M2 Écologie,Évolution,Biométrie 
UE Description Statistique des Structures Biologiques

De l'analyse multi-variables à l'analyse multi-tableaux

D. Chessel & A.-B. Dufour

Notes de cours cssb7

On pose la question des tableaux multiples à partir de jeux de données exemplaires. Les analyses élémentaires des cubes de données introduisent la notion de compromis.

Table des matières

1 Introduction

Quand on introduit une partition dans l'ensemble des descripteurs ou dans celui des conditions d'acquisition, on passe de l'analyse des données des tableaux ou des couples de tableaux à celle des K-tableaux. On trouvera dans la librairie `ade4` nombre d'exemples de ce type au premier rang desquels celui de L. Friday [11] décrit dans :

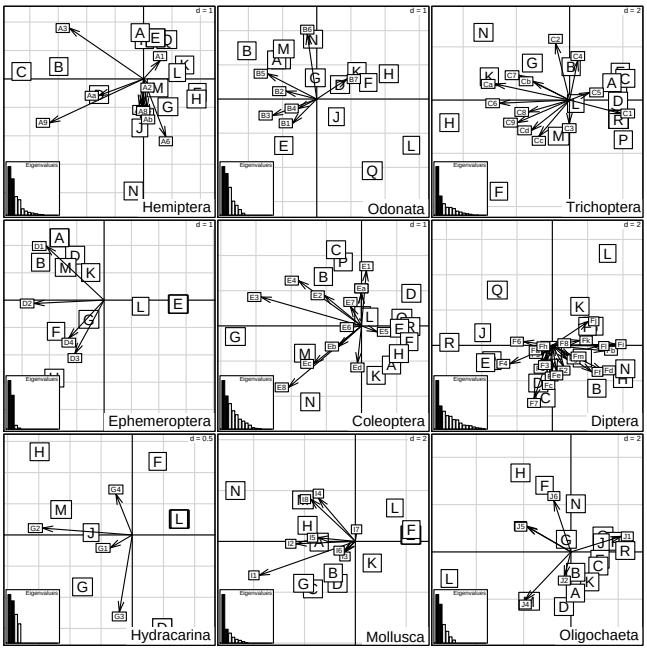
`http://pbil.univ-lyon1.fr/R/pps/pps078.pdf`

et celui de J. Verneaux [22] décrit dans :

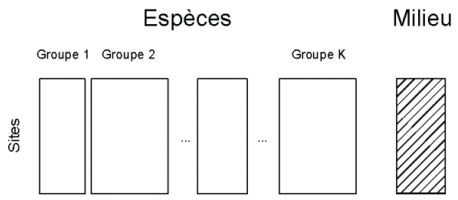
`http://pbil.univ-lyon1.fr/R/pps/pps047.pdf`

Le premier exemple représente la situation K-tableaux mêmes individus et la question de la valeur typologique des groupes d'espèces. Le second représente la situation K-tableaux mêmes variables et la question de la reproduction d'une même structure dans différentes conditions.

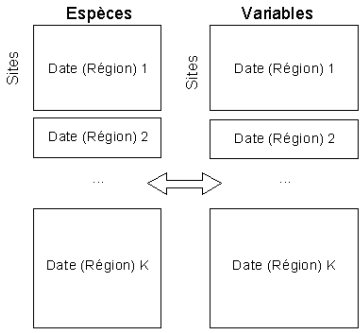
```
library(ade4)
data(friday87)
wfri <- data.frame(scale(friday87$fau[, -(75:77)], scal = FALSE))
wfri <- ktab.data.frame(wfri, friday87$fau.blo[-8], tabnames = friday87$tab.names[-8])
kplot(sepan(wfri), clab.r = 2, clab.c = 1)
```



Ces données sont exemplaires parce qu'elles représentent une multitude de situations écologiques, ce sont des exemples, et parce qu'elles sont parfaitement disponibles, ce sont des modèles. Les premières représentent la question $K+1$-tableaux :



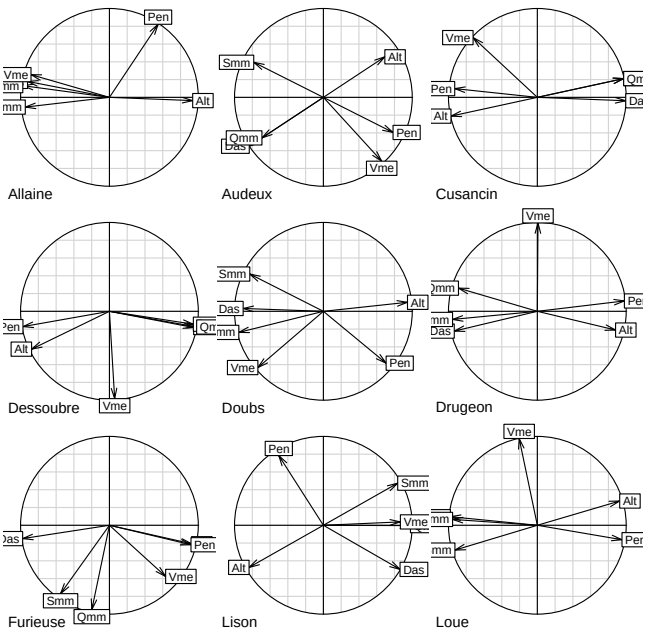
et la seconde la question $2K$-tableaux.



On s'intéressera donc pour commencer à la question des K-tableaux.

Ces situations indiquent qu'on manipule simultanément plusieurs analyses, par exemple 9 ACP centrées (ci-dessus) ou 9 ACP normées (ci-dessous).

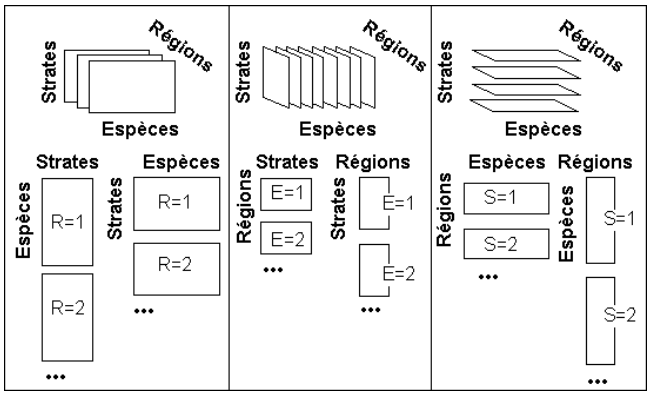
```
data(jv73)
w <- split(jv73$morpho, jv73$fac.riv)
w <- lapply(w, dudi.pca, scan = F)
par(mfrow = c(3, 3))
for (k in c(1:2, 5:7, 9:12)) s.corcircle(w[[k]]$co, clab = 1.5,
    sub = names(w)[k], csub = 2)
```



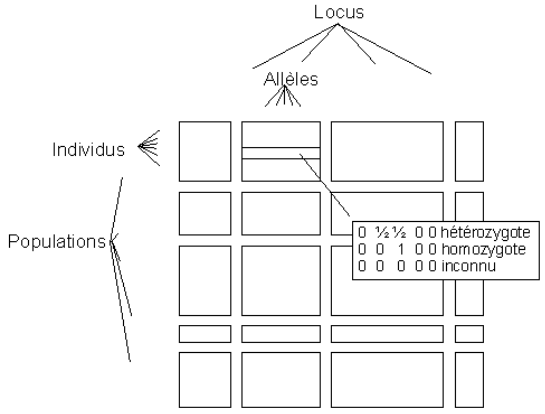
On voit immédiatement que la configuration "mêmes variables et mêmes individus" se présente naturellement. Dans ce champ, les cubes de données jouent un rôle central en exprimant la difficulté immédiate de la multiplicité des points de vue.

Les données de J. Blondel et H. Farré [2] sont un troisième cas d'école, parce qu'elles ont été publiées en toute transparence. 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). Les valeurs sont des effectifs de couples nicheurs pour 100 ha.

