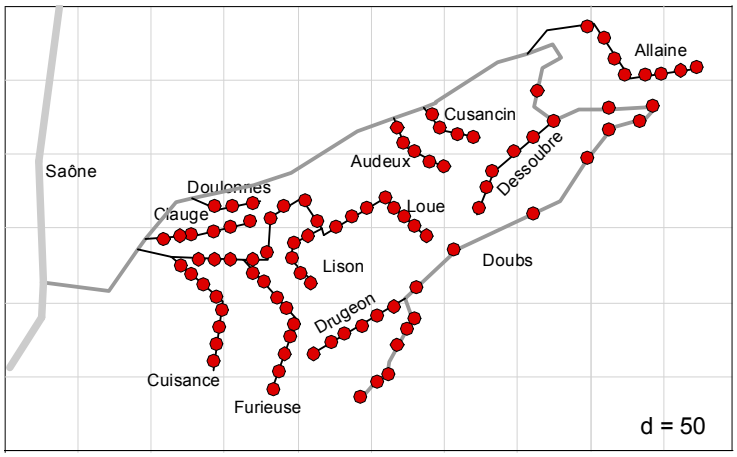
```
> lapply(bf88,dim)
$S1
[1] 79  4
…
$S6
[1] 79  4
```

Les valeurs sont des effectifs de couples nicheurs pour 100 ha. On doit penser que la forme de présentation des données, quelle qu'elle soit, permet d'accéder à un cube qui génère 6 manières d'assembler des tableaux ordinaires (ci-dessus).

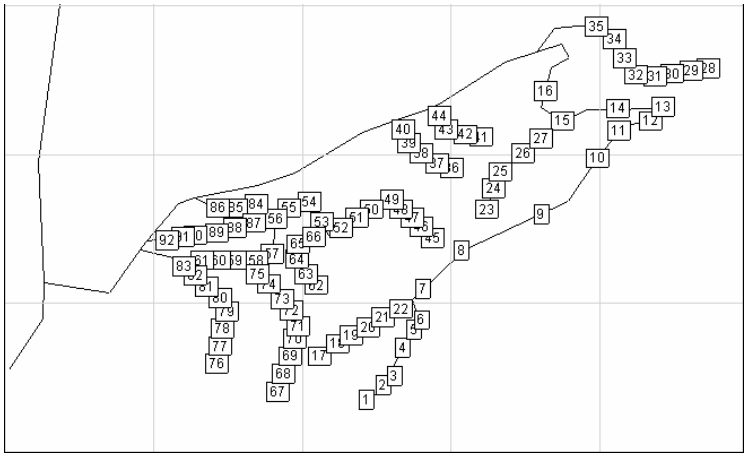
La liste donne le cube comme 6 data.frame à 79 lignes (espèces) et 4 colonnes (régions).

1.4. Groupes d'individus et groupes de variables : jv73

La liste 'jv73' (Verneaux 1973) contient l'information sur 92 stations réparties le long de 12 rivières. On utilise une partie des données récemment recodées par l'auteur. La figure qui suit donne le schéma de principe de la répartition des stations dans le réseau hydrographique très simplifié du bassin du Doubs :




```
> data(jv73)
> s.label(jv73$xy,contour=jv73$contour,clab=0.75, addaxes=F)
```



Le nom de la rivière sur laquelle on trouve chacune des stations est donné par un facteur :

```
> levels(jv73$fac.riv)
 [1] "Allaine"    "Audeux"     "Clauge"     "Cuisance"   "Cusancin"   "Dessoubre"
 [7] "Doubs"
"Doulonnes" "Drugeon"    "Furieuse"   "Lison"
"Loue"
```

Le tableau 'morpho' contient pour chacune des stations l'altitude, la distance à la source, la pente, la section mouillée, le débit moyen et la vitesse moyenne de l'eau :

```
> names(jv73$morpho)
[1] "Alt" "Das" "Pen" "Smm" "Qmm" "Vme"
```

Les variables retenues pour la description du milieu sont simplement recodées pour uniformiser les amplitudes de variation et rendre les distributions plus acceptables. Les variables peuvent prendre des valeurs comprises entre 0 et 9, valeurs associées aux numéros de classes (e classe d'amplitude égale, r raison géométrique) :

Repères géomorphologiques					
Nom	Code	Unités	Raison	Classe 0	Classe 9
Altitude	Alt	m	e = 50	< 150	> 1350
Distance à la source	Das	km	r = 2.5	< 0.4	> 810
Pente	Pen	°/°°	r = 2	< 0.1	> 25.6
Section mouillée	Smm	m2	r = 3	< 0.05	> 328
Débit moyen	Qmm	m3/s	r = 2	< 0.2	> 510.2
Vitesse (Qmm/Smm)	Vme	m/s	e = 0.1	< 0.1	> 0.9

Le tableau 'phychi' contient les variables physico-chimiques :

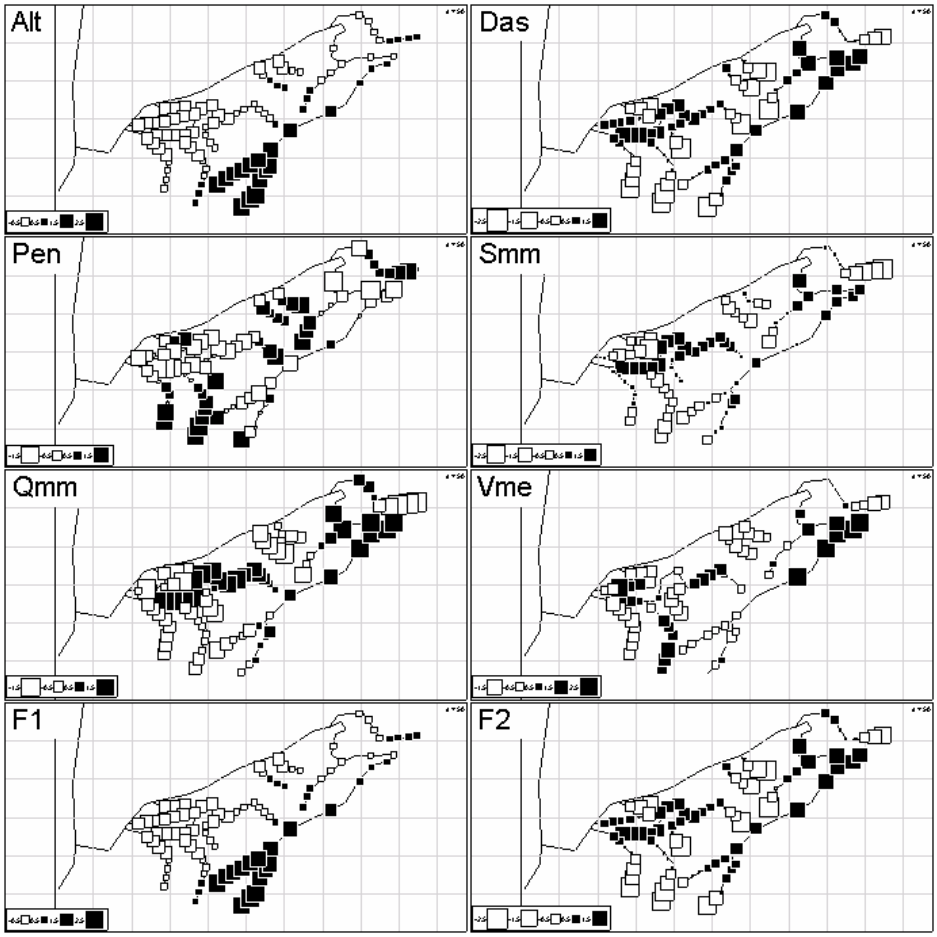
Paramètres physiques et chimiques					
Nom	Code	Unités	Raison	Classe 0	Classe 9
Température	Tmm	°C	e = 2	< 8	> 24
Conductivité	Con	µs/cm	e = 75	< 50	> 650
pH	pH		e = 0.4	< 6	> 9.2
Dureté	Dur	mg/l Ca	e = 15	< 10	> 130
Chlorures	Cl-	mg/l	e = 3	< 3	> 27
Sulfates	SO4--	mg/l	e = 10	< 10	> 90
Phosphates	PO4---	mg/l	r = 2	< 0.005	> 1.280
Nitrates	NO3-	mg/l	r = 2	< 0.05	> 12.8
Azote nitreux et ammonia	NO2-/NH4+	mg/l N	r = 2	< 0.004	> 1.024
Oxygène dissous	O2%		e = 5	< 60	> 100
Oxydabilité	OXY		e = 2	< 1	> 17
DBO	DBO	mg/l	e = 2	< 1	> 17

```
> names(jv73$phychi)
 [1] "Tmm"    "Con"    "pH"     "Dur"    "Cl-"    "SO4--"  "PO4---" "NO3-"
 [9] "N"      "O2%"    "OXY"    "DBO"
```

On a enfin l'abondance de 19 taxons (Poissons) :

```
> names(jv73$poi)
 [1] "Chb" "Tru" "Vai" "Loc" "Omb" "Bla" "Hot" "Tox" "Van" "Che" "Bar" "Lot"
[13] "Spi" "Gou" "Bro" "Per" "Tan" "Gar" "Lam"

Chabot, Truite, Vairon, Loche, Ombre, Blageon, Hotu, Toxostome, Vandoise,
Chevaine, Barbeau, Lotte, Spirlin, Goujon, Brochet, Perche, Tanche, Gardon,
Lamproie.

> par(mfrow=c(4,2))
> w <- scale(jv73$morpho)
> for(i in 1:6) s.value(jv73$xy,w[,i],
        contour=jv73$contour,ylim=c(0,300),sub=names(jv73$morpho)[i],
        possub="topleft",csub=3)
> w <- dudi.pca(morho,scann=F)$li
> s.value(jv73$xy,w[,1],
        contour=jv73$contour,ylim=c(0,300),sub="F1",possub="topleft",csub=3)
> s.value(jv73$xy,w[,2],
        contour=jv73$contour,ylim=c(0,300),sub="F2",possub="topleft",csub=3)
```

Seul le premier ensemble de descripteurs a une structure simple :



Mais elle montre cependant clairement que se mélangent des structures inter-rivières et des typologies de variation intra-rivières. On peut chercher à distinguer des groupes de stations ou, au contraire analyser le niveau de reproductibilité d'une structure unique d'une rivière à l'autre, sa permanence ou sa variabilité et détecter des causes éventuelles de sa perturbation. On s'interroge en particulier sur la solidité de la liaison de cette structure faunistique avec l'évolution multivariée des paramètres morphologiques et physico-chimiques relevés sur des stations régulièrement disposés sur plusieurs rivières.

Le but est donc d'employer des méthodes différentes pour des objectifs contradictoires.

2. L'importance des objectifs

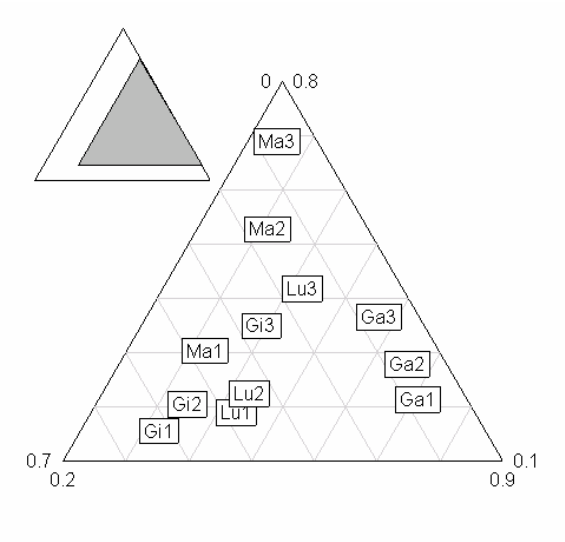
Quand on aborde l'analyse des données multi-tableaux, la principale difficulté est la maîtrise des objectifs poursuivis. La richesse implicite de la structure des données interdit un comportement basé sur la présentation formelle des données. Deux exemples qui ne demandent aucune introduction technique montrent cet aspect fondamental. Prenons le petit exemple numérique qui figure en fin de l'ouvrage de C. Lavit (1988) pour l'illustration du programme STATIS.

```
4 communes (Gignac, Ganges, Matellas, Lunel) donnent 3 variables communes
(taux des emplois agricoles, ouvriers et tertiaires) pour trois recensements
(1968, 1975 et 1982) :

51.88   32.55   15.57
44.94   34.59   20.47
```

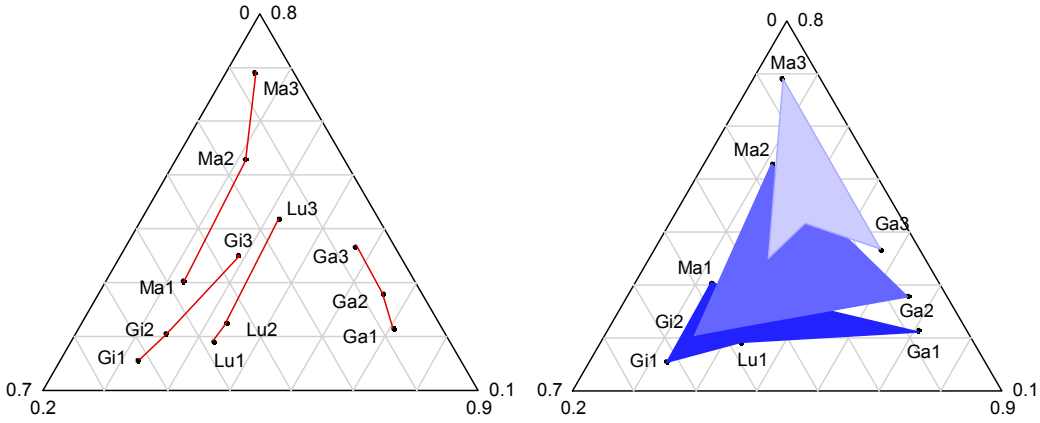
```
25.95   39.15   34.9

7.76    70.93   21.31
6.22    65.96   27.82
6.44    57.06   36.5

37.24   32.45   30.31
16.09   31.22   52.69
6.54    24.68   68.78

37.87   43.19   18.94
34.2    43.32   22.48
16.13   42.18   41.69
```

Installer les données "à la main". On représentent, techniquement toute l'information disponible avec :

```
> triangle.plot(exo1,label=lab,clab=1, labeltriangle=F)
```



Or cette figure, interprétée, en donne deux autres :

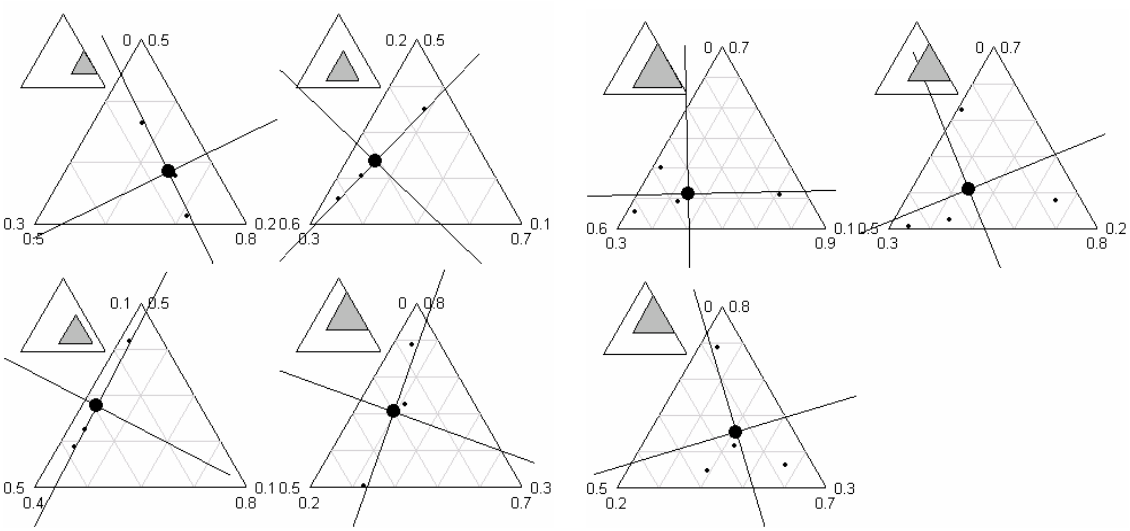


A gauche, on a dessiné les évolutions par commune, ce qui conduit à une question sur *la typologie des évolutions*. A droite on a dessiné les typologies de communes par dates, ce qui conduit à une question sur *l'évolution des typologies*. Ici les données sont en dimension deux. On voit tout du même endroit. Quand il y a de nombreuses variables, les deux représentations finalisées peuvent être totalement différentes.

```
> com <- factor(rep(c("Gi","Ga","Ma","Lu"),c(3,3,3,3)))
> dat <- factor(rep(c("D1","D2","D3"),4))
> plan <- cbind.data.frame(com,dat)
> plan
```

```
   com dat
1   Gi  D1
2   Gi  D2
3   Gi  D3
4   Ga  D1
5   Ga  D2
6   Ga  D3
7   Ma  D1
8   Ma  D2
9   Ma  D3
10  Lu  D1
11  Lu  D2
12  Lu  D3
```

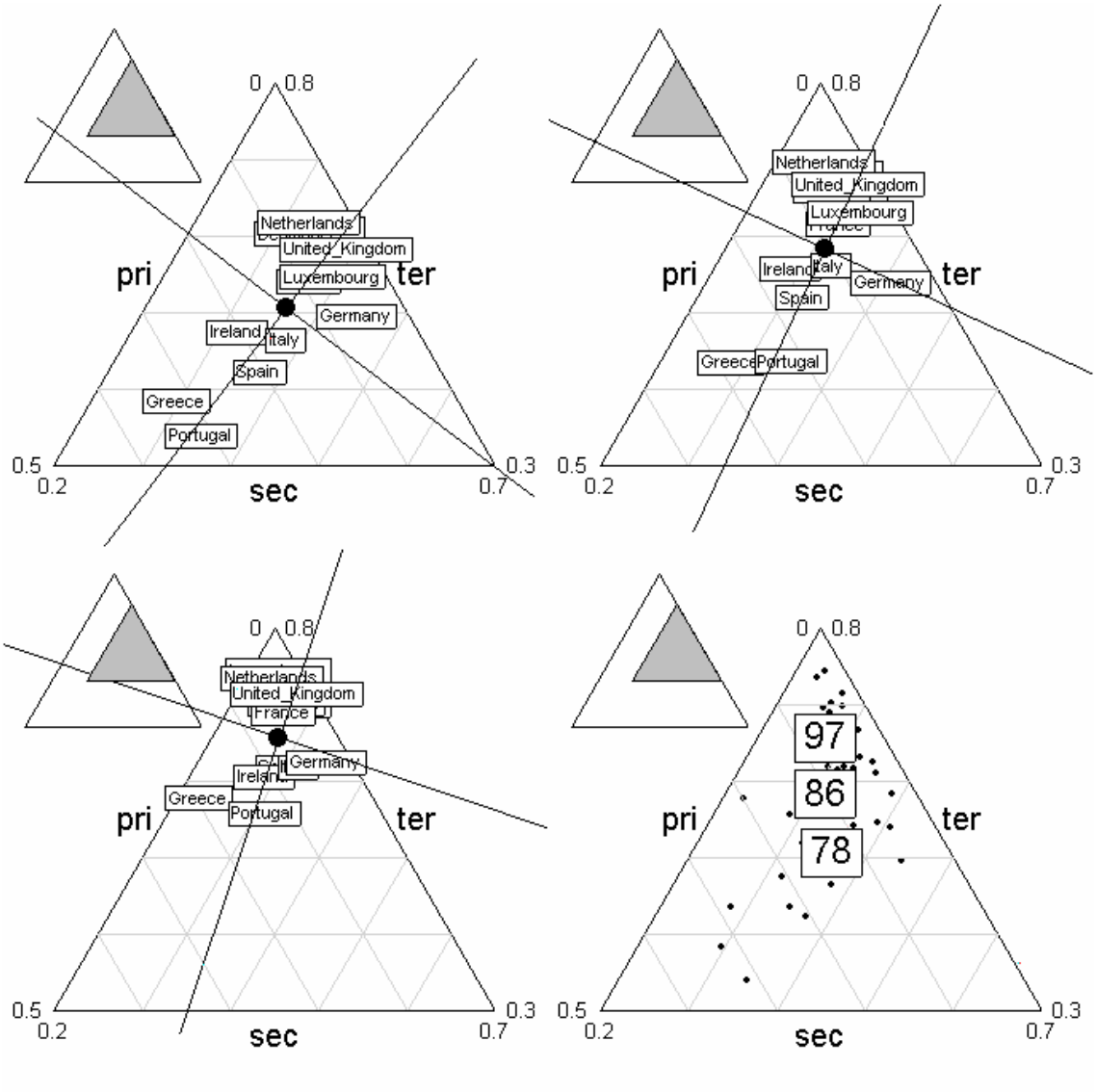
```
> for(i in 1:4) triangle.plot(exo1[com==levels(com)[i],],
      sub=as.character(levels(com)[i]),addax=T, labeltriangle=F)
> for(i in 1:3) triangle.plot(exo1[dat==levels(dat)[i],],
      sub=as.character(levels(dat)[i]),addax=T, labeltriangle=F)
```



Ce graphe représente à gauche les quatre analyses séparées associées au premier problème, à droite il donne les trois analyse séparées associées au second problème. Les deux types sont très différents. Il vaut mieux décider avant de commencer ce que l'on cherche.