A défaut de méthodes appropriées, on peut toujours présenter les données dans une analyse à un tableau. La question qui vient immédiatement est lequel. Il y a, en effet, 6 manières de faire d'un cube un tableau pour se ramener au cas précédent et on a peu de chances de tomber sur la bonne au hasard :



On aborde donc des questions qui ne sont pas simples. La génétique des populations propose même une structure doublement K-tableaux :



l'écologie des communautés a posé des questions expérimentales pertinentes avant d'avoir des solutions techniques adaptées : c'est un mérite qui lui a coûté cher.

2 La classe `ktab`

2.1 Rangement des objets

Qu'il s'agisse de mêmes individus ou de mêmes variables ou des deux, il n'y a qu'une manière de ranger les objets pour des raisins pratiques (figure 1 **A**). Ceci implique quelques précautions pour l'utilisateur. Au plan théorique, on voit donc que trois situations se présentent. K tableaux peuvent avoir les mêmes individus (figure 1 **B**) :

$$(\mathbf{X}_1, \mathbf{Q}_1, \mathbf{D}) \quad (\mathbf{X}_2, \mathbf{Q}_2, \mathbf{D}) \quad \cdots \quad (\mathbf{X}_K, \mathbf{Q}_K, \mathbf{D})$$

ou les mêmes variables (figure 1 **C**) :

$$(\mathbf{X}_1, \mathbf{Q}, \mathbf{D}_1) \quad (\mathbf{X}_2, \mathbf{Q}, \mathbf{D}_2) \quad \cdots \quad (\mathbf{X}_K, \mathbf{Q}, \mathbf{D}_K)$$

ou les mêmes marges :

$$(\mathbf{X}_1, \mathbf{Q}, \mathbf{D}) \quad (\mathbf{X}_2, \mathbf{Q}, \mathbf{D}) \quad \cdots \quad (\mathbf{X}_K, \mathbf{Q}, \mathbf{D})$$

ou être des tables de contingences avec une marge en commun (figure 1 **D**) ce qui pose une question spécifique que nous rencontrerons plus loin. On considère les K-tableaux comme une forme de données unique qui peut avoir des significations diverses : Un objet de la classe `ktab` est une liste de *data frame* ayant en commun

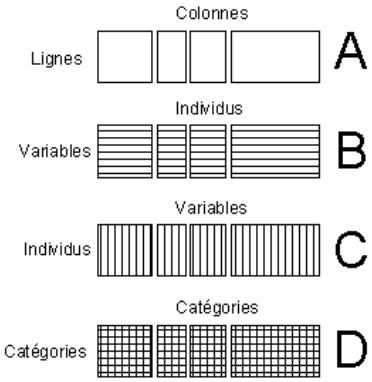
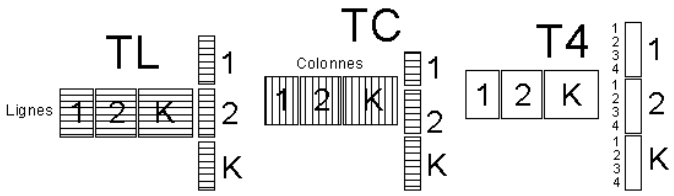


FIG. 1 – K-Tableaux : un stockage commun pour des situations variées.

les `row.names` et les poids. Chaque rectangle représente un *data frame* de la liste. Chacun de ces *data frame* est le tableau d'un schéma de dualité à deux pondération qui s'écrit $(\mathbf{X}_k, \mathbf{Q}_k, \mathbf{D})$ le $\mathbf{D}$ étant commun. Les noms des lignes des $\mathbf{X}_k$ sont les mêmes. Les $\mathbf{Q}_k$sont rangés dans un vecteur commun. Une liste de la classe `ktab` comprend donc :

- `blo` le vecteur du nombre de colonnes des tableaux
- `lw` le vecteur des poids des lignes commun à tous les tableaux
- `cw` le vecteur des poids des colonnes
- `TL` data.frame à deux composantes pour gérer le rangement des paramètres associés aux lignes des tableaux

– `TC` data.frame à deux composantes pour gérer le rangement des paramètres associés aux colonnes des tableaux
– `T4` data.frame à deux composantes pour gérer le rangement des paramètres à 4 composantes associés à un tableau



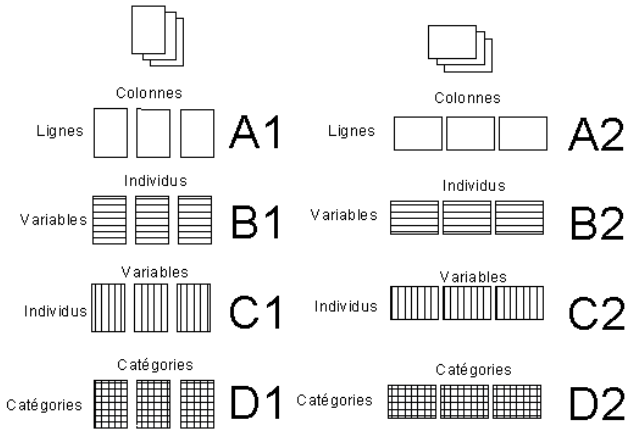
Les composantes `TL`, `TC` et `T4` contiennent des facteurs permettant de manipuler des informations associées aux tableaux par le schéma :

– `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.

2.2 Signification et objectifs

Un objet de la classe `ktab` est donc l'assemblage de K analyses séparées, qu'elles soient appariées par les lignes ou par les colonnes ou par les deux. Ceci signifie que les traitement préalables des tableaux (e.g. centrés par colonnes si les colonnes sont des variables, normalisés par lignes si les lignes sont des variables, doublement centrés si les tableaux sont des tables de contingence, ...) sont faits avant la constitution de la liste.

Ceci est particulièrement important pour les doubles appariements :



Dans ce cas B1 et C2 sont strictement équivalents, de même que C1 et B2, de même que D1 et D2. Ceci est purement formel et ne préjuge en rien de ce qu'on veut faire. On peut introduire la question en disant que globalement :

– un tableau peut être une variable en ce sens que plusieurs colonnes peuvent être nécessaires pour enregistrer l'information d'un type donné ;
– un tableau peut être un individu en ce sens que plusieurs lignes peuvent être nécessaires pour enregistrer l'information d'un type donné ;
– un tableau peut être une structure en ce sens qu'il définit à lui seul une typologie de variables et/ou d'individus.

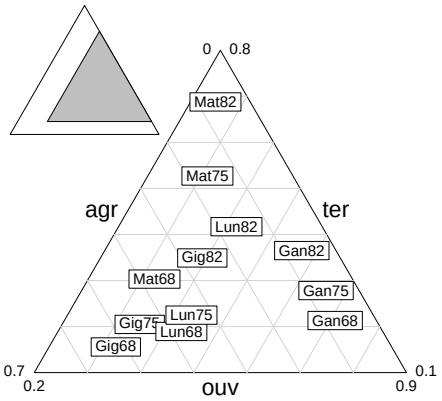
Ainsi une distribution de granulométrie donnée par 20% argile, 70% limon et 10% sable prendra 3 colonnes mais ne contiendra qu'une variable. Au contraire, une mesure de p variables sur n individus appartenant à la même population sera multivariée (chaque colonne est utile indépendamment des autres) mais ne représentera qu'un individu (une population, un groupe, un échantillon).

Au contraire, enfin, un tableau faunistique représente un mode de dispersion des taxons entre les stations, mais rien n'empêche de reproduire la mesure à une autre date et le tableau représente alors une structure. Rien ne se ressemble moins sur le fond que deux K-tableaux de nature différente, alors que formellement ce sont des objets très proches.

On devra donc abandonner définitivement l'idée que la forme des données impose la forme du traitement. Bien au contraire, il est ici possible de faire exactement le contraire de ce dont on a besoin !

La principale difficulté vient d'un argument théorique. Il y a plusieurs méthodes (ACP, AFC, ACM, ...) associées à un seul tableau qui résolvent plusieurs problèmes (inertie, codage, auto-modélisation) à l'aide d'un seul algorithme qui relève d'un seul modèle (*duality diagram*). Cette unicité est perdue dans un couple de tableaux, plus encore avec K tableaux.

La méthode n'existe pas et n'existera jamais. La question centrale est celle 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 [15] pour l'illustration du programme STATIS.

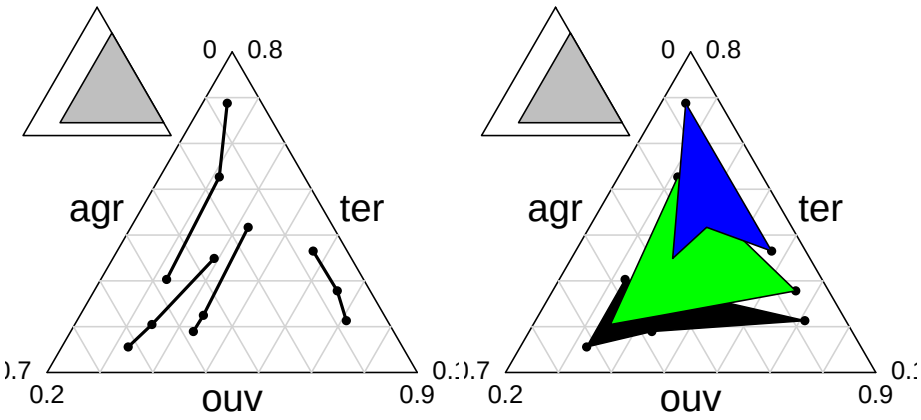


On a pour 4 communes (Gignac, Ganges, Matellas, Lunel) la mesure de 3 variables communes (taux des emplois agricoles, ouvriers et tertiaires) lors de trois recensements (1968, 1975 et 1982).

```
exo1 <- matrix(c(51.88, 32.55, 15.57, 44.94, 34.59, 20.47, 25.95,
    39.15, 34.9, 37.87, 43.19, 18.94, 34.2, 43.32, 22.48, 16.13,
    42.18, 41.69, 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), ncol = 3, byr = T)
exo1 <- as.data.frame(exo1)
names(exo1) <- c("agr", "ouv", "ter")
com <- as.factor(rep(c("Gig", "Lun", "Gan", "Mat"), c(3, 3, 3, 3)))
rec <- as.factor(rep(c("68", "75", "82"), 4))
row.names(exo1) = paste(com, rec, sep = "")
triangle.plot(exo1, lab = row.names(exo1), clab = 1, labeltriangle = T)
```

Or cette figure, interprétée, en donne deux autres.

```
par(mfrow = c(1, 2))
par(mar = rep(0, 4))
w <- triangle.plot(exo1)
l1 <- split(as.data.frame(w), com)
lapply(1:4, function(k) lines(l1[[k]], lwd = 2))
w <- triangle.plot(exo1)
l2 <- split(as.data.frame(w), rec)
lapply(1:3, function(k) polygon(l2[[k]], col = palette(rainbow(3))[k]))
palette("default")
```

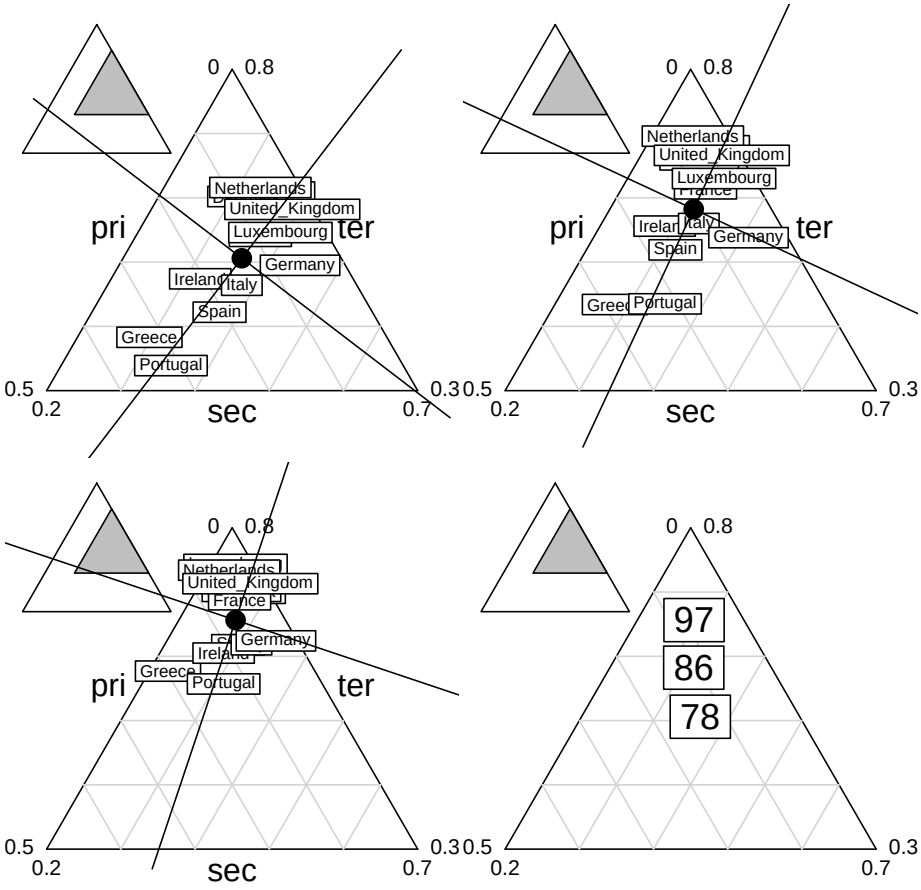


A gauche, on a dessiné les évolutions par commune, ce qui conduit à une question sur *la typologie des évolutions*. A droite on a dessiné les typologies de communes par dates, ce qui conduit à une question sur *l'évolution des typologies*. Ici les données sont en dimension deux. On voit tout du même endroit. Quand il y a de nombreuses variables, les deux représentations associées aux deux objectifs peuvent être totalement différentes. 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)
moy <- t(data.frame(lapply(euro123[1:3], function(x) apply(x, 2,
    mean))))
moy
```

```
           pri       sec       ter
in78 0.1340000 0.3600000 0.5060000
in86 0.1046667 0.3130000 0.5823333
in97 0.0672500 0.2764167 0.6563333
```

```
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 puisque 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.

Ces questions ont été posées à propos de l'évolution du tapis végétal [19]. De la morphométrie à l'écosystémie, la biologie évolutive est un domaine privilégié de réflexion sur la variabilité.

2.3 Comment faire des `ktab` ?

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

1. `ktab.list.dudi` : avec une liste de schémas
2. `ktab.list.dudi` : avec une liste de data.frame

3. `ktab.data.frame` : avec un data.frame
4. `ktab.within` : à partir d'une analyse intra-classe

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

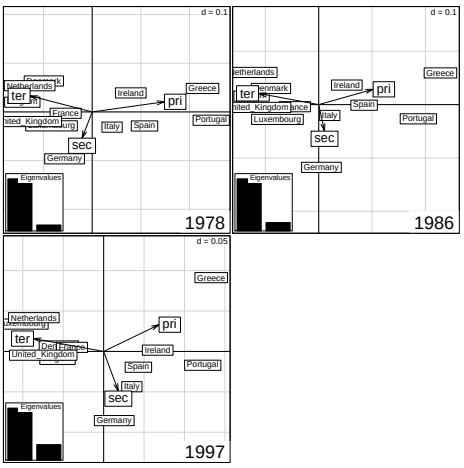
```
data(euro123)
pca1 <- dudi.pca(euro123$in78, scale = F, scann = F)
pca2 <- dudi.pca(euro123$in86, scale = F, scann = F)
pca3 <- dudi.pca(euro123$in97, scale = F, scann = F)
kta1 <- ktab.list.dudi(list(pca1, pca2, pca3), tabnames = c("1978",
    "1986", "1997"))
kta1
```

Identifier les éléments. 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 manière 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)
T78 <- data.frame(scalewt(euro123$in78, scale = F))
T86 <- data.frame(scalewt(euro123$in86, scale = F))
T97 <- data.frame(scalewt(euro123$in97, scale = F))
kta2 <- ktab.list.df(list(T78, T86, T97), tabnames = c("1978", "1986",
    "1997"))
kta2
```

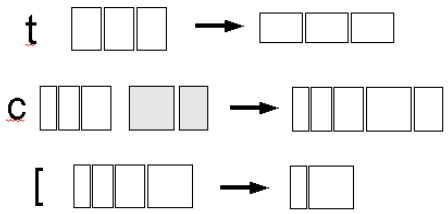
Observer qu'on obtient, avec les valeurs par défaut, une pondération unitaire des lignes, ce qui n'est pas le cas du précédent. Cette option est surtout faite pour associer des tableaux dont les lignes sont des variables (data.frame transposés). La troisième est la plus simple : elle part d'un tableau et d'un vecteur de nombre de colonnes par blocs :

```
w <- cbind.data.frame(euro123$in78, euro123$in86, euro123$in97)
w <- scalewt(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. Les opérations possibles sur les K-tableaux sont schématisées par :



3 Les variables floues

Un archétype des tableaux ayant valeur expérimentale de variables est celui des variables floues. Quand on veut noter le lien entre une espèce et l'altitude, on peut dire 670 m pour signifier que l'espèce se trouve en moyenne à 670 m. On peut dire 1|3|2|0|0 sur le code [0,400], [400,800], [800,1200], [1200,1600] et [1600,2000] pour dire qu'une fois sur 6 on trouve l'espèce dans la première classe, 1 fois sur 2 dans la seconde et 1 fois sur 3 dans la troisième. Le profil écologique est une variable floue.