Un autre exemple est disponible dans la liste 'euro123' :

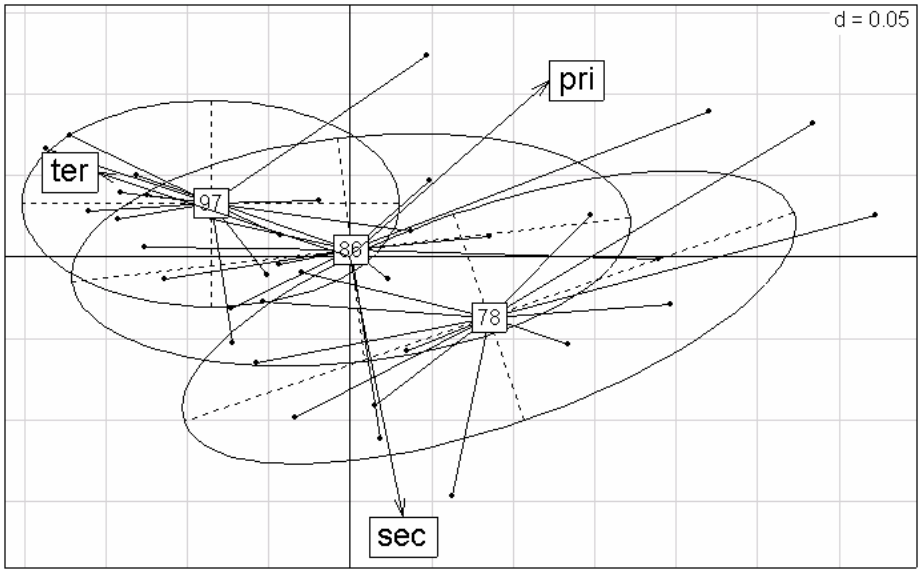
```
> data(euro123)
> par(mfrow=c(2,2))
> triangle.plot(euro123[[1]],min3=c(0,0.2,0.3),max3=c(0.5,0.7,0.8),
      clab=1,label=row.names(euro123[[1]]),addax=T)
> triangle.plot(euro123[[2]],min3=c(0,0.2,0.3),max3=c(0.5,0.7,0.8),
      clab=1,label=row.names(euro123[[1]]),addax=T)
> triangle.plot(euro123[[3]],min3=c(0,0.2,0.3),max3=c(0.5,0.7,0.8),
      clab=1,label=row.names(euro123[[1]]),addax=T)
> triangle.plot(rbind.data.frame(euro123[[1]],euro123[[2]],euro123[[3]]))

> data.frame(lapply(euro123[1:3],function(x) apply(x,2,mean)))
     in78   in86    in97
pri 0.134 0.1047 0.06708
sec 0.360 0.3130 0.27517
ter 0.506 0.5823 0.64308

> moy <- t(data.frame(lapply(euro123[1:3],function(x) apply(x,2,mean))))
> moy
         pri    sec    ter
in78 0.13400 0.3600 0.5060
in86 0.10467 0.3130 0.5823
in97 0.06708 0.2752 0.6431
> par(new=T)
> triangle.plot(moy,min3=c(0,0.2,0.3),max3=c(0.5,0.7,0.8),
      label=c("78","86","97"),clab=2)
```



La figure pose clairement la question de l'axe de l'évolution ou de l'évolution de l'axe. L'évolution de la moyenne est une question. La déformation de la structure en est une autre. La difficulté de la problématique n'implique pas nécessairement celle de la mise en œuvre technique puisqu'ici on voit en permanence toute l'information. La statistique *K*-tableaux introduit la notion de moyenne et de variabilité non plus d'une valeur mais d'une typologie. Est en cause la variabilité de la variabilité. C'est une question nouvelle.

La liste 'euro123' est formé de trois tableaux *que nous connaissons complètement*. Nous pouvons assembler ces 3 tableaux par les colonnes :

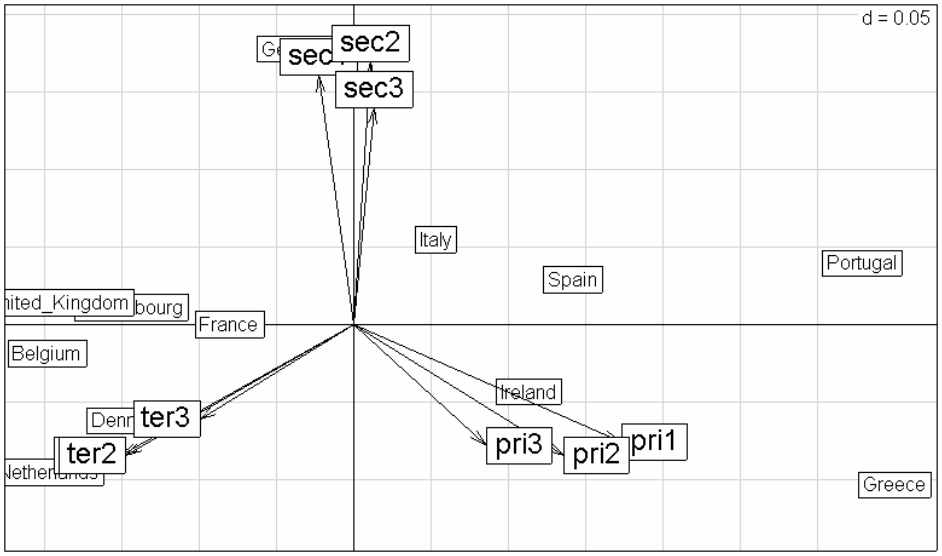
```
> euro <- rbind.data.frame(euro123[[1]],euro123[[2]],euro123[[3]])
> plan <- euro123[[4]]
> row.names(euro) <- row.names(plan)
> pca1 <- dudi.pca(euro,scal=F,scan=F)
> s.class(pca1$li,plan$an)
> s.arrow(pca1$c1/5,add.plot=T,clab=1.5)
```



L'axe principal enregistre le maximum de variabilité : il le fait en prenant pour partie la variabilité temporelle (entre dates) et la variabilité spatiale (entre pays).

Nous pouvons assembler ces 3 tableaux par les lignes :

```
> eurobis <- cbind.data.frame(euro123[[1]],euro123[[2]],euro123[[3]])
>                                 names(eurobis)                                    <-
paste(rep(c("pri","sec","ter"),3),rep(1:3,rep(3,3)),sep="")
> eurobis
                pri1  sec1  ter1  pri2  sec2  ter2  pri3  sec3  ter3
Belgium        0.032 0.359 0.609 0.028 0.291 0.681 0.027 0.275 0.698
Denmark        0.079 0.319 0.602 0.059 0.282 0.659 0.037 0.262 0.701
Spain          0.206 0.372 0.422 0.161 0.320 0.519 0.083 0.299 0.618
France         0.092 0.368 0.540 0.073 0.313 0.614 0.046 0.266 0.688
Greece         0.320 0.297 0.383 0.285 0.281 0.434 0.198 0.225 0.577
Ireland        0.206 0.320 0.474 0.157 0.287 0.556 0.109 0.286 0.605
Italy          0.155 0.381 0.464 0.109 0.331 0.560 0.065 0.317 0.618
Luxembourg     0.062 0.392 0.546 0.040 0.330 0.630 0.024 0.233 0.743
Netherlands    0.054 0.330 0.616 0.049 0.255 0.696 0.037 0.229 0.734
Portugal       0.313 0.348 0.339 0.217 0.348 0.435 0.133 0.310 0.557
Germany        0.061 0.444 0.495 0.053 0.409 0.538 0.029 0.347 0.624
United_Kingdom 0.028 0.390 0.582 0.025 0.309 0.666 0.019 0.268 0.713

> pca2 <- dudi.pca(eurobis,scal=F,scan=F)
> s.label(pca2$li)
> s.arrow(pca2$c1/3,add.p=T,clab=1.5)
```



L'utilisation de la même fonction donne des résultats radicalement différents. *Étudier en détail cette analyse*. Nous avons 12 lignes, 9 colonnes et :

```
> pca2$eig
[1] 3.588e-02 5.261e-03 7.882e-04 3.179e-04 1.021e-04 2.783e-05
```

Pourquoi ?

Envoyer des données structurées dans une procédure simple, c'est simplement ne pas afficher d'objectifs et laisser s'imposer des problématiques non maîtrisées. On vient de faire ici une demi analyse inter-classe, une demi analyse intra-classe et une quasi analyse factorielle multiple sans le savoir.

3. Structure de *K*-tableaux et analyses séparées

On a distingué les données pour lesquelles un groupe de lignes est un individu, les données pour lesquelles un groupe de colonnes est une variable, enfin les données pour lesquelles un groupe de lignes ou un groupe de colonnes est une structure. Pour éviter de réécrire les fonctions, la structure retenue est dans la figure 1.

En A, la disposition physique de l'information est dans un tableau unique. Cela ne préjuge pas de la signification. En B, l'information est un ensemble commun de variables mesurées sur plusieurs paquets d'individus. En C, l'information est un ensemble commun d'individus sur lesquels on a mesuré plusieurs paquets de variables. En D, les données sont des tables de contingence ou des tableaux de masses positives ou nulles qui relèvent de l'analyse des correspondances. Ces tableaux sont appariés par les lignes.

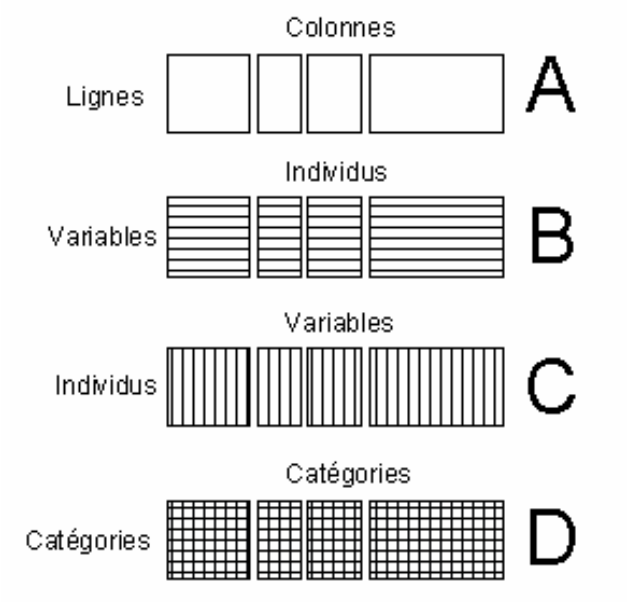


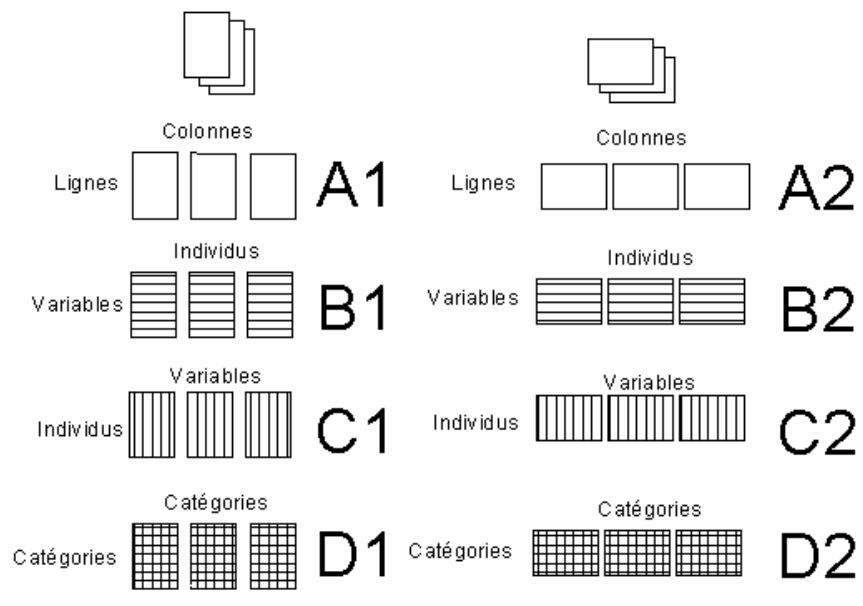
Figure 1

Dans ce point de vue, un *K*-tableaux est une association de *K* analyses qui ont la pondération des lignes en commun. Les fonctions de ade4 :

sepan : analyses séparées

afm : analyse factorielle multiple acom : analyse de co-inertie multiple statis : STATIS sur opérateurs

admettent toute en entrée la forme A, qu'elle soit du type B, C ou D. Il faut donc maîtriser la constitution du point de départ, les résultats pouvant être radicalement différents entre les emplois de la même analyse sur plusieurs préparations initiales différentes. La question est particulièrement critique pour les cubes de données dans lesquels les sous-tableaux ont en commun les lignes et les colonnes :



Le nombre de possibilités a encore augmenté. Les fonctions de ade4 :

pta : analyse triadique partielle foucart : AFC de Foucart

considèrent le cube et donneront le même résultat pour B1 et C2, C1 et B2 ou D1 et D2.

Il y a quatre stratégies de constitution de *K*-tableaux :

ktab.list.dudi : avec une liste de schémas ktab.list.dudi : avec une liste de data.frame ktab.data.frame : avec un data.frame ktab.within : à partir d'une intra-classe

La plus transparente est de faire *K* schémas de dualités séparés et de les assembler par ktab.list.dudi. La fonction vérifie que l'assemblage est possible, en particulier que les pondérations sont compatibles.

```
> data(euro123)
> pca1 <- dudi.pca(euro123$in78,scale=F,scann=F)
> pca2 <- dudi.pca(euro123$in86,scale=F,scann=F)
> pca3 <- dudi.pca(euro123$in97,scale=F,scann=F)
> kta1 <- ktab.list.dudi(list(pca1,pca2,pca3),
      tabnames=c("1978","1986","1997"))

> kta1
class: ktab
```

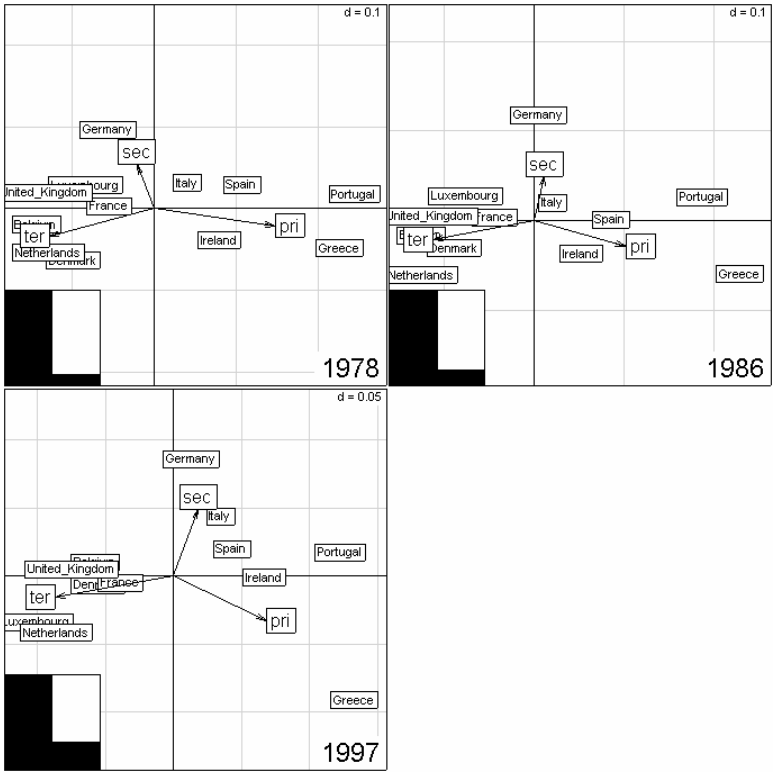
```
tab number:      3
  data.frame nrow ncol
1 1978       12   3
2 1986       12   3
3 1997       12   3

  vector length mode    content
4 $lw    12     numeric row weigths
5 $cw    9      numeric column weights
6 $blo   3      numeric column numbers
7 $tabw  0      NULL    array weights

   data.frame nrow ncol content
8  $TL        36   2    Factors Table number Line number
9  $TC        9    2    Factors Table number Col number
10 $T4        12   2    Factors Table number 1234

11 $call: ktab.list.dudi(obj = list(pca1, pca2, pca3), tabnames = c("1978",
    "1986", "1997"))

names :
1978 : pri sec ter
1986 : pri sec ter
1997 : pri sec ter

Col weigths :
1978 : 1 1 1
1986 : 1 1 1
1997 : 1 1 1

Row weigths :
0.08333  0.08333  0.08333  0.08333  0.08333  0.08333  0.08333  0.08333  0.08333
0.08333 0.08333 0.08333
```

Pour voir l'intégralité du contenu :

```
> unclass(kta1)
```

Pour refaire les analyses séparées :

```
> kplot(sepan(kta1),mfr=c(2,2),clab.c=1.5)
```



La seconde est de partir d'une liste de data.frame et de définir au moment de l'assemblage les poids des lignes et des colonnes :

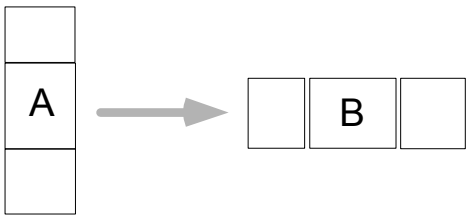
```
> names(euro123)
[1] "in78" "in86" "in97" "plan"
> T78 <- data.frame(scale.wt(euro123$in78,scale=F))
> T86 <- data.frame(scale.wt(euro123$in86,scale=F))
> T97 <- data.frame(scale.wt(euro123$in97,scale=F))
> kta2 <- ktab.list.df(list(T78,T86,T97),tabnames = c("1978","1986", "1997"))
```

Observer qu'on obtient, avec les valeurs par défaut, une pondération unitaire des lignes, ce qui n'est pas le cas du précédent. Cette option est surtout faite pour associer des tableaux dont les lignes sont des variables (data.frame transposés).

La troisième est la plus simple : elle part d'un tableau et d'un vecteur de nombre de colonnes par blocs :

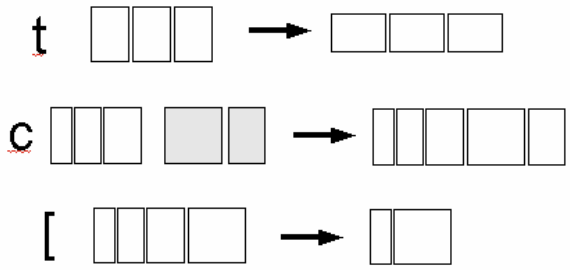
```
> w <- cbind.data.frame (euro123$in78, euro123$in86, euro123$in97)
> w <- scale.wt(w,scale=F)
> w <- data.frame(w)
> kta3 <- ktab.data.frame(w,rep(3,3),tabnames = c("1978","1986", "1997"))
> kta3
```

La dernière part d'une analyse intra-classe avec le schéma de principe :

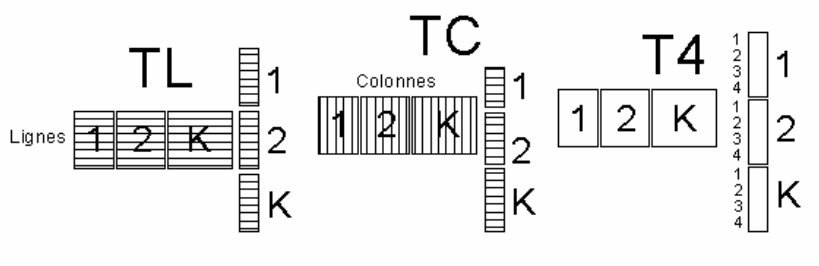



```
> w <- rbind.data.frame (euro123$in78, euro123$in86, euro123$in97)
> pca1 <- dudi.pca(w,scal=F,scan=F)
> wit1 <- within(pca1, factor(rep(c("1978","1986","1997"),rep(12,3))),scan=F)
> ktaprovi <- ktab.within(wit1, colnames=rep(row.names(euro123$in78),3))
> kta4 <- t(ktaprovi)
```

Noter l'obligation de transposer le *K*-tableaux pour se ramener au cas précédent. La fonction 'ktab.within' s'impose pour les tables de contingence (on utilise alors la fonction 'within.coa' dans 'within'). Les opérations possibles sur les *K*-tableaux sont schématisées par :



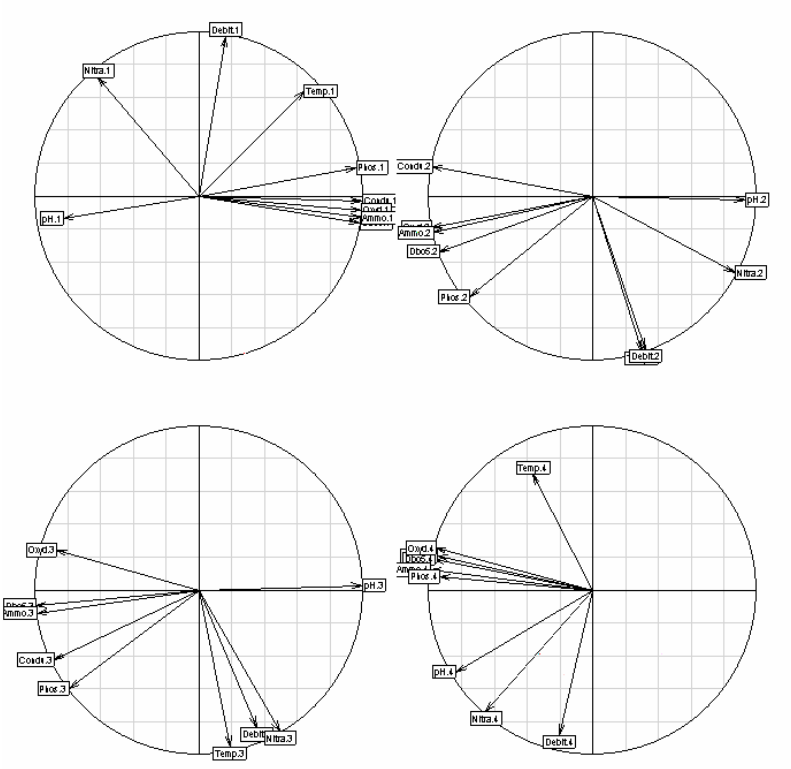
Les composantes TL, TC et T4 contiennent des facteurs permettant de manipuler des informations associées aux tableaux par le schéma :



TL : facteurs numéros de tableau, numéros de la ligne du tableau pour gérer des valeurs associées à chaque ligne de chaque tableau. TC : facteurs numéros de tableau, numéros de la colonne du tableau pour gérer des valeurs associées à chaque colonne de chaque tableau. T4 : facteurs numéros de tableau, numéros de 1 à 4 pour gérer 4 valeurs associées à chaque tableau.

Les analyses séparées sont alors exécutées par 'sepan' qui fournit une liste et directement un plot et un kplot.

```
data(meaudret)
w1 <- split(meaudret$mil,meaudret$plan$dat)
ll <- lapply(w1,dudi.pca,scann=F)
kta <- ktab.list.dudi(ll,rownames <- paste("Station",1:5,sep=""))
par(mfrow=c(2,2))
sep1 <- sepan(kta)
for(j in 1:4) {
   s.corcircle(sep1$Co[sep1$TC[,1]==j,])
}
```

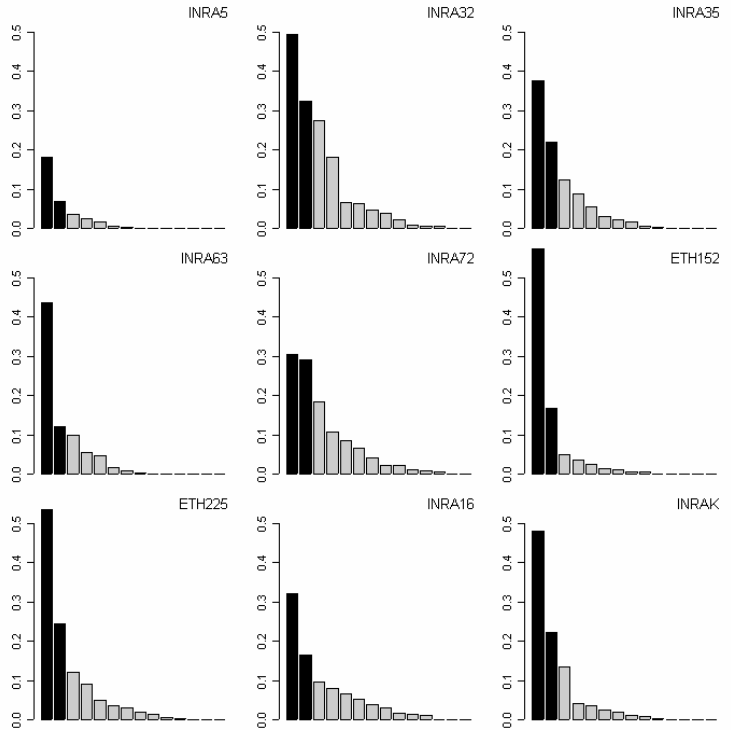


Les composantes, qui ont un sens aléatoire, ne sont pas coordonnées. Elles le seront dans les méthodes suivantes.

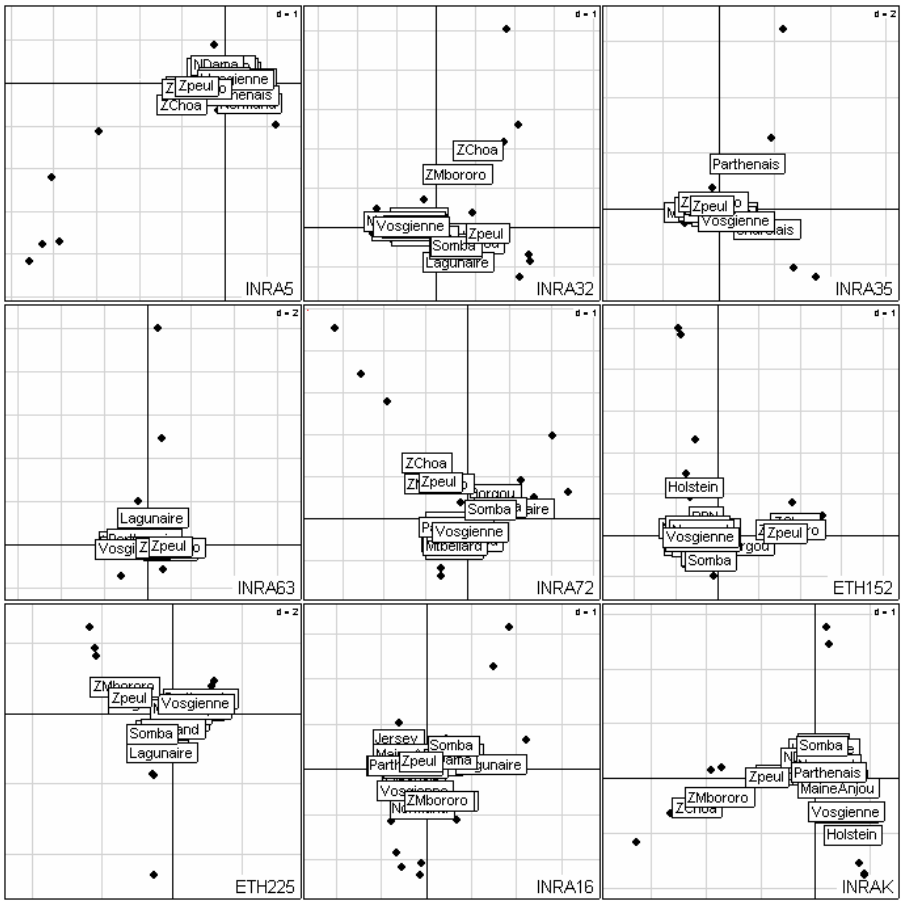
Reprendre le §7 de la fiche stage 2 :

```
data(microsatt)
w <- dudi.coa(data.frame(t(microsatt$tab)),scann=F)
```

```
loci.fac <- factor(rep(microsatt$loci.names,microsatt$loci.eff))
wit <- within(w,loci.fac,scann=F)
microsatt.ktab <- ktab.within(wit)
plot(sepan(microsatt.ktab)) # 9 AFC séparées
```




```
> kplot.sepan.coa(sepan(microsatt.ktab),
     show=F,clab.c=0,mfrow=c(3,3),clab.r=1.5)
```



Les fortes différences de capacité typologique entre marqueurs se voient sur les analyses séparées. Les méthodes suivantes devront aider à les quantifier. On aura soin de vérifier dans tous les cas le contenu du *K*-tableaux, les pondérations, la nature du centrage des tableaux et les noms des composantes (lignes, colonnes, tableaux).

4. Approche élémentaire des cubes

La plus simple des méthodes multi-tableaux pour les cubes est appelée à tort analyse triadique dans (Thioulouse and Chessel 1987) et à raison analyse triadique partielle dans (Kroonenberg 1989) ou Pré-STATIS (ou STATIS sur les X) dans (Leibovici 1993) ou PCA-SUP (PCA of a derived two-way supermatrix) dans (Kiers 1991) et envisagée à l'origine dans (Tucker 1966).

L'objectif est de définir la structure commune à *K* tableaux ayant les mêmes lignes et les mêmes colonnes. La fonction utile est 'pta'.

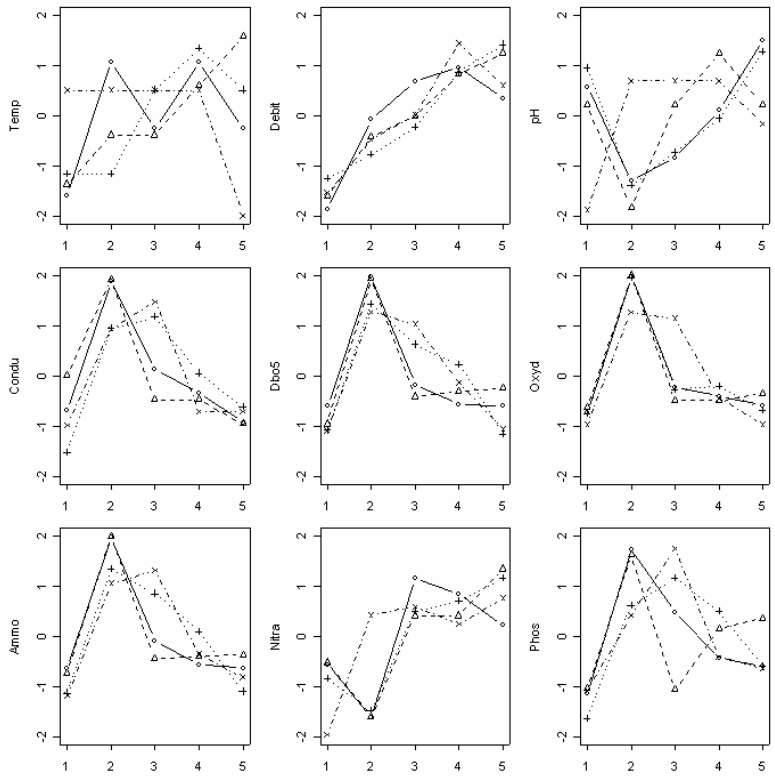
```
> data(meaudret)
> summary(meaudret$plan)
     dat    sta
 autumn:5   S1:4
 spring:5   S2:4
 summer:5   S3:4
 winter:5   S4:4
            S5:4
```

Les mesures portent sur 5 stations pendant 4 périodes, d'où un ensemble de 4 tableaux pour l'étude de la stabilité temporelle de la structure spatiale. On cherche d'abord à représenter les données.

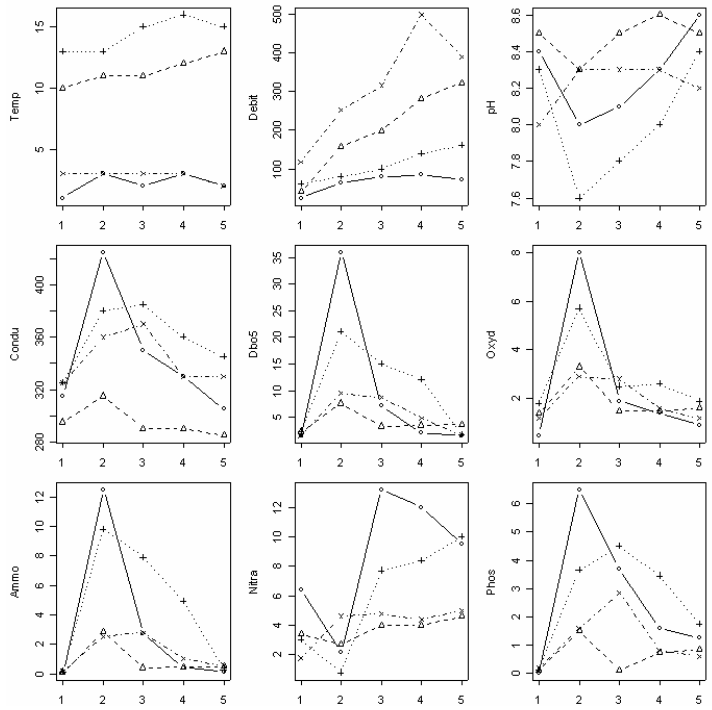
```
> pca0 <- within.pca(meaudret$mil,meaudret$plan$dat,scal="partial",scann=F)
```

Les tableaux sont normalisés par dates :

```
> pca0$tab[meaudret$plan$dat=="summer",]
       Temp   Debit       pH    Condu    Dbo5    Oxyd     Ammo   Nitra    Phos
su_1 -1.167 -1.2684  0.93541 -1.52973 -1.0875 -0.7664 -1.13791 -0.8554 -1.6393
su_2 -1.167 -0.7763 -1.40312  0.94483  1.4232  1.9509  1.32552 -1.4868  0.6112
su_3  0.500 -0.2296 -0.73497  1.16980  0.6176 -0.2787  0.84250  0.4937  1.1547
su_4  1.333  0.8638 -0.06682  0.04499  0.2148 -0.2090  0.07983  0.6946  0.4834
su_5  0.500  1.4105  1.26949 -0.62989 -1.1681 -0.6967 -1.10994  1.1539 -0.6100

> coplot(pca0$tab$Temp~as.numeric(meaudret$plan$sta)|meaudret$plan$dat,
      show.g=F)

par(mfrow=c(3,3))
par(mar=c(2.1,4.1,1.1,1.1))
xsite <- as.numeric(meaudret$plan$sta)
date <- meaudret$plan$dat
for (var in 1:9) {
   plot(xsite,pca0$tab[,var],type="n",ylab=names(pca0$tab)[var],ylim=c(-2,2))
   for (saison in 1:4) {
      sai <- levels(date)[saison]
      points(pca0$tab[meaudret$plan$dat==sai,var],
           lty=saison,pch=saison,type="b")
   }
}
```

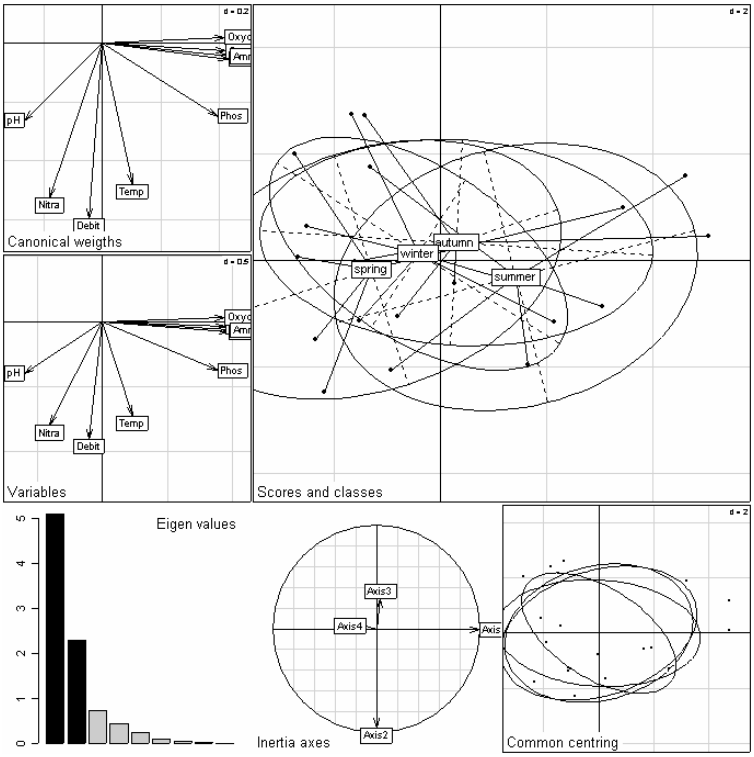



```
par(mfrow=c(3,3))
par(mar=c(2.1,4.1,1.1,1.1))
xsite <- as.numeric(meaudret$plan$sta)
date <- meaudret$plan$dat
for (var in 1:9) {
   plot(xsite,meaudret$mil[,var],type="n",ylab=names(pca0$tab)[var])
   for (saison in 1:4) {
      sai <- levels(date)[saison]
      points(meaudret$mil[meaudret$plan$dat==sai,var],
           lty=saison,pch=saison,type="b")
   }
}
```



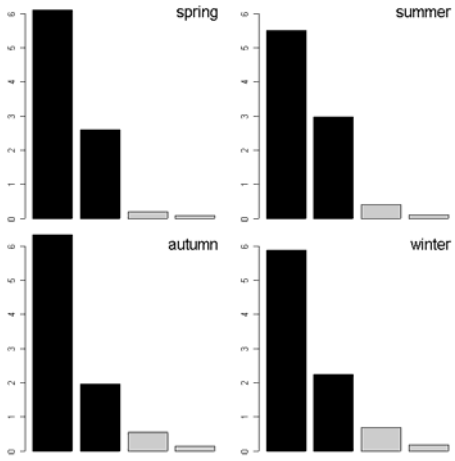
La normalisation par date est l'introduction d'un point de vue. Chaque date donne un tableau d'ACP normée.

```
> plot(pca0)
```

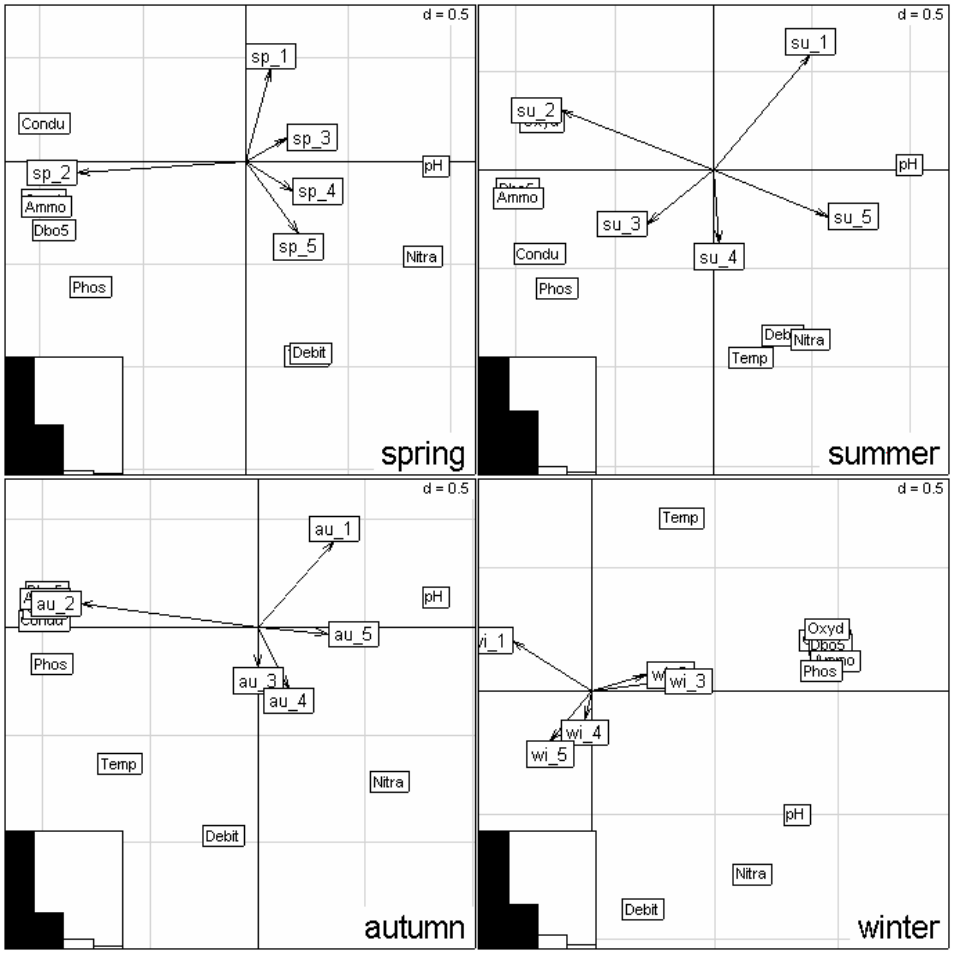


Dépouiller l'analyse intra-classes.

```
> kta0 <- ktab.within(pca0)
> plot(sepan(kta0))
```




```
> kplot(sepan(kta0))
```