La littérature naturaliste du siècle passé est devenu un nouvel espace experimental en permettant l'extraction (*data mining*) de l'information sur des grands ensembles de taxa.

Sur plusieurs variables et plusieurs espèces, on obtient des tableaux de variables floues :

	1	2	3
1	1	3	0
2	1	3	0
3	3	0	0
4	1	3	0
5	3	0	0
6	3	0	0
7	0	1	3
8	2	2	0
9	0	1	3
10	1	3	0
11	0	0	0
12	1	3	0
13	1	3	0
14	1	3	0
15	3	0	0
16	1	3	0
17	1	3	0

	1	2	3
1	0.250	0.750	0.000
2	0.250	0.750	0.000
3	1.000	0.000	0.000
4	0.250	0.750	0.000
5	1.000	0.000	0.000
6	1.000	0.000	0.000
7	0.000	0.250	0.750
8	0.500	0.500	0.000
9	0.000	0.250	0.750
10	0.250	0.750	0.000
11	0.000	0.000	0.000
12	0.250	0.750	0.000
13	0.250	0.750	0.000
14	0.250	0.750	0.000
15	1.000	0.000	0.000
16	0.250	0.750	0.000
17	0.250	0.750	0.000

	1	2	3
1	0.250	0.750	0.000
2	0.250	0.750	0.000
3	1.000	0.000	0.000
4	0.250	0.750	0.000
5	1.000	0.000	0.000
6	1.000	0.000	0.000
7	0.000	0.250	0.750
8	0.500	0.500	0.000
9	0.000	0.250	0.750
10	0.250	0.750	0.000
11	0.339	0.407	0.254
12	0.250	0.750	0.000
13	0.250	0.750	0.000
14	0.250	0.750	0.000
15	1.000	0.000	0.000
16	0.250	0.750	0.000
17	0.250	0.750	0.000

L'exemple de B. Statzner et ses collègues [18] est documenté dans :

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

```
data(bsetal97)
w = bsetal97$biol.blo[1:3]
biol.fuzzy = prep.fuzzy.var(bsetal97$biol, bsetal97$biol.blo)
```

```
17 missing data found in block 1
14 missing data found in block 2
28 missing data found in block 3
8 missing data found in block 4
5 missing data found in block 5
19 missing data found in block 6
10 missing data found in block 7
5 missing data found in block 8
2 missing data found in block 9
12 missing data found in block 10
```

```
w1 = biol.fuzzy[, 1:19]
ww1 = 1:sum(w)
ww0 = seq(from = 0, by = 4, len = length(w))
ww0 = rep(ww0, w) + ww1
phy = taxo2phylog(as.taxo(bsetal97$taxo))
row.names(w1) = names(phy$leaves)
table.phylog(w1, phy, x = ww0, clabel.r = 0.5, clabel.col = 0.75,
    csi = 0.5, clabel.n = 0)
```

Il y a identité formelle entre les trois classes de données :

- Traits biologiques en variables floues : espèces en lignes, modalités en colonnes, chaque paquet de modalités définit un trait biologique ou écologique (figure 2).
- Fréquences alléliques : populations en lignes, allèles en colonnes, chaque paquet d'allèles est attaché à un locus donné.
- Usage du code génétique : séquences en lignes, codons en colonnes, chaque paquet de codons regroupe les synonymes codant pour un même acide aminé.

Ces quelques exemples invitent à se poser des questions : peut-on comparer des typologies ? Existe-t'il une moyenne de typologies ? Mesure-t'on la variabilité entre typologies ? Les traits biologiques sont-il liés entre eux ? Qu'est ce que la corrélation entre tableaux ? Peut-on repérer des tableaux qui ne servent à rien, des locus non structurant, des groupes faunistiques non indicateurs ? L'accès le plus facile se fait par les analyses inter et intra-classes.

4 Les analyses inter et intra-classes

4.1 Tableaux de modèles et d'erreurs

Dès qu'on introduit un facteur en face d'une variable on décompose la variance de cette variable en deux composantes inter (la variabilité des moyennes par classes) et intra (la variabilité moyenne dans une classe). Tout ceci s'étend à une matrice de variances covariances sans difficulté. $\mathbf{X}$ est un tableau, $\mathbf{X}$ est le tableau centré associé, $\mathbf{C} = w\mathbf{Y}^T\mathbf{Y}$ sa matrice de covariances (w est $1/n$ ou $1/(n-1)$ suivant les sensibilités et cela n'a aucune importance). La régression sur le facteur est une projection qui remplace les données par les moyennes par classe. Les données se décompose en un modèle et une erreur

$$\mathbf{Y} = \widehat{\mathbf{Y}} + (\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{M} + \mathbf{E}$$

équation qui ne devrait pas poser de problème mais n'est pas complètement stupide si on rajoute que :

$$\mathbf{C} = w(\mathbf{M} + \mathbf{E})^T(\mathbf{M} + \mathbf{E}) = w\mathbf{M}^T\mathbf{M} + w\mathbf{E}^T\mathbf{E}$$

La variance (la matrice de covariance) des données égale velle des modèles plus celle des erreurs parce que la covariance (la matrice de covariance) entre les données et les erreurs est nulle. C"est la chose essentielle à savoir !