On a 4 analyses qui ont bien des points communs. L'analyse triadique partielle fait le bilan de ses points communs. Il s'agit d'abord de typologie moyenne ou *compromis*. Deux tableaux sont directement comparables, puisqu'ils portent sur les mêmes individus (stations) et les mêmes variables (descripteurs). Attention :

```
> pta0 <- pta(kta0)
Error in pta(kta0) : non equal col.names among array

> row.names(kta0)
[1] "Temp"  "Debit" "pH"    "Condu" "Dbo5"  "Oxyd"  "Ammo"  "Nitra" "Phos"
> col.names(kta0)
 [1] "sp_1" "sp_2" "sp_3" "sp_4" "sp_5" "su_1" "su_2" "su_3" "su_4" "su_5"
[11] "au_1" "au_2" "au_3" "au_4" "au_5" "wi_1" "wi_2" "wi_3" "wi_4" "wi_5"
> tab.names(kta0)
[1] "spring" "summer" "autumn" "winter"
```

Les tableaux ont les mêmes lignes, mais pas les mêmes noms de colonnes. La fonction vérifie ce point.

```
> col.names(kta0) <- rep(paste("sta",1:5,sep=""),4)
> pta0 <- pta(kta0)
Select the number of axes: 2
```

Notons n le nombre de lignes et p le nombre de colonnes de chacune des analyses séparées, $\mathbf{D}_n$ et $\mathbf{D}_p$ les normes associées. On peut calculer un produit scalaire entre tableaux :

$$Covv(\mathbf{X}_k, \mathbf{X}_j) = Trace\left(\mathbf{X}_k^t \mathbf{D}_n \mathbf{X}_j \mathbf{D}_p\right) = Trace\left(\mathbf{X}_j^t \mathbf{D}_n \mathbf{X}_k \mathbf{D}_p\right)$$

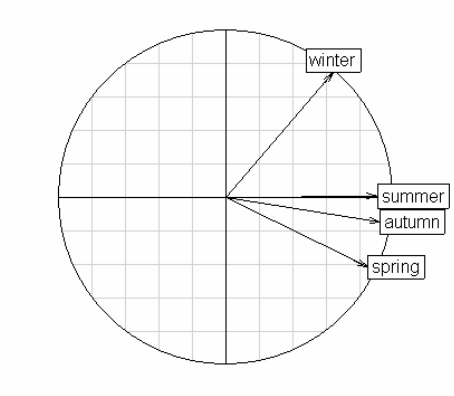
D'où le coefficient de corrélation entre deux tableaux :

$$RV(\mathbf{X}_k, \mathbf{X}_j) = \frac{Covv(\mathbf{X}_k, \mathbf{X}_j)}{\sqrt{Vav(\mathbf{X}_k)}\sqrt{Vav(\mathbf{X}_j)}}$$

```
> pta0
Partial Triadic Analysis
class:pta dudi
table number: 4
row number: 9   column number: 5

     **** Interstructure ****

eigen values: 2.812 0.7541 0.2537 0.18
 $RV            matrix          4       4      RV coefficients
 $RV.eig        vector          4              eigenvalues
 $RV.coo        data.frame      4       4      array scores
 $tab.names     vector          4              array names
…

> pta0$RV
       spring summer autumn winter
spring 1.0000 0.6935 0.7886 0.2835
summer 0.6935 1.0000 0.7672 0.5340
autumn 0.7886 0.7672 1.0000 0.4795
winter 0.2835 0.5340 0.4795 1.0000
```

Les RV sont élevés, mais la structure du tableau 4 est manifestement la plus éloignée du groupe des 3 autres. La matrice des RV est diagonalisée. On obtient une image euclidienne des tableaux pour ce produit scalaire.

```
> pta0$RV.eig
[1] 2.8121 0.7541 0.2537 0.1800
> s.corcircle(pta0$RV.coo)
```



Dans le cas présent, ce produit scalaire est simplement, entre les tableaux *j* et *k* la somme des corrélations des couples de variables identiques :

$$Covv(\mathbf{X}_k, \mathbf{X}_j) = Trace\left(\mathbf{X}_k^t \mathbf{D}_n \mathbf{X}_j \mathbf{D}_p\right) = \sum_{i=1}^{p} corr\left(\mathbf{X}_k^i, \mathbf{X}_j^i\right)$$

Le coefficient RV est donc simplement la moyenne des corrélations des couples de variables identiques :

$$Vav(\mathbf{X}_k) = Covv(\mathbf{X}_k, \mathbf{X}_k) = Trace\left(\mathbf{X}_k^t \mathbf{D}_n \mathbf{X}_k \mathbf{D}_p\right) = \sum_{i=1}^{p} corr\left(\mathbf{X}_k^i, \mathbf{X}_k^i\right) = p$$

$$\Downarrow$$

$$RV(\mathbf{X}_k, \mathbf{X}_k) = \frac{Covv(\mathbf{X}_k, \mathbf{X}_k)}{\sqrt{Vav(\mathbf{X}_k)}\sqrt{Vav(\mathbf{X}_k)}} = \frac{\sum_{i=1}^{p} corr\left(\mathbf{X}_k^i, \mathbf{X}_k^i\right)}{p}$$

Le coefficient RV est la corrélation entre tableaux comme moyenne des corrélations entre variables (pour les couples de variables identiques). Ce coefficient pourrait être négatif, ce qui indiquerait un lien nul voire inversé entre deux tableaux et l'analyse devrait alors s'en tenir là.

La diagonalisation a pour fonction d'attribuer à chaque tableau un poids :

```
> pta0$tabw
[1] 0.5067 0.5404 0.5510 0.3843
```

Le poids attribué au tableau 4 est moindre que celui des trois autres. La somme des carrés de ces poids est égale à 1 :

```
> sum(pta0$tabw^2)
[1] 1
```

dans la même logique que les coefficients des variables (loadings) donnant les composantes principales d'une ACP normée (combinaisons linéaires de variances maximales).

La combinaison des tableaux utilisant ces poids est un nouveau tableau de synthèse combinant les tableaux initiaux à proportion de leurs apports à la description de la structure commune dite compromis. On considère le tableau

$$\mathbf{Y} = \sum_{k=1}^{K} \alpha_k \mathbf{X}_k$$

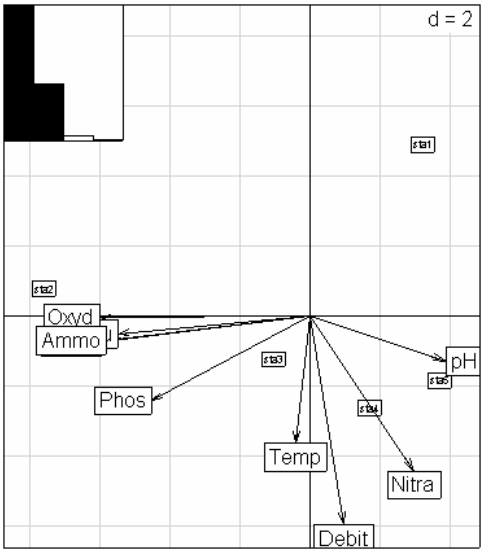
On peut adjoindre à ce tableau les pondérations lignes et colonnes communes à chaque terme du *K*-tableaux et calculer l'inertie du triplet obtenu :

$$Trace\left(\mathbf{Y}^t \mathbf{D}_n \mathbf{Y} \mathbf{D}_p\right) = Trace\left(\left(\sum_{k=1}^{K} \alpha_k \mathbf{X}_k^t\right) \mathbf{D}_n \left(\sum_{j=1}^{K} \alpha_j \mathbf{X}_j\right) \mathbf{D}_p\right)$$
$$= \sum_{\substack{j=1 \\ k=1}}^{K} \alpha_j \alpha_k Trace\left(\mathbf{X}_k^t \mathbf{D}_n \mathbf{X}_j \mathbf{D}_p\right) = \mathbf{a}^t \mathbf{R}_{\mathbf{v}} \mathbf{a}$$

Il s'en suit que les poids retenus maximisent, sous la contrainte "somme des carrés de ces poids égale à 1" l'inertie du tableau combiné. Ce nouveau tableau, dont le contenu importe peu (ce sont des combinaisons des valeurs des tableaux initiaux avec des coefficients tous positifs), a pour fonction de définir des axes et des composantes, donc des vecteurs de $\mathbb{R}^n$ et de $\mathbb{R}^p$, qui expriment la structure compromis. Le programme est donc consacré essentiellement à une recherche d'un compromis inter-tableaux et à l'étude de la structure de ce compromis. Le tableau compromis (tab) et les deux pondérations (cw et lw) ont été conservés et les objets 'pta' sont de la classe 'dudi' :

```
 vector length mode    content
 $tabw  4      numeric array weights
 $cw    5      numeric column weights
 $lw    9      numeric row weights
 data.frame nrow ncol content
 $tab       9    5    modified array
```

```
> s.dudi(pta0,permute=T)
```



Les valeurs propres de ce compromis sont :

```
      **** Compromise ****

eigen values: 17.2 7.298 0.6099 0.2008

> inertia.dudi(pta0,row=T)
$TOT
  inertia   cum  ratio
1 17.2011 17.20 0.6796
2  7.2975 24.50 0.9680
3  0.6099 25.11 0.9921
4  0.2008 25.31 1.0000

$row.rel
       Axis1 Axis2 con.tra
Temp    -115 -8494     581
Debit    256 -9514    1397
pH      8779  -957     640
Condu -9662   -80    1172
Dbo5  -9827  -152    1342
Oxyd  -9703    -1    1399
Ammo  -9846  -138    1300
Nitra  2845 -6295    1165
Phos  -7686 -2178    1004
```

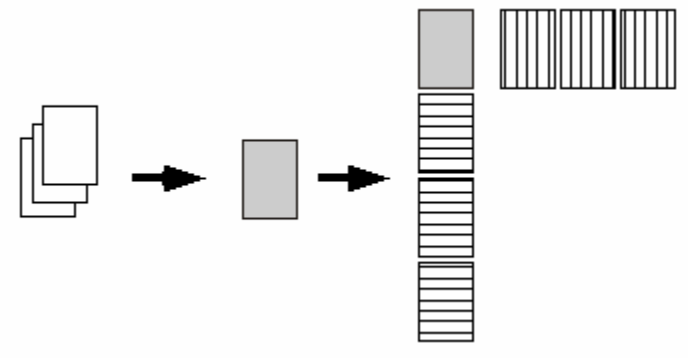
On a gardé deux axes et les cosinus carrés (au sens du même produit scalaire) entre chacun des tableaux et le compromis réduit à nf composantes :

```
 $nf: 2 axis-components saved

 $eig   4      numeric eigen values
 $cos2  4      numeric cosine^2 between compromise and arrays

> pta0$cos2
     1      2      3      4
0.8497 0.9062 0.9239 0.6444
```

On obtient des coordonnées pour les lignes et les colonnes et les scores normés associés :

```
 $li        9    2    row coordinates
 $l1        9    2    row normed scores
 $co        5    2    column coordinates
 $c1        5    2    column normed scores
```

Outre le calcul d'un tableau combinant tous les tableaux de départ et son analyse, on a la configuration très particulière :



En gris le tableau compromis dont l'analyse permet le projection en individus supplémentaires de toutes les lignes et la projection en variables supplémentaires de toutes les colonnes. On peut même projeter sur un plan des axes principaux du compromis les axes principaux de chaque tableau et sur le plan des composantes principales du compromis les composantes principales de chacun des tableaux.

```
     **** Intrastructure ****

 data.frame nrow ncol content
 $Tli       36   2    row coordinates (each table)
 $Tco       20   2    col coordinates (each table)
 $Tcomp     16   2    principal components (each table)
 $Tax       16   2    principal axis (each table)
 $TL        36   2    factors for Tli
 $TC        20   2    factors for Tco
 $T4        16   2    factors for Tax Tcomp
```

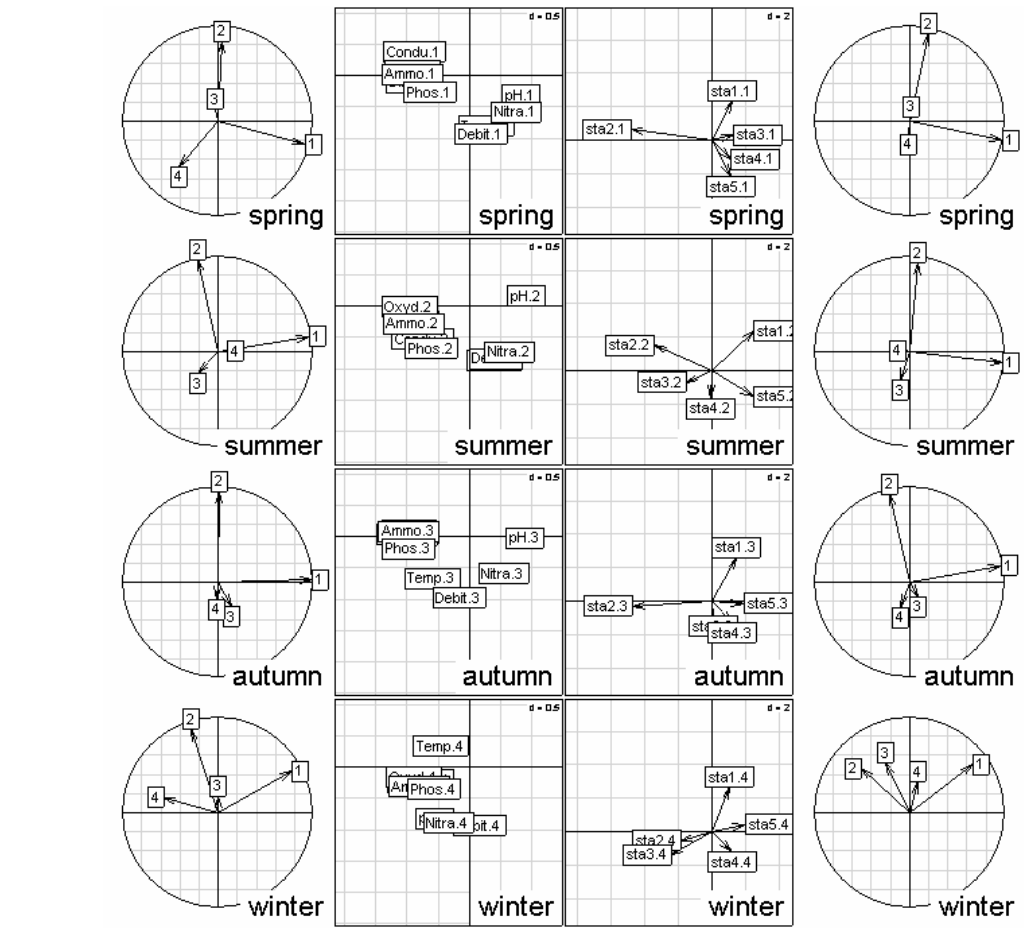
L'ensemble de ces projections est directement accessible dans le 'kplot' de cette analyse :

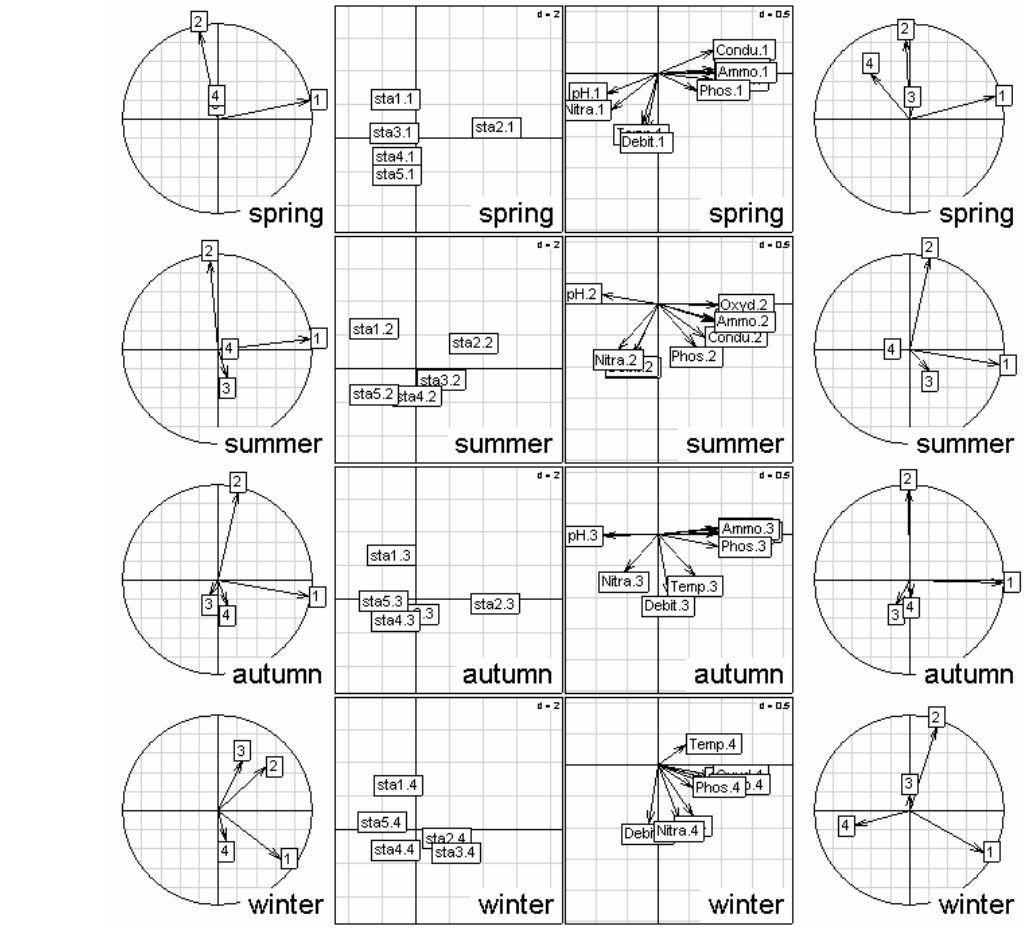
```
> kplot (pta0,clab=1.5,csub=3)
```

Observer (page suivante en haut) que les colonnes sont traitées comme des variables et les lignes comme des individus alors que l'origine du *K*-tableau donnait une disposition inverse. La transposition pure et simple de chaque tableau, qui ne change strictement rien sur le fond, rétablit la situation naturelle :

```
> kplot (pta(t(kta0),scan=F),clab=1.5,csub=3)
```

On voit ainsi *K* analyses simultanées exprimées dans le même cadre géométrique (page suivante en bas).