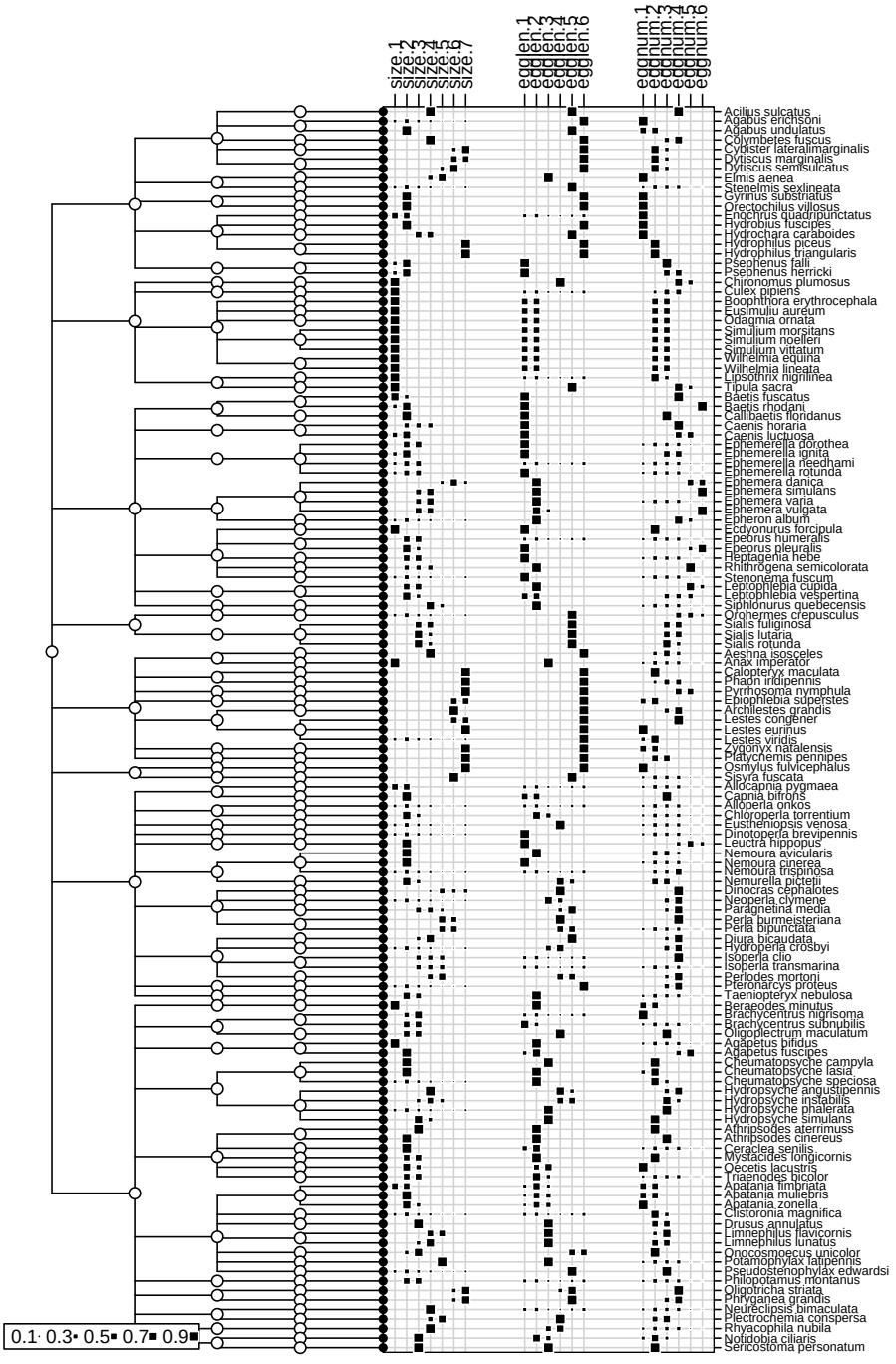


FIG. 2 – Traits biologiques, taxonomie et phylogénie en mélange : une grande classe de données qui relèvent de la statistique K-tableaux.

```
data(jv73)
Y = as.data.frame(scale(jv73$morpho))
M <- as.data.frame(lapply(Y, function(x) predict(lm(x ~ jv73$fac.riv))))
E <- as.data.frame(lapply(Y, function(x) residuals(lm(x ~ jv73$fac.riv))))
cbind(round(cov(Y, Y), 2), round(cov(E, M), 2))
```

```
      Alt   Das   Pen   Smm   Qmm   Vme Alt Das Pen Smm Qmm Vme
Alt  1.00 -0.14  0.21 -0.14 -0.09  0.03   0   0   0   0   0   0
Das -0.14  1.00 -0.79  0.66  0.71  0.45   0   0   0   0   0   0
Pen  0.21 -0.79  1.00 -0.59 -0.58 -0.23   0   0   0   0   0   0
Smm -0.14  0.66 -0.59  1.00  0.86  0.27   0   0   0   0   0   0
Qmm -0.09  0.71 -0.58  0.86  1.00  0.63   0   0   0   0   0   0
Vme  0.03  0.45 -0.23  0.27  0.63  1.00   0   0   0   0   0   0
```

```
cbind(round(cov(M, M), 2), round(cov(E, E), 2))
```

```
      Alt   Das   Pen   Smm   Qmm   Vme   Alt   Das   Pen   Smm   Qmm   Vme
Alt  0.69  0.19 -0.06  0.03  0.11  0.17  0.31 -0.34  0.27 -0.17 -0.20 -0.14
Das  0.19  0.33 -0.29  0.32  0.40  0.29 -0.34  0.67 -0.50  0.34  0.31  0.17
Pen -0.06 -0.29  0.34 -0.30 -0.34 -0.19  0.27 -0.50  0.66 -0.29 -0.24 -0.04
Smm  0.03  0.32 -0.30  0.65  0.61  0.20 -0.17  0.34 -0.29  0.35  0.24  0.07
Qmm  0.11  0.40 -0.34  0.61  0.71  0.41 -0.20  0.31 -0.24  0.24  0.29  0.22
Vme  0.17  0.29 -0.19  0.20  0.41  0.50 -0.14  0.17 -0.04  0.07  0.22  0.50
```

Nous avons ici une illustration de la linéarité très forte. Le tableau des données **Y** est la somme du modèle et de l'erreur. Dans **M** chaque colonne est le résultat d'une application linéaire (projection) portant sur la colonne correspondante de **Y**. Mais chaque ligne de **M** est déduit de la ligne correspondante de **Y** (déplacement vers le centre de gravité de la classe) qui ne l'est pas. Faire l'analyse de **Y** c'est faire l'analyse globale. Faire celle de **M** c'est faire l'analyse inter-classe, faire celle de **E** c'est faire l'analyse intra-classe.

```
glob1 <- dudi.pca(jv73$morpho, scale = T, scan = F)
bet1 <- between(glob1, jv73$fac.riv, scan = F)
wit1 <- within(glob1, jv73$fac.riv, scan = F)
```

Observer que l'inertie de la première est de 5 (le nombre de variable) dont 54% est de nature inter-classe et 46% de nature intra-classe. En gros la moitié de la variabilité totale est prédictible par le nom de la rivière. Mais observer que la réduction de dimension dans les trois cas est quasiment la même :

```
round(glob1$eig/sum(glob1$eig), 2)
```

```
[1] 0.57 0.18 0.13 0.09 0.03 0.01
```

```
round(bet1$eig/sum(bet1$eig), 2)
```

```
[1] 0.61 0.22 0.10 0.06 0.00 0.00
```

```
round(wit1$eig/sum(wit1$eig), 2)
```

```
[1] 0.64 0.19 0.07 0.05 0.04 0.01
```

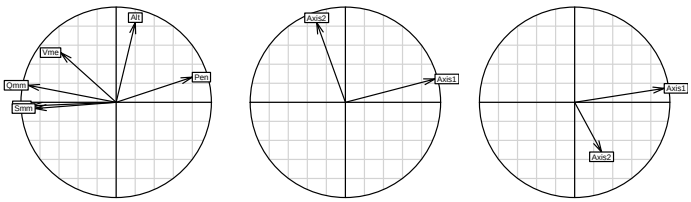
Observer que les coordonnées sont très voisines :

```
a = glob1$li[, 1]
b = bet1$ls[, 1]
c = wit1$ls[, 1]
round(cor(data.frame(a, b, c)), 2)
```

```
     a    b    c
a 1.00 0.98 0.98
b 0.98 1.00 0.93
c 0.98 0.93 1.00
```

Le lien entre les variables est de même nature entre les stations d'une rivière et entre les moyennes par rivières. Il s'agit d'un élément mécanique : une rivière d'altitude a un débit plus faible et une pente plus forte qu'une rivière moins élevée. Retrouver que l'altitude varie beaucoup entre rivières. Mais l'amont d'une rivière a un débit plus faible et une pente plus forte que l'aval, quelle que soit l'altitude. On retrouve ça sur la figure plus technique qui suit :

```
par(mfrow = c(1, 3))
s.corcircle(glob1$co)
s.corcircle(bet1$as)
s.corcircle(wit1$as)
```



Dans $\mathbb{R}^{92}$ les variables (normalisée, de longueur 1) sont projetées sur le plan des composantes. Dans $\mathbb{R}^5$ les axes d'inertie (normés) de l'analyse inter (les axes d'inertie du nuage des centrs de gravité) sont projetés sur les axes d'inertie globaux. Enfin, les axes d'inertie de l'analyse intra (celle du nuage formé des sous-nuages recentrés) sont projetés sur les axes d'inertie globaux.

Cette introduction n'a qu'un seul objectif : souligner qu'un paquet de tableaux superposés est une structure de données qui conduit à l'emploi de méthodes profondément différentes dans leurs objectifs. Si les colonnes communes sont des variables, on peut séparer les populations par leur moyenne (inter-classes, discriminante).