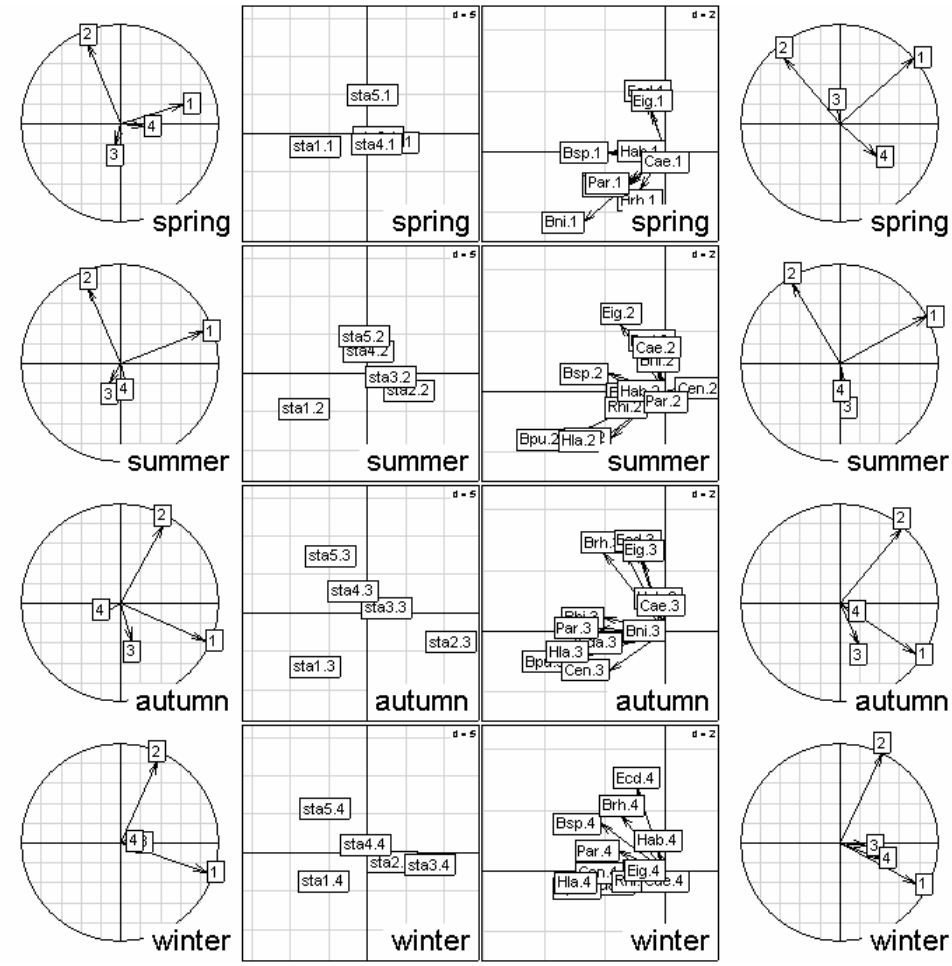

spring
Condu.1
Ammo.1
Phos.1
pH.1
Nitra.1
Debit.1
sta1.1
sta2.1
sta3.1
sta4.1
sta5.1
summer
Oxyd.2
Ammo.2
Phos.2
pH.2
Nitra.2
sta2.2
sta3.2
sta4.2
autumn
Ammo.3
Phos.3
pH.3
Temp.3
Nitra.3
Debit.3
sta1.3
sta2.3
sta5.3
sta4.3
winter
Temp.4
Phos.4
sta1.4
sta5.4
sta2.4
sta3.4
sta4.4


spring
sta1.1
sta3.1
sta2.1
sta4.1
sta5.1
Condu.1
Ammo.1
Phos.1
pH.1
Nitra.1
Debit.1
summer
sta1.2
sta2.2
sta3.2
sta5.2
sta4.2
pH.2
Oxyd.2
Ammo.2
Condu.2
Nitra.2
Phos.2
autumn
sta1.3
sta5.3
sta2.3
sta4.3
pH.3
Ammo.3
Phos.3
Nitra.3
Temp.3
Debit.3
winter
sta1.4
sta5.4
sta2.4
sta4.4
sta3.4
Temp.4
Phos.4
Nitra.4

L'analyse triadique partielle est donc très simple dans ses fondements comme dans son usage. Ici l'essentiel tient dans la permanence de deux axes établissant une boucle (station 1 non perturbée, pollution dans la station 2, restauration progressive sur 3-4-5 liée à l'augmentation du débit), l'affaiblissement de cette structure en hiver et une restauration plus rapide au printemps.

Interpréter cette analyse sur la faune (Trichoptères) :

```
> pca1 <- dudi.pca(meaudret$fau,scal=F,scan=F)
> wit1 <- within(pca1,meaudret$plan$dat,scan=F)
> kta1 <- ktab.within(wit1)
> col.names(kta1) <- rep(paste("sta",1:5,sep=""),4)
> kta1 <- t(kta1)
> kplot(pta(kta1,scan=F),,clab=1.5,csub=3)
```



La version analyse des correspondances dite "AFC de Foucart" (Foucart 1978, 1983, 1984) n'a pas un support aussi canonique (du fait de la nom permanence des pondérations) mais conserve cette facilité d'interprétation pour un objectif cependant relativement complexe.

Le jeu de données utilisées est proposé par J. Blondel et H. Farré (1988). Il illustre un des problèmes fondamentaux de l'écologie factorielle. En confrontant un cortège faunistique à un paramètre de structure de l'habitat, on définit la notion de profil écologique ou de niche écologique. Quand on recommence la même opération à une

autre date ou dans une autre région, la relation binaire faune-milieu devient une relation ternaire faune-milieu-région. Les données sont dans 'bf88' :

```
> data(bf88)
> names(bf88)
[1] "S1" "S2" "S3" "S4" "S5" "S6"
> is.list(bf88)
[1] TRUE
> lapply(bf88,dim)
$S1
[1] 79  4

$S2
[1] 79  4

$S3
[1] 79  4

$S4
[1] 79  4

$S5
[1] 79  4

$S6
[1] 79  4
```

Elles forment une liste de 6 data.frame. Il s'agit de mesurer la variabilité du cortège avifaunistique entre 4 régions (Pologne, Bourgogne, Provence et Corse), le long du gradient de fermeture de la végétation vu par six strates d'échantillonnage (1- végétation buissonnante basse (hauteur < 1 m) à 6- forêts de plus de 20 m de hauteur).

La relation entre cortège faunistique et architecture de la végétation définit dans chaque région une structure de tableau donc une analyse. Cette analyse est une analyse des correspondances, sans contestation (typologie d'espèces par leur courbe de réponse inter-strates, typologie de strates par leur profil spécifique). Dans une région, le tableau est de type espèces-strates. Il y a quatre tableaux de ce type, donc une structure moyenne et des divergences régionales autour de cette structure.

La relation entre cortège faunistique et zones biogéographiques définit pour chaque strate de végétation une structure de tableau donc une analyse. Cette analyse est une analyse des correspondances (typologie d'espèces par leur distribution géographique, typologie de régions par leur contenu spécifique). Dans une strate, le tableau est de type espèces-régions. Il y a six tableaux de ce type, donc une structure moyenne et des divergences, fonction de la végétation, autour de cette structure.

L'abondance d'une espèce dans chaque région et chaque strate définit un modèle de répartition, demandant l'analyse d'un tableau homogène par une analyse simplement centrée. Il y a 79 tableaux de ce type. Que signifierait la notion de modèle moyen ? Le plus simple est de se référer à l'espèce sans signification écologique, uniformément présente dans chaque strate et chaque région. On peut penser à une typologie de modèles (courbes de réponse bivariées).

Ces indications sont incitatives à une réflexion préliminaire dans l'étude des cubes de données. Il convient, en effet, de garder son calme, tant un cube de données peut supporter potentiellement d'approches statistiques. Comme nous allons le voir, l'intention peut conduire à des résultats radicalement différents, sans que la validité des opérations soit mise en cause. La première chose à faire est de distinguer ce qui relève de l'observation de ce qui relève de l'organisation de l'information. Ici, nous avons

deux effets fixes, à savoir la végétation et la région. On aurait pu étudier un autre facteur écologique, par exemple l'altitude, et un autre corpus biogéographique, par exemple plusieurs massifs montagneux. Deux des arêtes du cube de données sont l'expression de l'intention de l'observation. La troisième, au contraire n'est pas maîtrisée. C'est la liste des espèces observables ou observées. Son contenu est fourni par les écosystèmes étudiés.

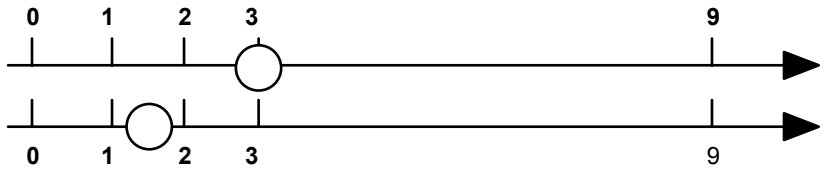
Lorsque les trois marges sont des effets fixes, par exemple mesure d'un paramètre dans 4 types de végétation, dans 3 classes d'altitude et dans 5 régions, les données forment un cube vrai. On peut désirer modéliser la variable x en fonction des 3 facteurs contrôlés, voire étudier les interactions ternaires : c'est le domaine des analyses à trois modes et plus (Kroonenberg 1983) ; ade4 ne contient aucune proposition dans ce domaine. Lorsque deux marges sont des effets fixes, il y a deux grands types d'objectifs. Le premier est celui des variables explicatives : construire un modèle de l'effet strate-régions pour chacune des espèces (effet simple A ou B, effet additif A+B, effet partiel A sachant B ou B sachant A, ...). Le second est celui de la comparaison de structures, c'est-à-dire de l'effet d'un facteur sur la structure engendré par l'autre.

T. Foucart part de la constatation qu'on peut aussi bien concevoir une table de contingence comme un tableau d'ACP particulière que comme une matrice de covariance particulière. Mais dans un cas comme dans l'autre, la question des pondérations interdit de généraliser STATIS (aussi bien sur les tableaux que sur les opérateurs). Il propose une opération qui n'a pas l'esthétique mathématique de STATIS mais qui est efficace. Il note :

> *Dans cet article, nous avons proposé des définitions de tendance et de structure susceptibles d'être utilisées dans l'étude des suites de tableaux de probabilités indexées par le temps. Si la technique simple d'analyse des évolutions des tendances repose effectivement sur la définition que nous en avons donnée, il n'en est malheureusement pas de même en ce qui concerne les techniques d'étude des évolutions de la structure : nous ne sommes pas partis des équivalences entre structures pour mettre au point les méthodes qui ont été décrites. Si ce manque de cohérence nuit à la qualité de notre exposé, il ne diminue en rien l'intérêt de ces équivalences et l'efficacité de ces méthodes.*

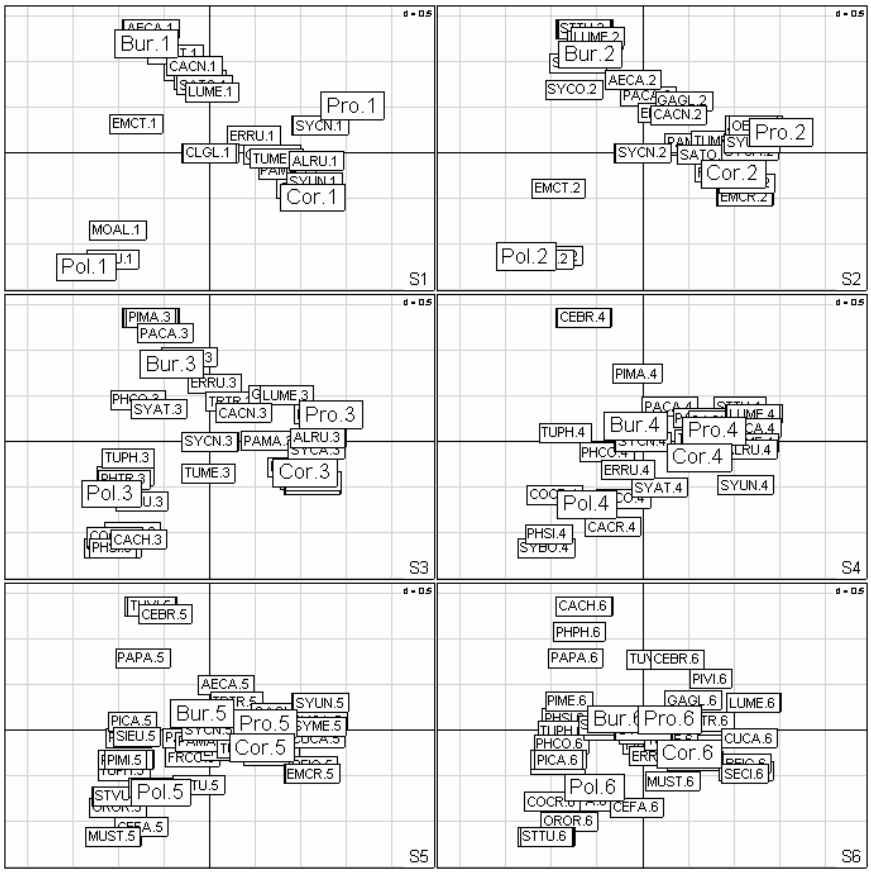
La difficulté vient de la présentation classique de STATIS : interstructure, compromis, intrastructure. En termes d'un seul tableau, tout se passe comme si on disait matrice de corrélation, moyenne, variance. L'interstructure définit une typologie de structure, le compromis définit une structure moyenne et l'intrastructure représente la variabilité autour de la moyenne (centrage). C'est comme si on voulait définir l'analyse en composantes principales avant la moyenne. Nous avons souligné que STATIS définit essentiellement la moyenne, et ce, de façon élaborée. En effet, on calcule généralement une moyenne en définissant au préalable les poids utilisés. Dans STATIS on définit une moyenne en calculant les poids pour que cette moyenne soit la meilleur possible.

Cela conduit à éliminer le rôle des points douteux. Par exemple, la moyenne à pondération uniforme des 5 valeurs ci-dessous vaut 3, mais si on considère que la cinquième valeur est bizarre, on dit que cette moyenne vaut 1.5. On fait évidemment cette opération, dans STATIS, sur les structures et non sur les valeurs. On fait effectivement cette opération sur des valeurs en ACP non centrée.



Foucart propose simplement de faire le compromis en prenant une moyenne uniformément pondérée des tableaux. Soient K tableaux d'AFC. Le $k^{\text{ème}}$ tableau a, comme les autres, I lignes et J colonnes. Son terme général est x_{ij}^k et la somme de toutes les valeurs est $x_{..}^k$. Le tableau de fréquences associé est $\mathbf{P}_k = \left[x_{ij}^k \middle/ x_{..}^k \right]$. La moyenne est $\mathbf{P} = (1/K) \sum_k \mathbf{P}_k$.

On fait l'analyse des correspondances de **P**, structure compromis utilisant une pondération uniforme et l'intrastructure consiste à projeter en individus supplémentaires les lignes et les colonnes des K tableaux de départ. Les pondérations de l'AFC du compromis servent de référence générale. Si on n'a aucune raison de pondérer inégalement les tableaux, cette analyse ne pose aucun problème de signification pour l'utilisateur.

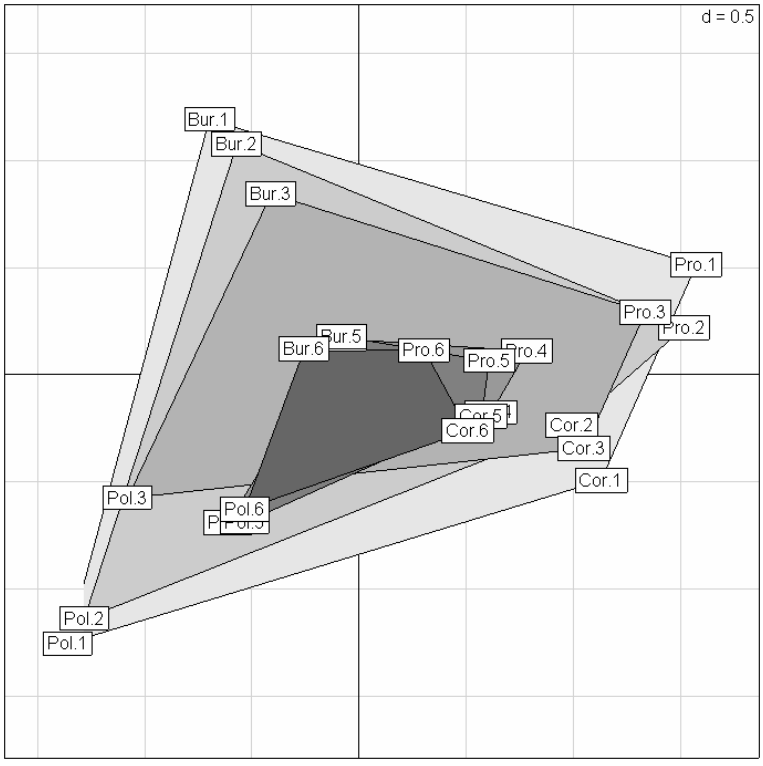



```
> fou1 <- foucart(bf88,scan=F,nf=3)
> fou1
Foucart's  COA
class: foucart coa dudi
$call: foucart(X = bf88, scannf = F, nf = 3)
table  number: 6

$nf: 3 axis-components saved
$rank: 3
eigen values: 0.5278 0.3591 0.3235
```

```
blo     vector           6             blocks
 vector length mode    content
 $cw    4      numeric column weights
 $lw    79     numeric row weights
 $eig   3      numeric eigen values

 data.frame nrow ncol content
 $tab       79   4    modified array
 $li        79   3    row coordinates
 $l1        79   3    row normed scores
 $co        4    3    column coordinates
 $c1        4    3    column normed scores

     **** Intrastructure ****

 data.frame nrow ncol content
 $Tli       474  3    row coordinates (each table)
 $Tco       24   3    col coordinates (each table)
 $TL        474  2    factors for Tli
 $TC        24   2    factors for Tco

> kplot(fou1,clab.r=1.5,clab.c=2.5) # ci-dessus
```

On remarquera qu'il faut conserver les 3 axes disponibles pour décrire la structure des tableaux à 4 colonnes. L'essentiel est dans :

```
> s.label(fou1$Tco)
> polygon(fou1$Tco[fou1$TC[,1]==1,1:2],col=grey(0.9))
> polygon(fou1$Tco[fou1$TC[,1]==2,1:2],col=grey(0.8))
> polygon(fou1$Tco[fou1$TC[,1]==3,1:2],col=grey(0.7))
> polygon(fou1$Tco[fou1$TC[,1]==4,1:2],col=grey(0.6))
> polygon(fou1$Tco[fou1$TC[,1]==5,1:2],col=grey(0.5))
> polygon(fou1$Tco[fou1$TC[,1]==6,1:2],col=grey(0.4))
> s.label(fou1$Tco,add.p=T)
```



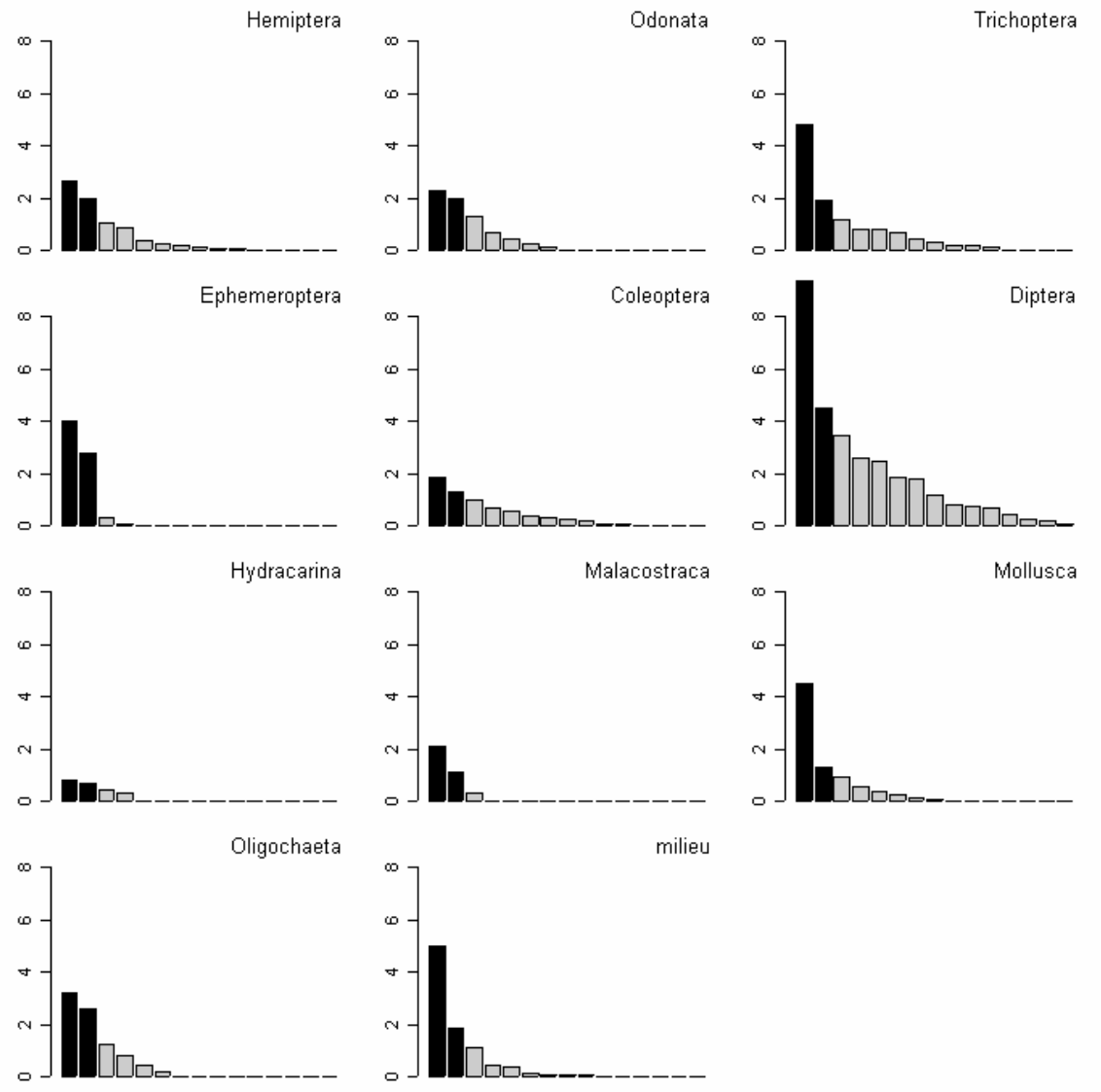
Il y a deux niveaux (1-3, 4-6, comme dans tous les exemples articulant ouverture de la végétation et avifaune) et la variabilité inter-régionale est forte et constante dans le second, faible et constante dans le premier. Il y a convergence des avifaunes des milieux forestiers. Quand les tableaux sont appariés par une seule dimension, lignes ou colonnes, la situation est moins favorable.