On peut rechercher au contraire un compromis entre leur structure interne (intra-classes, STATIS, *cf. infra*). Inter-classes et intra-classes sont des jumelles au plan théorique (deux ACPVI projetant soit sur un espace soit sur son orthogonal). Elles sont opposées au plan expérimental (ce qui sépare contre ce qui rapproche).

Intra-classes et STATIS sont fort éloignées au plan théorique, elle sont voisines au plan expérimental. Plus qu'ailleurs il est indispensable de savoir ce que l'on veut.

Ici, c'est l'analyse intra-classe qui nous intéresse, pare qu'elle aborde chaque rivière comme un système. Si on ne s'intéresse qu'à cet aspect, sans vouloir le coordonner à une analyse globale ou inter-associée, deux fonctions sont à connaître.

4.2 Transformation de Bouroche

La première concerne les variables numériques. Ce qui précède utilise une normalisation des données (élimination des questions d'unité en plaçant les moyennes à 0 et les variances à 1) puis un recentrage par groupe (élimination de l'effet moyen). Il s'en suit que la variance totale d'une variable est sa variance inter-classe. On peut amplifier la correction soit en replaçant la variance à 1 dans chaque groupe (chaque variable joue dans chaque groupe un rôle équivalent) soit en reportant la variance totale après recentrage à [1] (chaque variable joue globalement le même rôle en pouvant avoir des variances par groupes très diverse). On utilise dans ce but la fonction `within.pca` avec les option `scaling = "partial"` ou `scaling = "global"`. Dans le premier cas, on fait des ACP normée par blocs, dans le second on fait des ACP centrées par blocs après une

normalisation sur variables recentrées (transformation de Bouroche [3]). En utilisant simplement `dudi.pca` et `within` on fait des ACP centrées par blocs avec ou sans normalisation préalable. `within` s'applique à tout type de variables.

Cette discussion est un peu pénible mais indique une difficulté. C'est encore une question d'intention. Par exemple, la variabilité entre les stations d'une rivière n'est pas constante dans le temps. Elle s'exprime plus ou moins en fonctions des événements hydrologiques. Si on pense que c'est un fait parasite, on remet les variances à l'unité. Si on pense que c'est un fait important, on le laisse s'exprimer dans des variances inégales. J. Verneaux, dans un recodage intégral des données en classes (voir la fiche citée) veut maîtriser la signification des valeurs.

On le comprend. Toutes les méthodes sont très sensibles aux transformations préalables. Il faut en prendre le plus grand soin. Les critères expérimentaux doivent ici prendre le pas sur les contraintes techniques.

4.3 Les AFC intra-classes

La seconde concerne la notion d'intra-classe en AFC [1]. C'est une question assez difficile sur le fond pour une raison très simple. Dans un tableau à n lignes et p colonnes décomposé en bloc de n_1, n_2, ..., n_K lignes, l'ACP du total donne un poids unité à chaque variable, les ACP de chacun des tableaux font de même et la coordination des différentes analyses se fait simplement. Dans la même situation, en AFC, il en va tout autrement. L'AFC du tout utilise la pondération marginale du tableau complet, chaque AFC de chaque morceau utilise sa propre pondération marginale et devient difficilement comparable avec les autres.

On doit imposer une marge commune et recentrer les blocs en conséquence. Les différents morceaux sont alors des AFC sur modèles. Ceci concerne les tableaux faunistiques avec les les mêmes espèces ou les mêmes sites ou (encore pire) les deux. Ceci concerne aussi les tableaux populations-allèles par locus et de manière générale les tables de contingence à marges communes. La fonction utile est `within.coa`, le descriptif du calcul est p. 27-29 dans :

`http://pbil.univ-lyon1.fr/R/fichestd/tdr62.pdf`

Noter encore que le tableau d'AFC peut être doublement partitionné. On trouve cette situation avec des espèces par groupes et des relevés par date [5] ou avec des séquences par espèces et des codons par acides aminés [17]. On trouvera la fonction associée (`witwit`) une référence utile avec [4], un exemple dans :

`http://pbil.univ-lyon1.fr/R/querep/qrd.pdf`

et un descriptif à :

`http://pbil.univ-lyon1.fr/R/thematique/them6.pdf`

5 Approches élémentaires des cubes de données

5.1 Analyses triadiques partielles

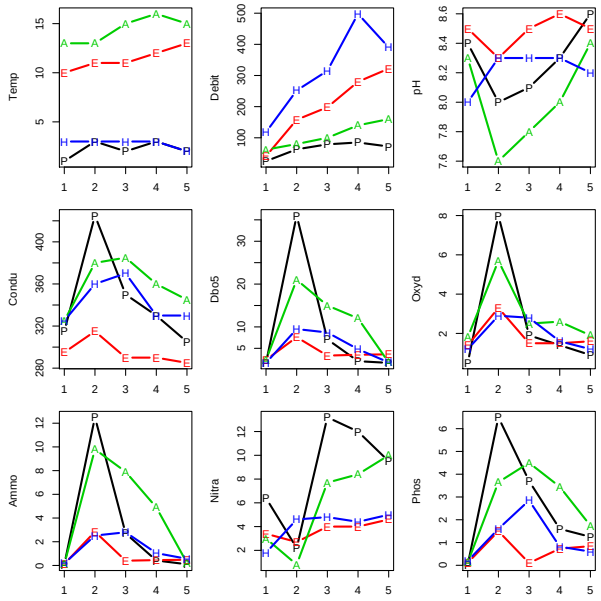
La plus simple des méthodes multi-tableaux pour les cubes est appelée à tort analyse triadique dans [20] et à raison analyse triadique partielle dans [14] ou

Pré-STATIS (ou STATIS sur les X) dans [16] ou PCA-SUP (PCA of a derived two-way supermatrix) dans [12] et envisagée à l'origine dans [21]. 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
```

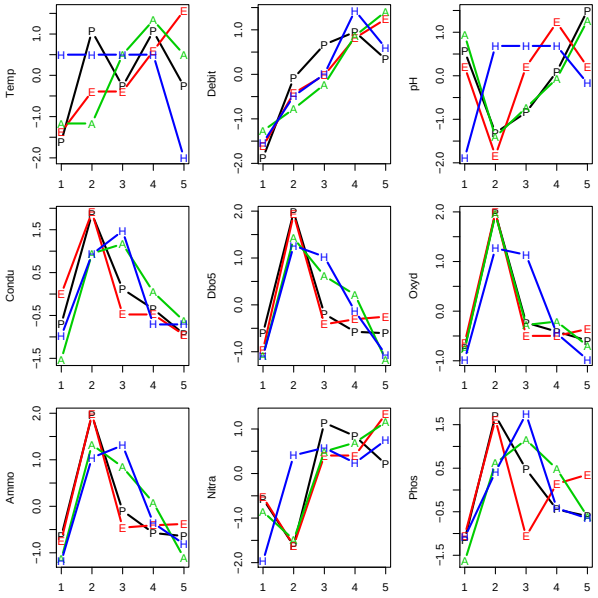
On a mesuré 9 variables, respectivement 1- Température (°C), 2- Débit (l/s), 3- pH, 4- Conductivité ($\mu\Omega$/cm), 5- Oxygène (% de saturation), 6- DBO5 (mg/l oxygène), 7- Ammoniaque (mg/l), 8- Nitrates (mg/l) et 9- Orthophosphates (mg/l) dans 5 stations d'un réseau hydrographique à 4 saisons (printemps, été, automne, hiver). Les 5 stations sont placées d'amont en aval.




```
par(mfrow = c(3, 3))
par(mar = c(2.1, 4.1, 1.1, 1.1))
site <- as.numeric(meaudret$plan$sta)
date <- meaudret$plan$dat
levels(date) <- c("P", "E", "A", "H")
for (var in 1:9) {
    z <- meaudret$mil[, var]
    plot(site, z, type = "n", ylab = names(meaudret$mil)[var])
    for (sai in 1:4) {
        w <- levels(date)[sai]
        points(meaudret$mil[date == w, var], pch = w, type = "b",
            col = palette()[sai], lwd = 2)
    }
}
```

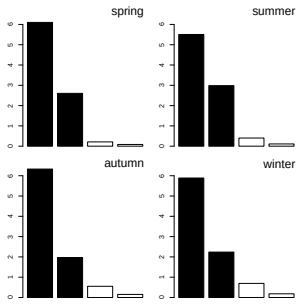
Les données contiennent une structure temporelle et une structure spatiale. En normalisant les données par date, on impose un point de vue.