5. Compromis : calcul et analyse par STATIS

Nous avons vu que les tableaux sont appariés par les lignes mais qu'on utilise cette dimension pour parler aussi bien des individus (multiplicité des ensembles de descripteurs) que des variables (multiplicité des ensembles de points). Dans un cas comme dans l'autre, on a affaire à *K* schémas de dualité.

```
> data(friday87)
> names(friday87)
[1] "fau"       "mil"       "fau.blo"   "tab.names"
> w1 <-
cbind.data.frame(scale.wt(friday87$fau,scale=F),scale.wt(friday87$mil))
> dim(friday87$mil)
[1] 16  9
> kta1 <- ktab.data.frame(w1,c(friday87$fau.blo,9),
     tabnames=c(friday87$tab.names,"milieu"))
> sep1 <- sepan(kta1)
> plot(sep1)
```

Chaque tableau définit une structure.



Selon les critères habituels, la mesure de la structure d'un tableau est son inertie totale. Avant cela le nombre de variables est essentiel.

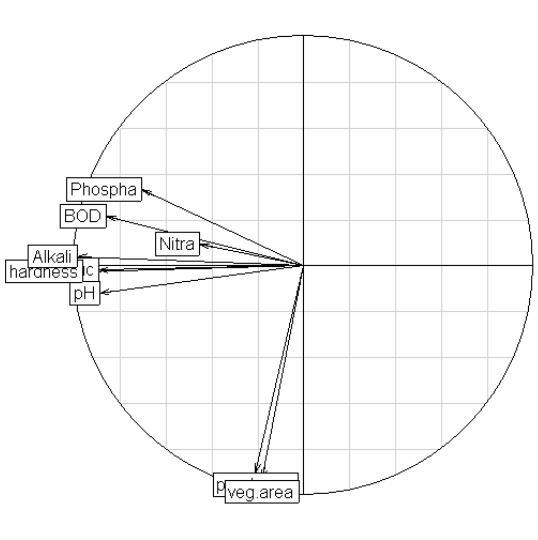
```
> kta1$blo
    Hemiptera       Odonata   Trichoptera Ephemeroptera    Coleoptera
           11             7            13             4            13
      Diptera   Hydracarina  Malacostraca      Mollusca   Oligochaeta
           22             4             3             8             6
       milieu
            9
```

Ces inerties sont :

```
> tapply(sep1$Eig,factor(rep(sep1$tab.names,sep1$rank)),sum)
   Coleoptera       Diptera Ephemeroptera     Hemiptera   Hydracarina
        6.582        30.391         7.230         7.516         2.230
 Malacostraca        milieu      Mollusca       Odonata   Oligochaeta
        3.547         9.000         8.055         6.996         8.426
  Trichoptera
       11.316
```

On reconnaît l'inertie de l'ACP normée qui est égale au nombre de variables et qui donne une structure simple à deux gradients (taille sur le facteur 2, charge minérale et organique sur le facteur 1) :

```
> s.corcircle(dudi.pca(friday87$mil,scan=F)$co)
```



On s'attend à ce qu'un groupe faunistique comportant de nombreuses espèces fournisse une inertie totale (somme des variances des abondances des taxons) plus grande. Il n'en est rien et la corrélation inertie-richesse est nulle. On peut dire qu'utilisés seuls, les groupes 1, 2, 5, 7, 8 et 10 n'auraient pas grand chose à dire sur une éventuelle typologie de relevés. On garderait pour les groupes 3, 6 et 9 le premier axe et pour le seul groupe 4 les deux premiers. Curieusement il n'y a que 4 espèces éphéméroptères et c'est le tableau qui semble le plus structuré.

STATIS propose de mesurer la valeur d'une analyse non par la somme des valeurs propres d'une analyse élémentaire (qui vaut l'inertie totale) mais par la somme des carrés de ces valeurs propres (carré de la norme d'Hilbert-Schmidt de l'opérateur d'inertie dit opérateur d'Escoufier associé à cette analyse). Il s'agit d'abord dans cette modification d'une conséquence de la logique algébrique sous-jacente. En fait, la signification écologique de cette innovation ne saurait échapper à un ... écologue. En effet, on sait que plus les valeurs propres d'une analyse sont différentes (c'est-à-dire, plus les premières valeurs propres sont grandes et les suivantes petites), meilleure est l'analyse (au sens que l'expression issue de la projection ne peut être un artefact). En passant de la somme (ou la moyenne) à la somme des carrés, on passe de la moyenne à

la variance, de l'abondance à la diversité. Plus l'inertie exprimée sur les premiers axes est grande, pour une valeur d'inertie donnée, plus la norme de l'opérateur est grande. Le groupe des éphéméroptères l'emporte sans conteste et c'est un autre groupe pauvre qui prend la seconde valeur.

```
> nvar <- kta1$blo[levels(factor(rep(sep1$tab.names,sep1$rank)))]
> inertia <- tapply(sep1$Eig,factor(rep(sep1$tab.names,sep1$rank)),sum)
> Vav <- tapply(sep1$Eig,factor(rep(sep1$tab.names,sep1$rank)),
      function(x) sum(x*x))

> cbind(nvar,inertia,Vav)
              nvar inertia     Vav
Coleoptera      13   6.582   7.221
Diptera         22  30.391 143.100
Ephemeroptera    4   7.230  24.105
Hemiptera       11   7.516  12.970
Hydracarina      4   2.230   1.378
Malacostraca     3   3.547   5.787
milieu           9   9.000  29.920
Mollusca         8   8.055  23.173
Odonata          7   6.996  11.355
Oligochaeta      6   8.426  19.299
Trichoptera     13  11.316  30.326
```

On retiendra que pour le $k^{\text{ième}}$ triplet statistique $\left(\mathbf{X}_k, \mathbf{D}, \mathbf{D}_k\right)$, on substitue à la notion habituelle d'inertie :

$$I_t = Trace\left(\mathbf{X}_k^t \mathbf{D}_k \mathbf{X}_k \mathbf{D}\right) = \sum_i \lambda_i$$

celle de variance vectorielle ou norme d'opérateur (Escoufier 1973, op. cit.) :

$$Vav(\mathbf{X}_k) = Trace\left(\mathbf{X}_k^t \mathbf{D}_k \mathbf{X}_k \mathbf{D} \mathbf{X}_k^t \mathbf{D}_k \mathbf{X}_k \mathbf{D}\right) = \sum_i \lambda_i^2$$

A la notion de mesure de structure, on ajoute la mesure de co-structure par le produit scalaire utilisé dans la norme précédente :

$$Covv(\mathbf{X}_k, \mathbf{X}_j) = Trace\left(\mathbf{X}_k^t \mathbf{D}_k \mathbf{X}_k \mathbf{D} \mathbf{X}_j^t \mathbf{D}_j \mathbf{X}_j \mathbf{D}\right)$$

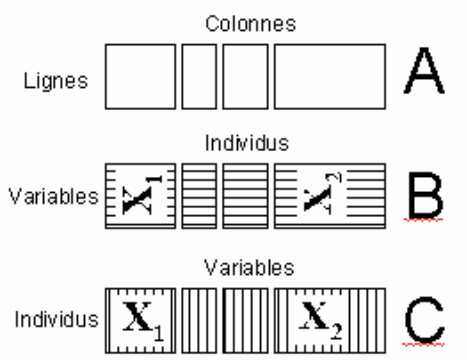
$$= Trace\left(\mathbf{X}_j \mathbf{D} \mathbf{X}_k^t \mathbf{D}_k \mathbf{X}_k \mathbf{D} \mathbf{X}_j^t \mathbf{D}_j\right) = \sum_i \mu_i$$

Il s'agit exactement de la co-inertie totale associée au couple de tableaux qui se décompose dans la somme des valeurs propres de l'analyse de co-inertie.

Il s'en suit qu'on peut mesurer la corrélation entre deux triplets par le coefficient de corrélation vectoriel :

$$RV(\mathbf{X}_k, \mathbf{X}_j) = \frac{Covv(\mathbf{X}_k, \mathbf{X}_j)}{\sqrt{Vav(\mathbf{X}_k)}\sqrt{Vav(\mathbf{X}_j)}}$$

Nous ne ferons pas la distinction habituellement pratiquée entre les opérateurs **VQ** et les opérateurs **WD** de la théorie des opérateurs d'Escoufier. Le programme 'statis' d'ade4 recouvre les deux cas de manière identique.



Dans la situation C, on considère deux tableaux d'ACP centrée appariés par les lignes-individus. On possède alors deux matrices comparables de produits scalaires $\mathbf{W}_1$ et $\mathbf{W}_2$ et les éléments qui précèdent s'écrivent :

$$Vav(\mathbf{X}_1) = Trace\left(\mathbf{X}_1\mathbf{D}_1\mathbf{X}_1^t\mathbf{D}\mathbf{X}_1\mathbf{D}_1\mathbf{X}_1^t\mathbf{D}\right) = Trace\left(\mathbf{W}_1\,\mathbf{D}\mathbf{W}_1\,\mathbf{D}\right)$$
$$Vav(\mathbf{X}_2) = Trace\left(\mathbf{X}_2\mathbf{D}_2\mathbf{X}_2^t\mathbf{D}\mathbf{X}_2\mathbf{D}_2\mathbf{X}_2^t\mathbf{D}\right) = Trace\left(\mathbf{W}_2\mathbf{D}\mathbf{W}_2\,\mathbf{D}\right)$$
$$Covv(\mathbf{X}_1,\mathbf{X}_2) = Trace\left(\mathbf{X}_1\mathbf{D}_1\mathbf{X}_1^t\mathbf{D}\mathbf{X}_2\mathbf{D}_2\mathbf{X}_2^t\mathbf{D}\right) = Trace\left(\mathbf{W}_1\,\mathbf{D}\mathbf{W}_2\,\mathbf{D}\right)$$

Dans la situation B, le centrage est opéré sur les lignes-variables et ce qui est comparable entre les deux tableaux concerne les matrices de covariances $\mathbf{V}_1$ et $\mathbf{V}_2$, en écrivant

$$Vav(\mathbf{X}_1) = Trace\left(\mathbf{X}_1^t\mathbf{D}_1\mathbf{X}_1\mathbf{D}\mathbf{X}_1^t\mathbf{D}_1\mathbf{X}_1\mathbf{D}\right) = Trace\left(\mathbf{V}_1\,\mathbf{D}\mathbf{V}_1\,\mathbf{D}\right)$$
$$Vav(\mathbf{X}_2) = Trace\left(\mathbf{X}_2^t\mathbf{D}_2\mathbf{X}_2\mathbf{D}\mathbf{X}_2^t\mathbf{D}_2\mathbf{X}_2\mathbf{D}\right) = Trace\left(\mathbf{V}_2\,\mathbf{D}\mathbf{V}_2\,\mathbf{D}\right)$$
$$Covv(\mathbf{X}_1,\mathbf{X}_2) = Trace\left(\mathbf{X}_1^t\mathbf{D}_1\mathbf{X}_1\mathbf{D}\mathbf{X}_2^t\mathbf{D}_2\mathbf{X}_2\mathbf{D}\right) = Trace\left(\mathbf{V}_1\,\mathbf{D}\mathbf{V}_2\,\mathbf{D}\right)$$

On a deux écritures mathématiques en voulant conserver l'identité lignes-individus et colonnes-variables pour le même calcul. On n'en conservera qu'une tout en gardant à l'esprit que dans l'une on compare des matrices de covariances (qui exprime de la ressemblance entre variables) alors que dans l'autre on compare des matrices de produits scalaires (qui exprime de la différence entre individus)

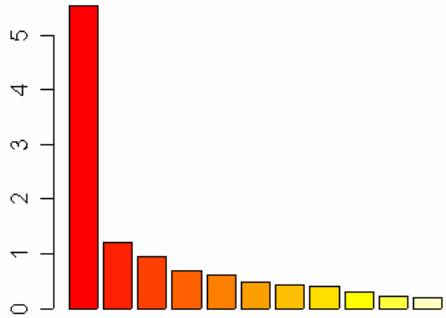
Chaque groupe de variables fait ici une typologie d'individus et STATIS mesure la corrélation entre ces typologies :

```
> statis1 <- statis(kta1)
Select the number of axes: 2

> statis1
STATIS Analysis
class:statis
table number: 11
row number: 16  total column number: 100

     **** Interstructure ****

eigen values: 5.527 1.202 0.943 0.6993 0.5999 ...
 $RV            matrix          11      11     RV coefficients
 $RV.eig        vector          11             eigenvalues
 $RV.coo        data.frame      11      4      array scores
 $tab.names     vector          11             array names
 $RV.tabw       vector          11             array weigths

RV coefficient
```

```
             Hemiptera Odonata Trichoptera Ephemeroptera Coleoptera Diptera
Hemiptera       1.0000
Odonata         0.4417  1.0000
Trichoptera     0.5271  0.5095      1.0000
Ephemeroptera   0.4391  0.5713      0.5435        1.0000
Coleoptera      0.4984  0.4062      0.4632        0.3054     1.0000
Diptera         0.4278  0.6396      0.6145        0.6239     0.4510  1.0000
Hydracarina     0.5022  0.3068      0.4327        0.3162     0.4726  0.3385
Malacostraca    0.3475  0.4323      0.5100        0.5965     0.3101  0.4944
Mollusca        0.4066  0.4915      0.4243        0.6104     0.4962  0.5282
Oligochaeta     0.3888  0.4105      0.4417        0.4074     0.2540  0.5138
milieu          0.3281  0.4078      0.4877        0.4444     0.2999  0.6658
             Hydracarina Malacostraca Mollusca Oligochaeta milieu
Hemiptera
Odonata
Trichoptera
Ephemeroptera
Coleoptera
Diptera
Hydracarina        1.0000
Malacostraca       0.4177       1.0000
Mollusca           0.5943       0.6381   1.0000
Oligochaeta        0.2839       0.3351   0.2417      1.0000
milieu             0.3143       0.3077   0.3736      0.6473      1
```

La matrice des RV est diagonalisée. On obtient une image euclidienne des tableaux pour ce produit scalaire. Mais contrairement au cas précédent, les RV sont toujours positifs ou nuls et le premier vecteur propre définit toujours une pondération des tableaux.

```
> barplot(statis1$RV.eig)
```



Cette forme est naturelle et dérive directement du caractère positif des RV. Si la première valeur propre n'était pas nettement dominante, on aurait l'indication d'une absence de compromis et donc d'une analyse sans objet. Les opérateurs d'inertie ont été normés avant d'en faire une combinaison linéaire. C'est un choix simplifié qui élimine le rôle en général écrasant des opérateurs "volumineux". On a donc débarrassé chaque tableau de son inertie et de son importance typologique en prenant comme compromis :

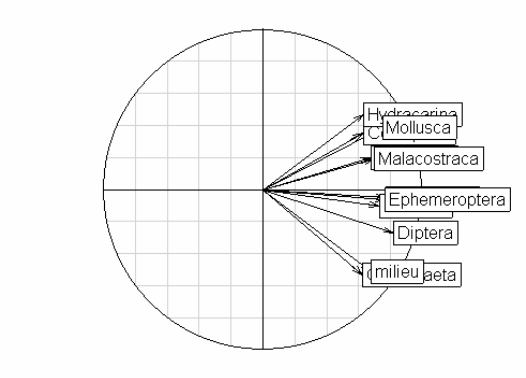
$$\mathbf{WD} = \sum_{k=1}^{K} \alpha_k \frac{\mathbf{W}_k\mathbf{D}}{\|\mathbf{W}_k\mathbf{D}\|}$$

```
> statis1$RV.tabw
 [1] 0.2872 0.3104 0.3280 0.3255 0.2669 0.3491 0.2669 0.2968 0.3190 0.2656
[11] 0.2881
```

Chaque tableau, quelle que soit sa structure initiale, participe à égalité à la constitution du compromis dont la diagonalisation fait une typologie moyenne.

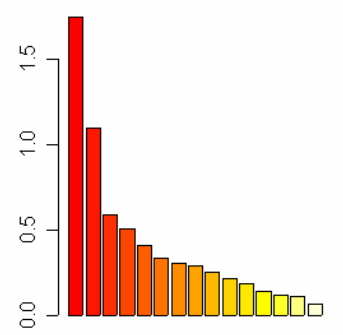
```
> s.corcircle(statis1$RV.coo)
```

La cohérence des tableaux n'est pas très grande :



Par contre le compromis est franchement de dimension 2. Le tableau de milieu ne tient pas une position centrale, ce qui laisse à penser que seule une partie de la typologie commune sera prise en compte.

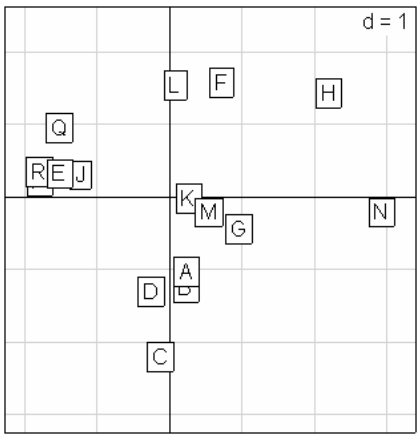
```
> barplot(statis1$C.eig)
```



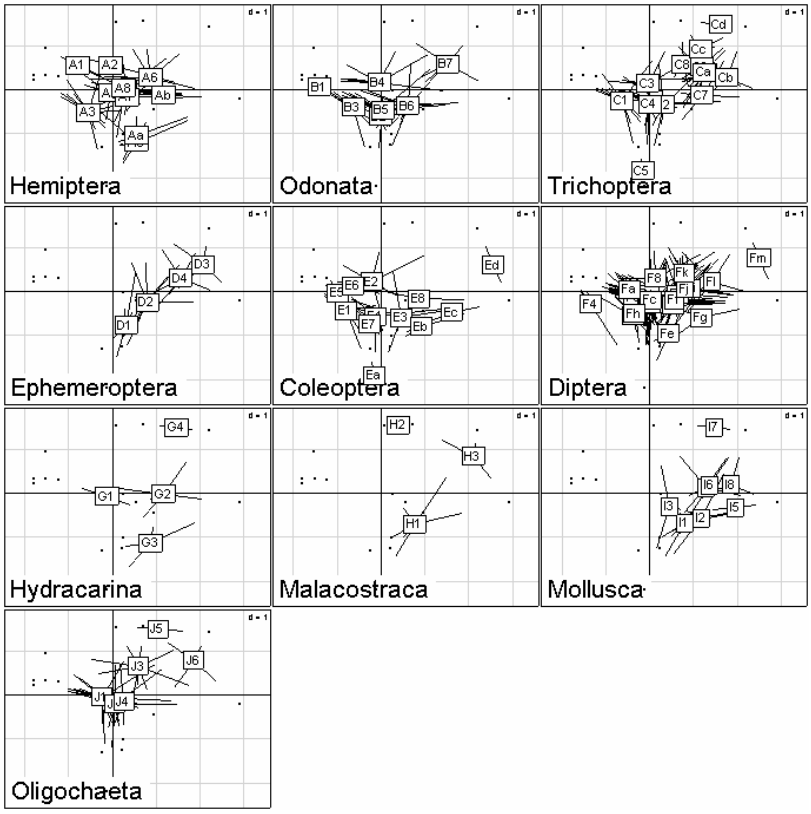

```
      **** Compromise ****

eigen values: 1.744 1.099 0.5934 0.512 0.4096 ...

 $nf: 2 axis-components saved
 $rank: 15
 data.frame nrow ncol content
 $C.li      16   2    row coordinates
 $C.Co      100  2    column coordinates
 $T4        44   2    principal vectors (each table)
 $TL        176  2    factors (not used)
 $TC        100  2    factors for Co
 $T4        44   2    factors for T4
```

L'analyse du compromis donne une typologie des stations :

```
s.label(statis1$C.li)
```

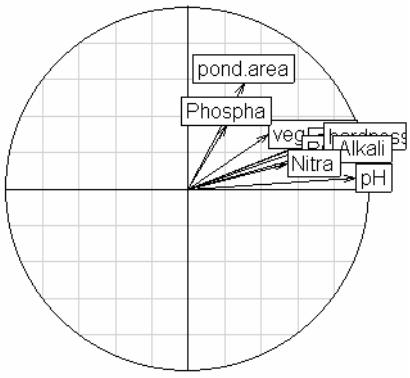


```
kta2 <- ktab.data.frame(friday87$fau,friday87$fau.blo)
par(mfrow=c(4,3))
for (j in 1:10) {
   s.distri(statis1$C.li,kta2[[j]],cell=0,axesell=F,cstar=.5
   ,clab=1.5,sub=names(kta2)[[j]],csub=3)}
```