```
pca0 <- withinpca(meaudret$mil, meaudret$plan$dat, scal = "partial",
    scann = F)
par(mfrow = c(3, 3))
par(mar = c(2.1, 4.1, 1.1, 1.1))
for (var in 1:9) {
    z <- pca0$tab[, var]
    plot(site, z, type = "n", ylab = names(pca0$tab)[var])
    for (sai in 1:4) {
        w <- levels(date)[sai]
        points(pca0$tab[date == w, var], pch = w, type = "b", col = palette()[sai],
            lwd = 2)
    }
}
```



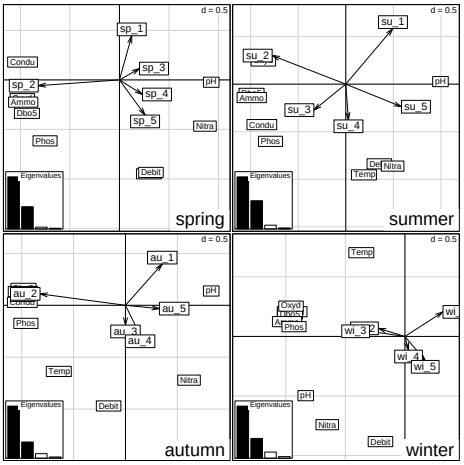
Chaque date donne un tableau d'ACP normée. Nous avons donc 4 ACP à interpréter. La fonction **sepan** est la plus simple et l'une des plus utile qui s'applique à un K-tableau.

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



On reconnaît les 4 graphes de valeurs propres. Les K-tableaux ont des méthodes qui comportent des fonctions kplot :

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



Le seul intérêt est de souligner l'absence de cohésion. On sait qu'un axe est lui-même ou son opposé avec exactement la même signification, comme à pile ou face.

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). Mais attention :

```
pta0 <- pta(kta0)
```

Observer le message d'erreur `Error in pta(kta0) : non equal col.names among array`. En effet :

```
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" "au_1"
[12] "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, scan = F)
```

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$RV
```

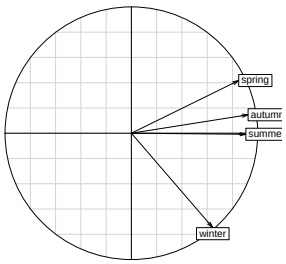
```
          spring    summer    autumn    winter
spring 1.0000000 0.6934558 0.7886185 0.2834592
summer 0.6934558 1.0000000 0.7671756 0.5340456
autumn 0.7886185 0.7671756 1.0000000 0.4794976
winter 0.2834592 0.5340456 0.4794976 1.0000000
```

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] 2.8121449 0.7541476 0.2536781 0.1800294
```

```
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] 0.5066788 0.5403976 0.5509640 0.3842991
```

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

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

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`. Vérifier l'existence de ces objets dans `pta0`.

Ce tableau compromis peut faire l'objet de toutes les manipulations possibles sur une analyse ordinaire, comme une analyse d'inertie :

```
inertia.dudi(pta0, row = T)$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 est :

```
pta0$cos2
```

```
        1         2         3         4
0.8496725 0.9062170 0.9239364 0.6444484
```

On obtient des coordonnées pour les lignes et les colonnes et les scores normés associés dans les composantes `li` et `co`, `l1` et `c1`. Outre le calcul d'un tableau combinant tous les tableaux de départ et son analyse, on a la configuration très particulière (figure 3).

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

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

On observera 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 et on préfèrera la figure plus lisible donnée par :

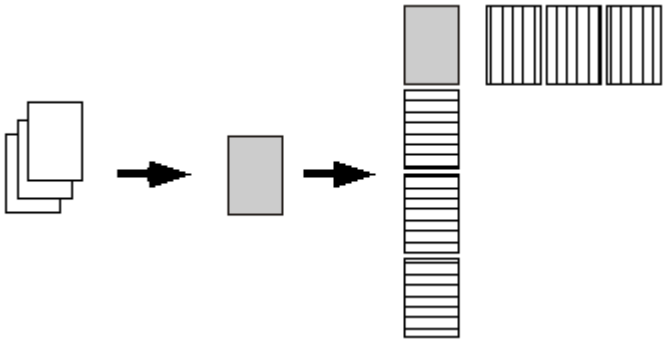
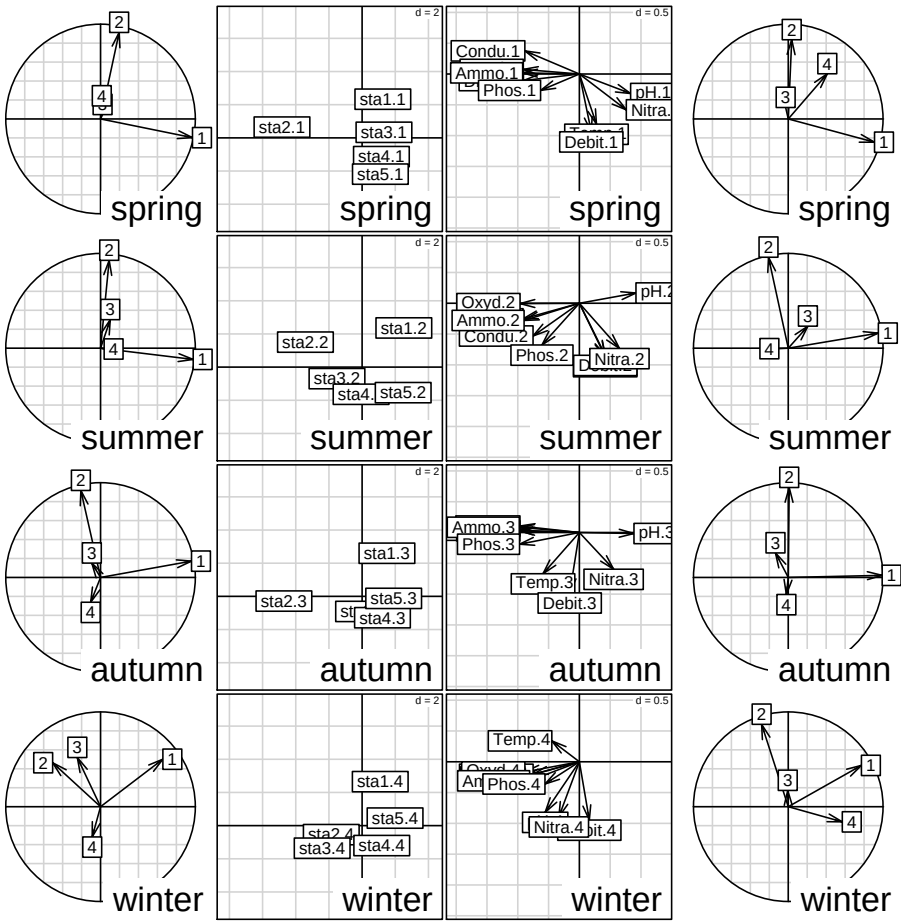