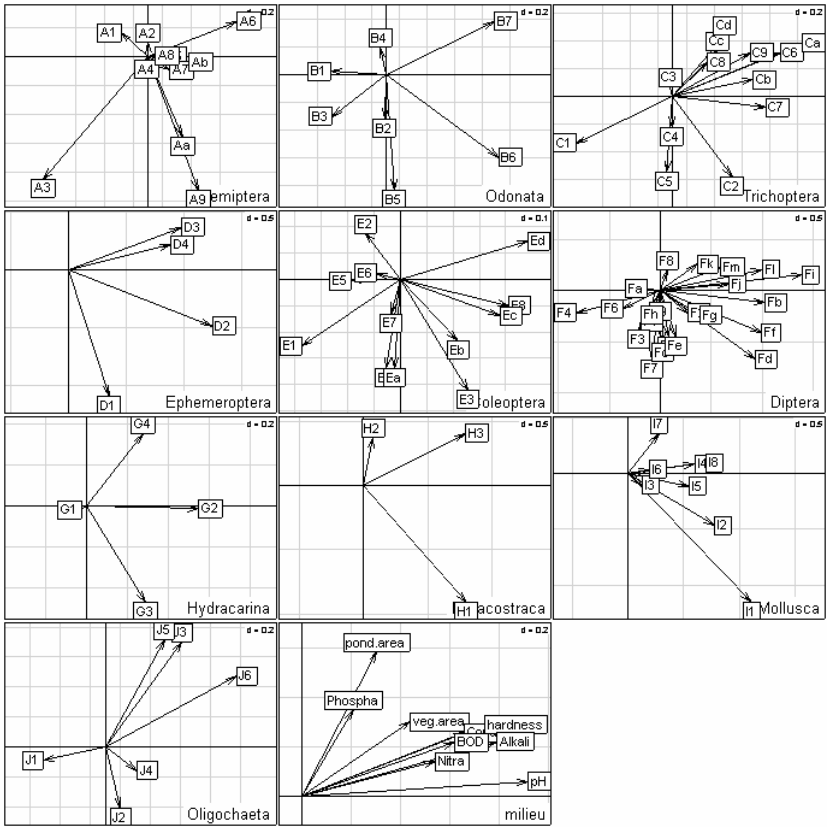
Les trichoptères, les éphéméroptères et les quatre derniers groupes sont largement absents des stations de gauche, alors que les hémiptères, les coléoptères, les diptères (dans une mesure moindre) ne présentent pas cet effet. En fait on a là l'effet conjugué de deux éléments : le nombre d'espèces augmente avec la taille du lac, d'une part, l'acidité des eaux étant le facteur limitant d'autre part :

```
> s.corcircle(statis1$C.Co[statis1$TC[,1]==11,])
```

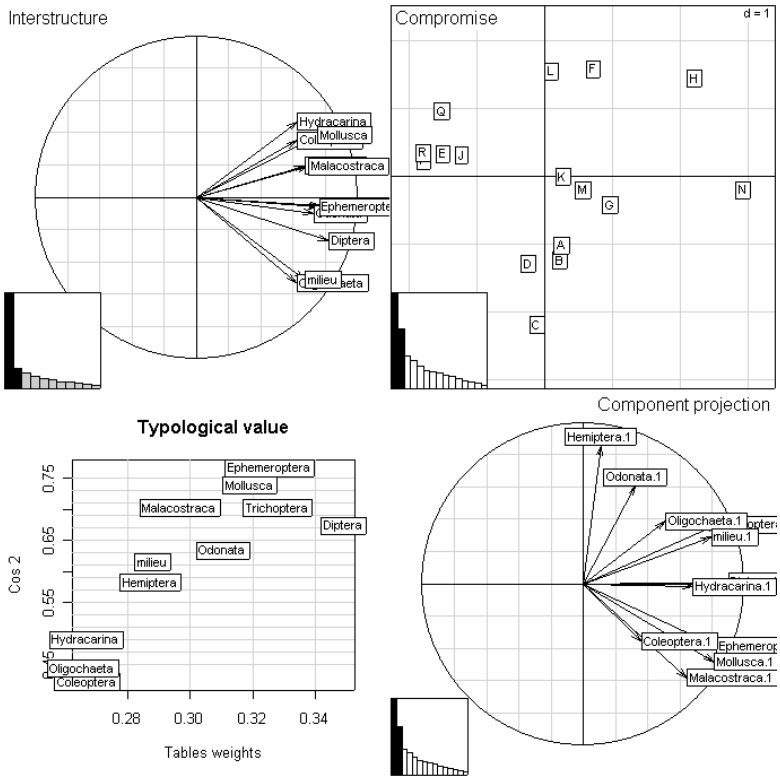


On retrouvera tous ces éléments dans les graphes génériques de 'statis' :

```
> kplot(statis1)
```




```
> plot(statis1)
```

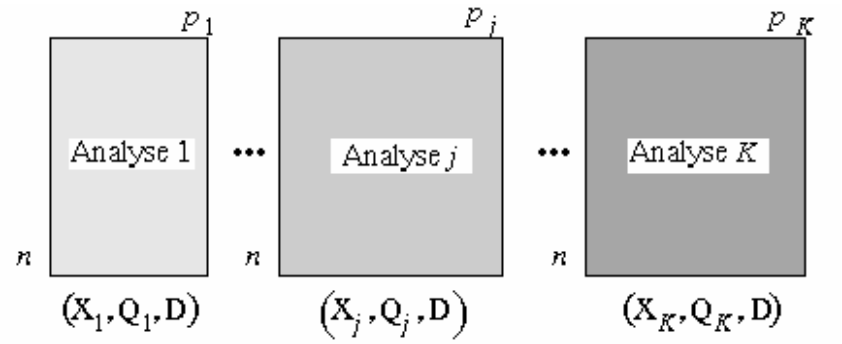


Un facteur écologique limitant (pH) est omniprésent dans cette observation, son rôle suivant les groupes d'espèces varie fortement. Foster (1995) apporte des arguments très cohérents avec ces résultats.

Nous n'avons pas introduit la représentation des trajectoires (chaque individu dans chaque tableau) car l'analyse factorielle multiple fait sensiblement mieux en la matière.

6. Compromis : sa constitution dans l'AFM

Contrairement à STATIS les auteurs de l'Analyse Factorielle Multiple n'ont envisagé que le cas des tableaux appariés par les individus alors que la théorie ne semble pas imposer ce point de vue unique. C'est cependant le cas de l'exemple en cours. On dispose de *K* tableaux ayant en commun les lignes-individus, chacun d'entre eux correspondant à un groupe de variables-colonnes. On dispose donc de *K* triplets statistiques $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ $(1 \leq k \leq K)$:



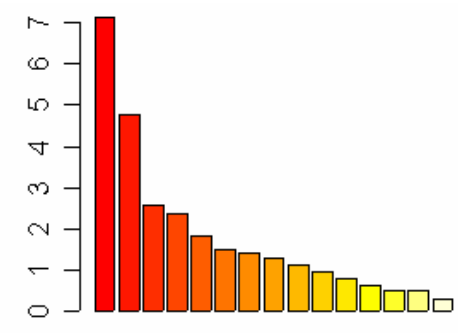
L'objectif est d'ordonner simultanément *K* tableaux, de réaliser *K* analyses portant sur un même ensemble d'individus, ou encore de comparer *K* groupes de variables définis sur le même ensemble d'individus comme indiqué dans le premier texte présentant la méthode (Escofier and Pagès 1982a, b).

L'AFM est basée sur l'analyse du tableau accolé $[\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K]$ avec les pondérations lignes et colonnes commune (pour les lignes) et concaténées (pour les colonnes). Mais une opération préliminaire essentielle est introduite pour uniformiser le rôle des tableaux dans l'analyse simultanée. On multiplie chaque tableau par un poids qui diminue l'importance des grands tableaux et augmente celle des petits.

```
> args(mfa)
function (X, option = c("lambda1", "inertia", "uniform", "internal"),
    scannf = TRUE, nf = 3)
```

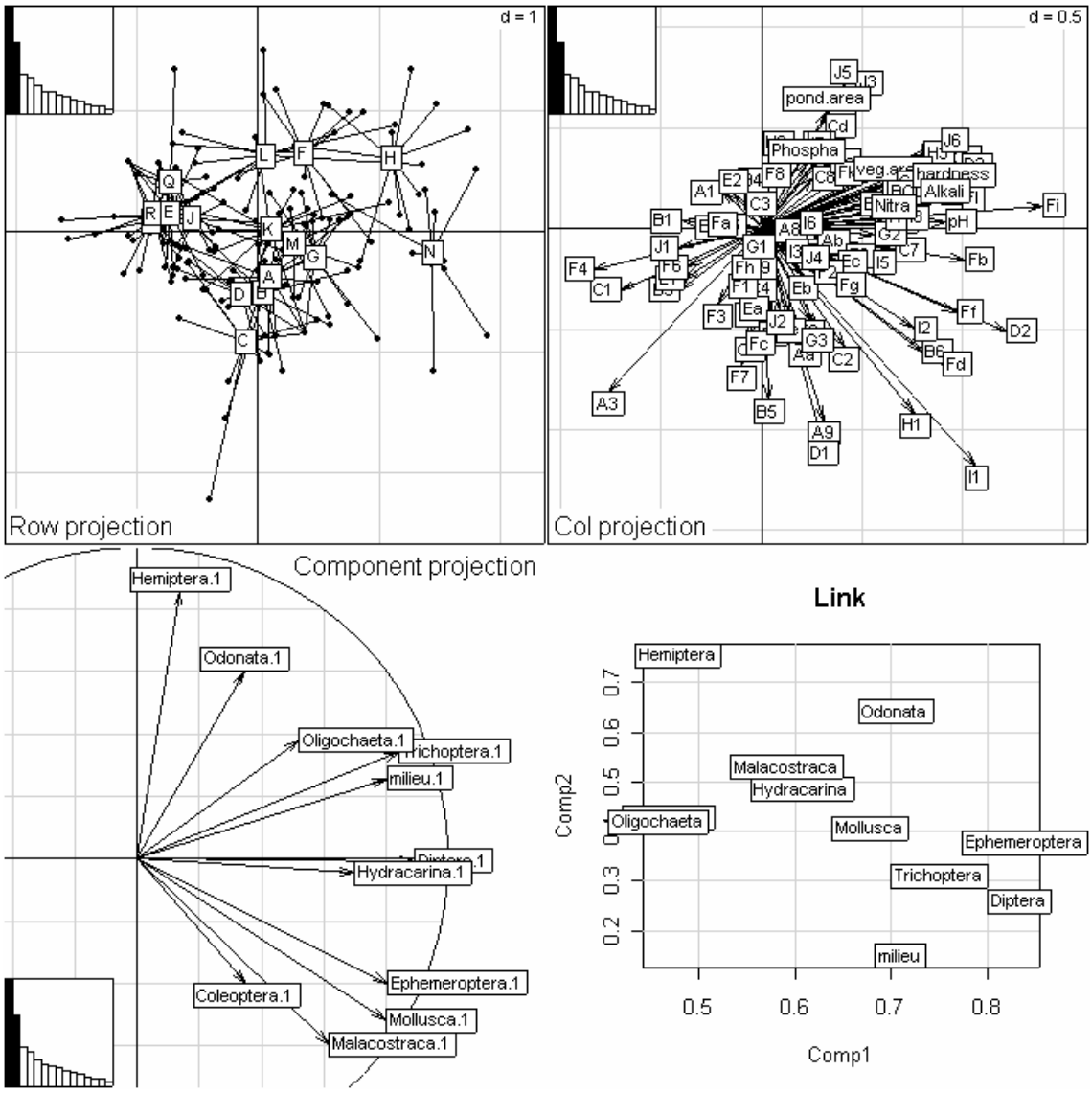
Ce poids est par défaut l'inverse de la première valeur propre ("lambda1") ou si on préfère l'inverse de l'inertie du tableau ("inertia"), ou encore un poids uniforme ("uniform") si on pense que cette opération ne doit pas avoir lieu ou enfin un poids prédéfini dans le *K*-tableaux par une composante 'tabw' ("internal"). Dans le premier cas, les tableaux juxtaposés et pondérés ont une même première valeur propre égale à 1, dans le second cas ils ont tous même inertie totale égale à 1.

```
> afm1 <- mfa(kta1)
Select the number of axes: 2
```



On retrouve d'entrée la dimension 2 du compromis de STATIS.

```
> plot(afm1)
```



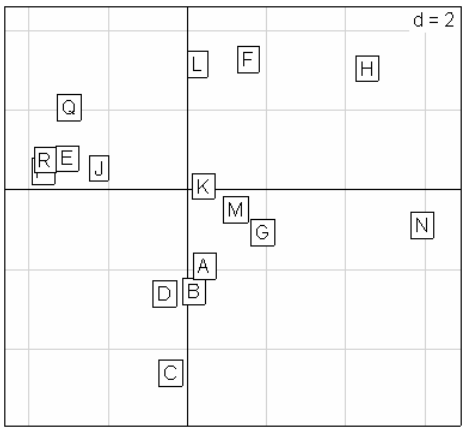
Comme on fait une analyse avec n individus et $p_1 + p_2 + ... + p_k$ variables, on a directement une carte des variables (en haut à droite) et un graphe de valeurs propres. Sur les composantes principales, on peut projeter les composantes principales de chaque tableau (ce qui était fait dans STATIS) et on observe (en bas à gauche) que dans $\mathbb{R}^n$, STATIS a construit un compromis pour l'analyser alors que l'AFM a trouvé des composantes principales pour constituer un compromis mais que les deux ont trouvé des plans très voisins.

Deux différences, à partir de ces points communs, sont sensibles.

La première introduit la représentation simultanée d'un point et de chacune de ses réalisation par tableaux.

Nous savons qu'en tant qu'analyse d'un triplet particulier l'AFMULT renvoie à une analyse duale des individus. Un individu est une ligne du tableau juxtaposant les *K* tableaux. Les lignes du tableau *k* sont dans l'espace $\mathbb{R}^{p_k}$ et seule l'ACOM s'intéresse directement à ces *K* nuages de points. Les lignes du tableau global sont dans $\mathbb{R}^p$ où $p = p_1 + p_2 + ... + p_K$. Le nuage des *n* points de $\mathbb{R}^p$ donne la carte ordinaire.

```
> s.label(afm1$li)
```



Sa similitude avec celle de STATIS est frappante alors que leurs calculs sont radicalement différents. C'est lié à une similitude profonde d'objectifs et à l'unité de la géométrie euclidienne.

La question est de savoir si l'on peut faire une carte des stations par tableau et examiner la cohérence des positions d'un point pour ses multiples représentations. La solution proposée par les auteurs de l'AFM est simple en théorie et remarquablement efficace.

En effet, elle utilise une propriété de base des espaces vectoriels de dimension finie. Une ligne du tableau global peut s'écrire comme somme de *K* vecteurs de $\mathbb{R}^p$.

Notons *p* le nombre total de variables, P l'ensemble des *p* premiers entiers, à savoir {1, 2, ..., *p*}. Notons $p_1, p_2, ..., p_K$ les effectifs des variables par tableau et $\mathrm{P}_1, \mathrm{P}_2, ..., \mathrm{P}_K$ les parties de P correspondant à chaque groupe. Cela permet d'écrire $j \in \mathrm{P}_k \ (1 \le j \le p)(1 \le k \le K)$ pour dire que la variable *j* dans la numérotation complète appartient au tableau *k*. $\mathbf{x}^j$ désigne la variable *j*, vecteur de $\mathbb{R}^n$ (*n* est le nombre d'individus), de poids initial m_j dans l'analyse globale (ici l'unité). $\mathbb{R}^n$ est muni du produit scalaire **D** associé à la pondération des individus.

Soit un vecteur $\mathbf{x} = (x_1, x_2, ..., x_p)$. Notons $\mathbf{x}[k] = (a_1, a_2, ..., a_p)$ le vecteur défini par ses composantes a_j et la règle $a_j = x_j$ si $j \in P_k$ et $a_j = 0$ sinon. Pour obtenir $\mathbf{x}[k]$, on part du vecteur **x** et on annule toutes les valeurs sauf celles des variables du groupe *k*. Il est évident que $\mathbf{x} = \mathbf{x}[1] + \mathbf{x}[2] + ... + \mathbf{x}[K]$. En partant du nuage des *n* points $\mathbf{x}_i$ (points définis par l'ensemble des variables) on obtient *K* nuages de *n* points du type $\mathbf{x}_i[k]$.

Ceci consiste à projeter sur le sous-espace engendré par les vecteurs de la base canonique de $\mathbb{R}^p$ de rang j tel que $j \in \mathrm{P}_k \ (1 \le j \le p)(1 \le k \le K)$.

Cette opération immerge les *K* nuages distincts dans un même espace où se trouve déjà le nuage de référence. La somme $\mathbf{x}_i = \mathbf{x}_i[1] + \mathbf{x}_i[2] + ... + \mathbf{x}_i[K]$ se conserve par projection et au prix d'une dilatation sans importance se réécrit :

$$\mathbf{x}_i = K \frac{\mathbf{x}_i[1] + \mathbf{x}_i[2] + ... + \mathbf{x}_i[K]}{K}$$

Dès qu'on possède un vecteur **u** normé pour la métrique :

$$\mathbf{Q} = \begin{bmatrix} \alpha_1\mathbf{Q}_1 & 0 & & 0 \\ 0 & \alpha_2\mathbf{Q}_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \alpha_K\mathbf{Q}_K \end{bmatrix}$$

on peut projeter sur **u** les *K* nuages de *n* points et prendre pour chaque point le centre de gravité de ses *K* représentations. Quelle propriété utiliser pour choisir le vecteur **u** ?

On veut exprimer les différences entre points donc maximiser la variance des projections des centres de gravité. C'est exactement ce que fait l'analyse du triplet modifié de l'AFMULT. En effet :

$$\mathbf{u} = \mathbf{u}[1] + \mathbf{u}[2] + ... + \mathbf{u}[K]$$

La coordonnée de la projection de $\mathbf{x}_i[k]$ sur **u** vaut :

$$\mathbf{u}^t\mathbf{Q}\mathbf{x}_i[k] = \left(\mathbf{x}_i[k]\right)^t \mathbf{Q}\mathbf{u} = \alpha_k \mathbf{x}_i{}^t\mathbf{Q}_k\mathbf{u}[k]$$

La coordonnée moyenne des représentations sur **u** du point *i* vaut alors :

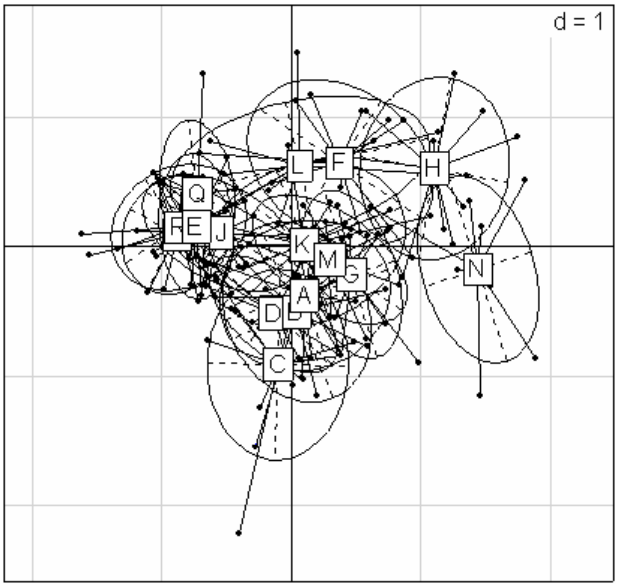
$$\mathbf{u}^t\mathbf{Q}\mathbf{x}_i[k] = \left(\mathbf{x}_i[k]\right)^t \mathbf{Q}\mathbf{u} = \frac{1}{K}\sum_{k=1}^{K} \alpha_k \mathbf{x}_i{}^t\mathbf{Q}_k\mathbf{u}[k] = \frac{1}{K}\mathbf{x}_i{}^t\mathbf{Q}\mathbf{u}$$

La variance de ces moyennes, ou variance inter est donc :

$$\frac{1}{K^2}\sum_{i=1}^{K} p_i \left(\mathbf{x}_i{}^t\mathbf{Q}\mathbf{u}\right)^2 = \frac{1}{K^2}\left\|\mathbf{X}\mathbf{Q}\mathbf{u}\right\|_{\mathbf{D}}^2$$

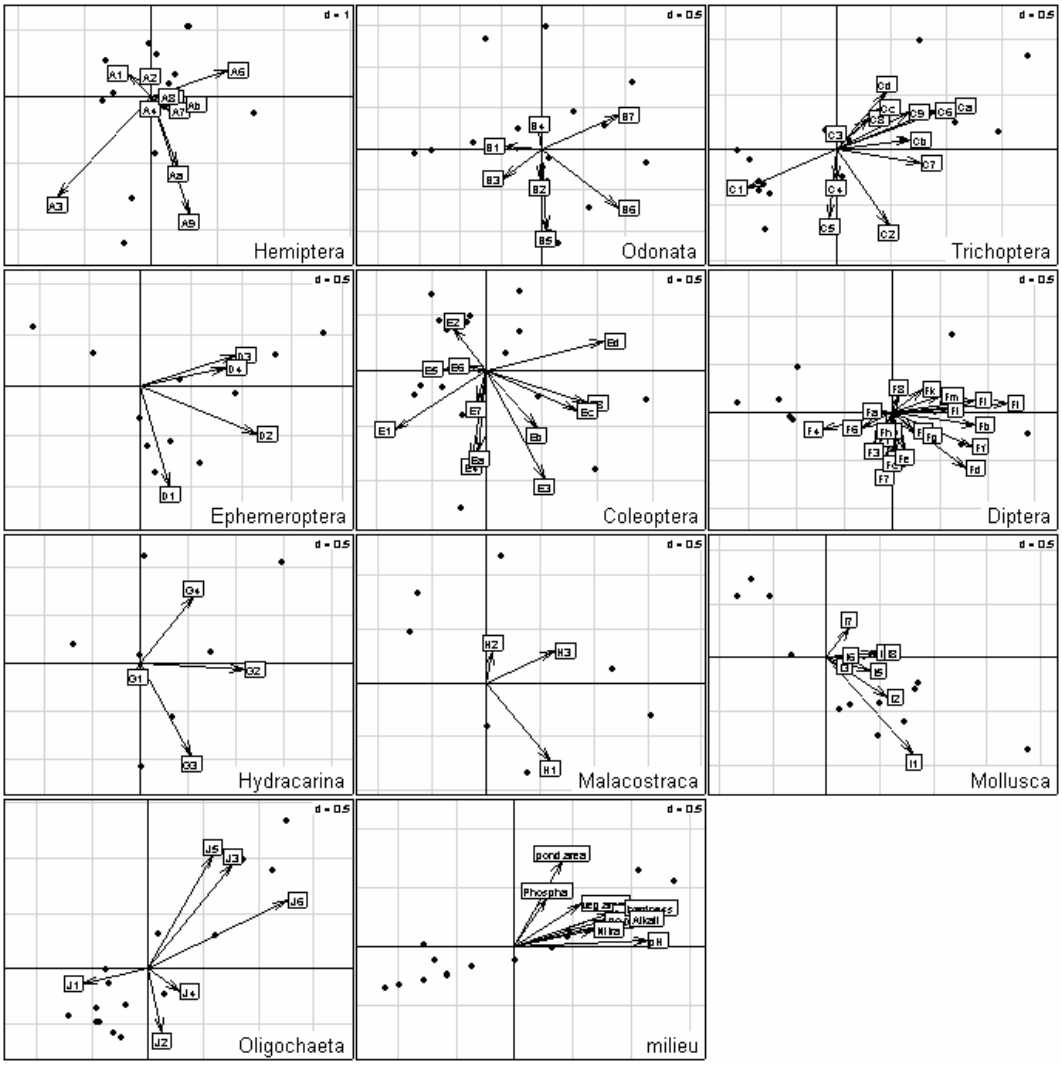
A une constante près, la solution est donnée directement par l'analyse du triplet de l'AFMULT (Escofier and Pagès 1984). Ceci donne la représentation simultanée.

```
> s.class(afm1$lisup,afm1$TL[,2],lab=row.names(afm1$tab))
```



On peut utiliser ces coordonnées par tableau soit pour les assembler (c'est la variance des positions moyennes qui est optimisée) et exprimer la cohérence des positions d'un même point, soit les dispatcher pour les représenter avec le nuage des variables correspondant pour observer ensemble les *K* analyses :

```
> kplot(afm1)
```



La compréhension du rôle de chacun des tableaux est renforcée.

La seconde différence est la mesure du lien entre un tableau et le compromis. Dans STATIS nous avons gardé le cosinus carré entre opérateurs et opérateur compromis réduit à ses axes conservés. Dans l'AFM c'est très différent.

Si **z** est un vecteur de $\mathbb{R}^n$ **D**-normé, l'inertie projetée associée aux p_k colonnes de $\mathbf{X}_k$ est :

$$I_k(\mathbf{z}) = \sum_{j \in P_k} p_j \left(\mathbf{x}^j \middle| \mathbf{z}\right)_{\mathbf{D}}^2 = \left\|\mathbf{X}_k^t \mathbf{D}\mathbf{z}\right\|_{\mathbf{Q}_k}^2 = \left(\mathbf{X}_k^t \mathbf{D}\mathbf{z} \middle| \mathbf{X}_k^t \mathbf{D}\mathbf{z}\right)_{\mathbf{Q}_k} = \left(\mathbf{X}_k \mathbf{D}\mathbf{X}_k^t \mathbf{Q}_k \mathbf{z} \middle| \mathbf{z}\right)_{\mathbf{D}} = \left(\mathbf{W}_k \mathbf{D}\mathbf{z} \middle| \mathbf{z}\right)_{\mathbf{D}}$$

On appellera valeur du lien entre le vecteur **z** et le tableau $\mathbf{X}_k$ cette inertie projetée ramenée au maximum potentiel, c'est-à-dire α_k. D'où la notation (Escofier and Pagès 1984, 1994) :

$$\mathsf{L}(\mathbf{z}, k) = \sum_{j \in P_k} p_j \alpha_k \left(\mathbf{x}^j \middle| \mathbf{z}\right)_{\mathbf{D}}^2 = \alpha_k \left\|\mathbf{X}_k^t \mathbf{D}\mathbf{z}\right\|_{\mathbf{Q}_k}^2 = \alpha_k \left(\mathbf{W}_k \mathbf{D}\mathbf{z} \middle| \mathbf{z}\right)_{\mathbf{D}}$$

Ce lien ne peut donc dépasser l'unité (pourcentage d'un optimum). La valeur globale du vecteur **z**, sur l'ensemble des nuages de variables (des tableaux) est la somme des liens, soit :

$$\mathsf{L}(\mathbf{z}) = \sum_{k=1}^{K} \mathsf{L}(\mathbf{z}, k) = \sum_{k=1}^{K} \left\langle \alpha_k \mathbf{W}_k \mathbf{D}\mathbf{z} \middle| \mathbf{z} \right\rangle_{\mathbf{D}} = \left\langle \left(\left(\sum_{k=1}^{K} \alpha_k \mathbf{W}_k \right) \mathbf{D}\mathbf{z} \middle| \mathbf{z} \right\rangle_{\mathbf{D}}$$

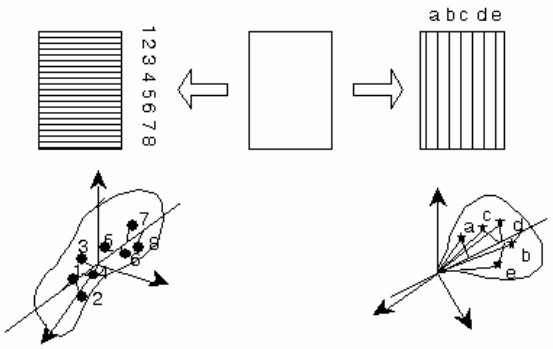
L'optimum est atteint pour la première composante principale normée du schéma surpondéré de l'AFMULT et cet optimum ne peut dépasser le nombre de tableau (somme de *k* valeurs inférieures ou égales à 1). On cherche ensuite un second vecteur **z**, orthogonal au précédent qui optimise la même quantité et on trouve la seconde composante de l'AFMULT. En cherchant un plan qui optimise l'inertie projetée, on trouve le sous-espace engendré par les deux premières composantes. Les liens sont représentés dans la figure de synthèse (p.41). Le milieu ne joue aucun rôle sur le deuxième composante principale qui est purement faunistique.

Cette analyse permet, en outre, de mélanger tableaux de variables quantitatives et qualitatives (Escofier and Pagès 1986). La procédure 'mfa' a cette généralité. Tout porte alors à croire la conclusion des auteurs (Escofier and Pagès 1989) :

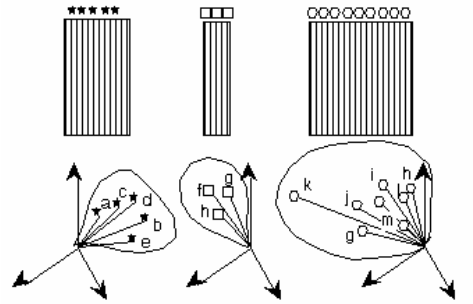
> *First applications of M.F.A. were highly encouraging as regards to the practical value of the method. ... The last example clearly shows that to take into account variable groups is not only a technical problem, susceptible of being solved by an appropriate method, but it also can be considered as a methodological problem which enriches the field of data Analysis.*

7. La co-inertie multiple

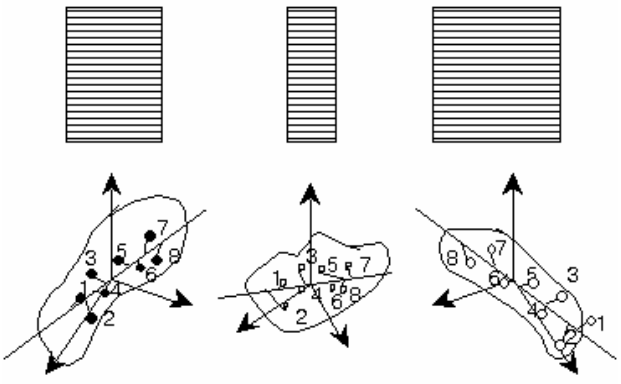
Très proche de la précédente, l'analyse de co-inertie multiple tra^^ite de la même situation. Un tableau individus-variables génère deux nuages de points :



K tableaux de données donnent donc *K* nuages de variables dans un même espace et ce fait est exploité par l'AFM :

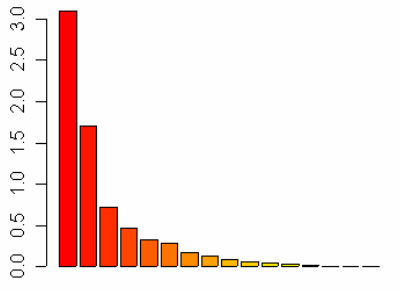


Ils donnent *K* opérateurs d'inertie dans un même espace et ce fait est exploité par STATIS. Ils donnent aussi *K* nuages de points appariés dans *K* espaces différents et ceci est pris en compte par l'ACOM :

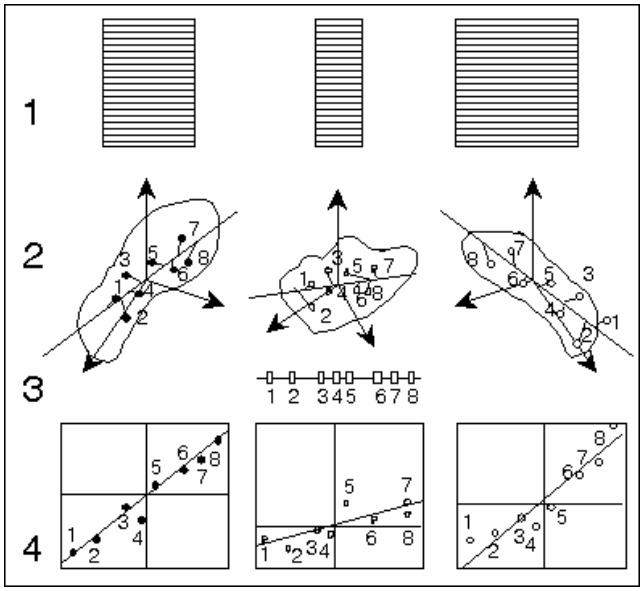


On voit qu'en partant des objets créés par les mêmes données plusieurs approches sont concurrentes. L'analyse linéaire d'un tableau a ceci de caractéristique : elle aborde les deux nuages de points dualement. C'est la base de son succès : tout résultat obtenu sur un des deux objets se retrouve exprimé directement sur l'autre et réciproquement. Mais dès qu'il y a plus d'un tableau, la symétrie lignes-colonnes est détruite et plusieurs voies sont ouvertes suivant qu'on aborde la question par l'une ou l'autre.

```
> acom1 <- mcoa(kta1)
Select the number of axes: 2
```



On a encore un compromis de dimension 2. Le principe de cette analyse se résume par :



1 — *K* tableaux ont les mêmes lignes.

2— Ils définissent *K* nuages de points dans *K* espaces euclidiens. Les points sont pondérés de la même manière dans chaque nuage. Dans chacun des espaces, on cherche un vecteur normé (axe) sur lequel on projette le nuage.

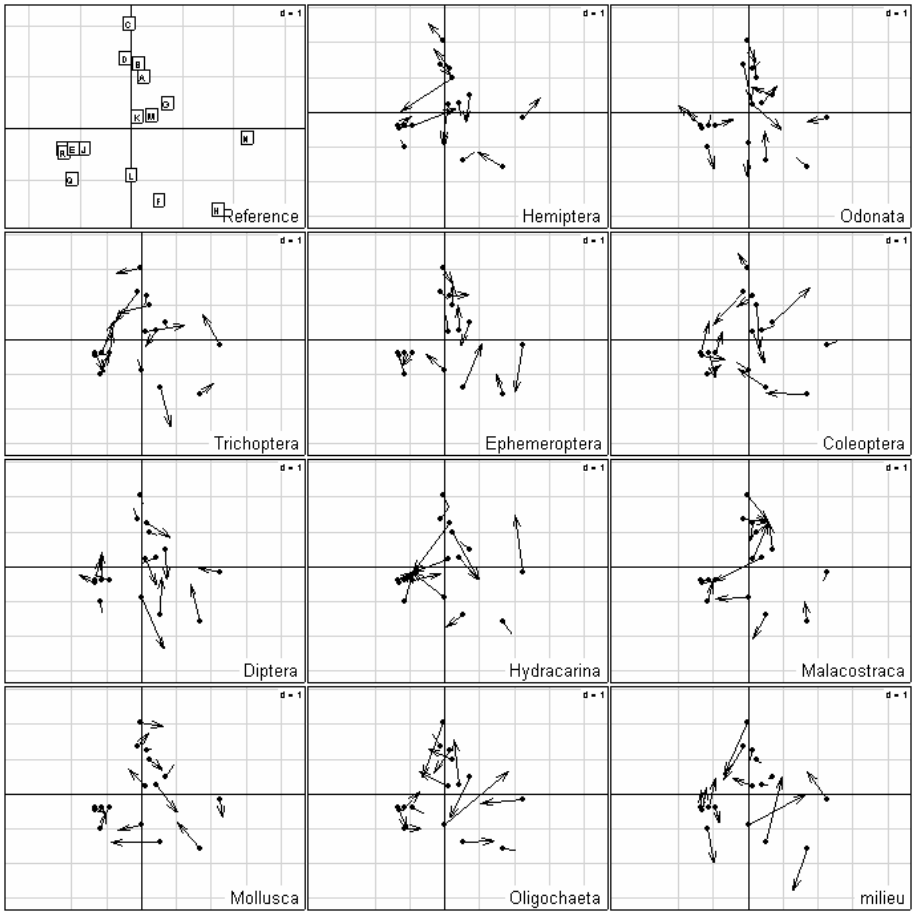
3 — On définit un code numérique de référence de variance unitaire.

4 — Les axes et le code de référence optimisent la somme pondérée des carrés des covariances entre le code de référence et les coordonnées de chaque projection. On recommence sous la contrainte d'orthogonalité sur les axes et sur les codes.

On fait donc, axe par axe, l'analyse d'inertie de chacun des tableaux en coordonnant les systèmes de coordonnées par des variables de synthèse (Chessel and Hanafi 1996). Le premier système est directement donné par la première composante de l'AFM. Ensuite on est plus précis sur la géométrie séparée des nuages *au prix de la perte de la représentation optimale des variables*.

Le 'kplot' de cette méthode exprime comment la projection sur un plan du nuages des lignes dans chaque espace reproduit la typologie de synthèse.

```
> kplot(acom1)
```



Comme cette voie est la seule qui permet de coupler un *K*-tableaux $[\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K]$ avec un tableau **Y**, c'est-à-dire de sortir le tableau du milieu pour piloter l'analyse simultanée des tableaux faunistiques, nous y reviendrons plus tard.

On retiendra que deux méthodes s'imposent par leur accessibilité et leur performance : c'est l'analyse triadique partielle pour les tableaux doublement appariés et l'analyse factorielle multiple pour les tableaux simplement appariés. Le concept de coefficients RV est fondamental pour qui voudra en outre faire simultanément l'analyse de co-inertie de *K* couples de tableaux $(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), ..., (\mathbf{X}_K, \mathbf{Y}_K)$ où les **X** portent sur les mêmes variables, les tableaux **Y** portent sur les mêmes variables et les couples sont appariés sur les mêmes individus.

Références

Blondel, J., and H. Farre. 1988. The convergent trajectories of bird communities along ecological successions in european forests. Œcologia (Berlin) **75**:83-93.

Chessel, D., and M. Hanafi. 1996. Analyses de la co-inertie de K nuages de points. Revue de Statistique Appliquée **44**:35-60.

Escofier, B., and J. Pagès. 1982a. Comparaison de groupes de variables définies sur le même ensemble d'individus. Rapport de recherche n°149, ISSN 0249-6399 INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le chesnay cedex, France.

Escofier, B., and J. Pagès. 1982b. Comparaison de groupes de variables. 2ème partie : un exemple d'applications. Rapport de recherche n°165 INRIA, Domaine de Voluceau-Rocquencourt, BP 105, 78153 Le chesnay cedex, France.

Escofier, B., and J. Pagès. 1984. L'analyse factorielle multiple : une méthode de comparaison de groupes de variables. Pages 41-55 *in* E. Diday and Coll., editors. Data Analysis and Informatics III. Elsevier, North-Holland.

Escofier, B., and J. Pagès. 1986. Le traitement des variables qualitatives et des tableaux mixtes par analyse factorielle multiple. Pages 179-191 *in* E. Diday and Coll., editors. Data Analysis and Informatics IV. Elsevier, North-Holland.

Escofier, B., and J. Pagès. 1989. Multiple factor analysis: results of a three-year utilization. Pages 277-285 *in* R. Coppi and S. Bolasco, editors. Multiway data analysis. Elsevier Science Publishers B.V., North-Holland.

Escofier, B., and J. Pagès. 1994. Multiple factor analysis (AFMULT package). Computational Statistics and Data Analysis **18**:121-140.

Foster, G. N. 1995. Evidence for pH insensitivity in some insects inhabiting peat pools in the loch fleet catchment. Chemistry and Ecology **9**:207-215.

Foucart, T. 1978. Sur les suites de tableaux de contingence indexés par le temps. Statistique et Analyse des données **2**:67-84.

Foucart, T. 1983. Une nouvelle approche de la méthode STATIS. Revue de Statistique Appliquée **31**:61-75.

Foucart, T. 1984. Analyse factorielle de tableaux multiples. Masson, Paris.

Friday, L. E. 1987. The diversity of macroinvertebrate and macrophyte communities in ponds. Freshwater Biology **18**:87-104.

Kiers, H. A. L. 1991. Hierarchical relations among three-way methods. Psychometrika **56**:449-470.

Kroonenberg, P. M. 1983. Three-mode principal component analysis. DSWO Press, Leiden.

Kroonenberg, P. M. 1989. The analysis of multiple tables in factorial ecology. III Three-mode principal component analysis:"analyse triadique complète". Acta Œcologica, Œcologia Generalis **10**:245-256.

Lavit, C. 1988. Analyse conjointe de tableaux quantitatifs. Masson, Paris.

Leibovici, B. 1993. Facteurs à mesures répétées et analyses factorielles : application à un suivi épidémiologique. Thèse de doctorat, Université Montpellier II.

Pegaz-Maucet, D. 1980. Impact d'une perturbation d'origine organique sur la dérive des macro-invertébérés benthiques d'un cours d'eau. Comparaison avec le benthos. Thèse de 3° cycle, Université Lyon 1.

Thioulouse, J., and D. Chessel. 1987. Les analyses multi-tableaux en écologie factorielle. I De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. Acta Œcologica, Œcologia Generalis **8**:463-480.

Tucker, L. R. 1966. Some mathemetical notes on three-mode factor analysis. Psychometrika **31**:279-311.

Verneaux, J. 1973. Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse d'état, Besançon.