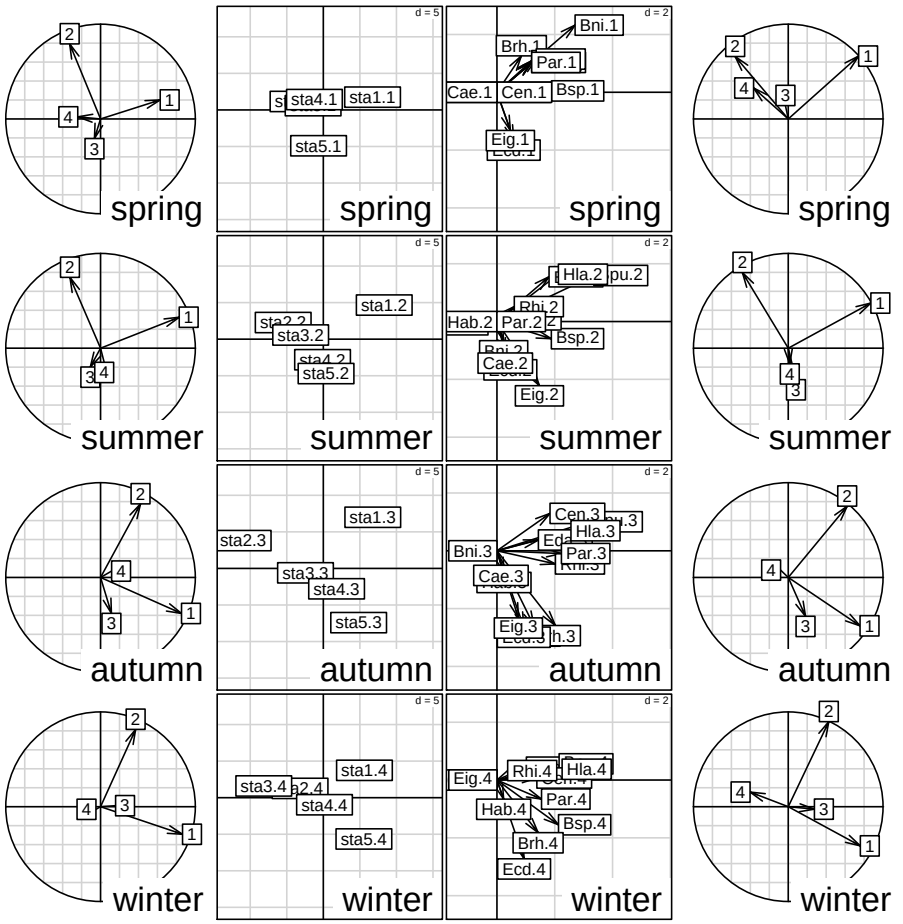
FIG. 3 – En gris le tableau compromis dont l'analyse permet le projection en individus supplémentaires de toutes les lignes et la projection en variables supplémentaires de toutes les colonnes. On peut même projeter sur un plan des axes principaux du compromis les axes principaux de chaque tableau et sur le plan des composantes principales du compromis les composantes principales de chacun des tableaux.

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



On voit ainsi K analyses simultanées exprimées dans le même cadre géométrique. 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)
```



L'ACP 3-modes fait partie des méthodes qui étendent les propriétés de reconstitution de données. On sait approximerestimer un tableau par :

$$x_{ij} = a_i b_j + r_{ij}$$

On peut donc chercher des modèles des cubes de données dont le plus simple est :

$$x_{ijk} = a_i b_j c_k + r_{ijk}$$

Une bibliographie considérable est associée à cette problématique [13] [6] [10]. Le site de référence de la *Three-Mode Company* est consacré à ce domaine :

`http://three-mode.leidenuniv.nl/index.html`

5.2 L'AFC de Foucart

La version analyse des correspondances dite "AFC de Foucart" [7] [8] [9] 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 [2]. 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"
```

Elles forment une liste de 6 data.frame de dimensions 79×4. 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 citées au paragraphe précédent.

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. On trouve des éléments pour cette stratégie dans la fiche sur l'ordination indirecte.

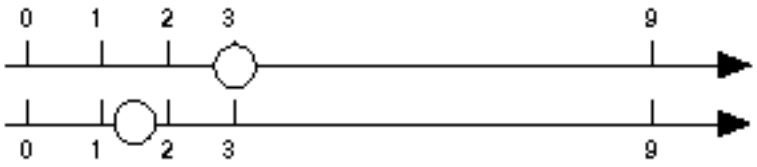
T. Foucart part de la constatation qu'on peut aussi bien concevoir une table de contingence comme un tableau d'ACP particulière que comme une matrice de covariance particulière. Mais dans un cas comme dans l'autre, la question des pondérations interdit de généraliser STATIS (aussi bien sur les tableaux 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 tableau $\mathbf{X}_k$ 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

$$\mathbf{P}_k = \left[\frac{x_{ij}^k}{x_{..}^k} \right]$$

La moyenne est :

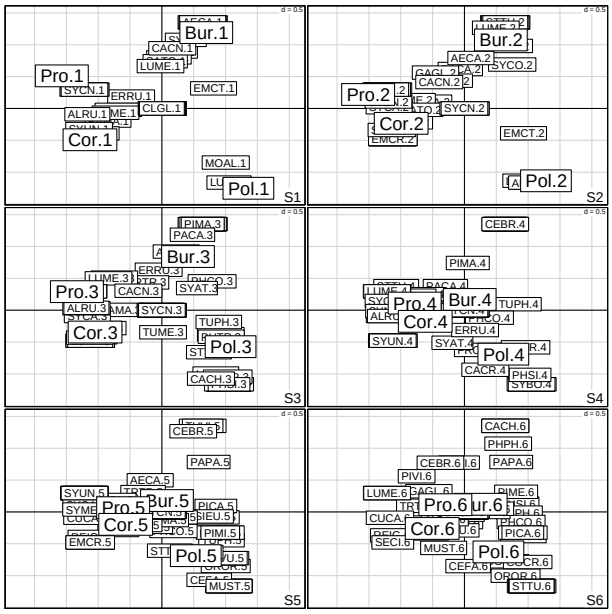
$$\mathbf{P} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{P}_k$$

On fait l'analyse des correspondances de $mathbf{P}$, 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)
```

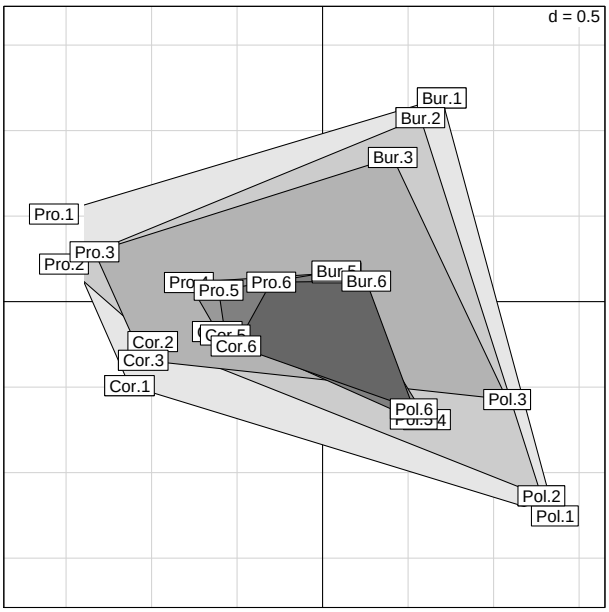
Les sorties de cette fonctions sont de même logique que celles de `pta`.

```
kplot(fou1, clab.r = 1.5, clab.c = 2.5)
```



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.

Références

[1] J.P. Benzécri. Analyse de l'inertie intra-classe par l'analyse d'un tableau de correspondances. *Les Cahiers de l'Analyse des données*, 8 :351–358, 1983.

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

[3] J.M. Bouroche. *Analyse des données ternaires : la double analyse en composantes principales*. PhD thesis, 1975.

[4] M. Bécue, J. Pagès, and C.-E. Pardo. Contingency table with a double partition on rows and colums. visualization and comparison of the partial and global structures. In J. Janssen and P. Lenca, editors, *Applied stochastic models and data analysis*, pages 355–364. ENST Bretagne, Brest, 2005.

[5] P. Cazes, D. Chessel, and S. Dolédec. L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. *Revue de Statistique Appliquée*, 36 :39–54, 1988.

[6] R. Coppi and S. Eds. Bolasco. *Multiway Data Analysis*. Elsevier Science Publishers B.V., North-Holland, 1989.

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

[8] T. Foucart. Une nouvelle approche de la méthode statis. *Revue de Statistique Appliquée*, 31 :61–75, 1983.

[9] T. Foucart. *Analyse factorielle de tableaux multiples*. Masson, Paris, 1984.

[10] A. Franc. *Etude algébrique des multitableaux : apports de l'algèbre tensorielle*. PhD thesis, 1992.

[11] L.E. Friday. The diversity of macro invertebrate and macrophyte communities in ponds. *Freshwater Biology*, 18 :87–104, 1987.

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

[13] P.M. Kroonenberg. *Three-mode principal component analysis*. DSWO Press, Leiden, 1983.

[14] P.M. Kroonenberg. The analysis of multiple tables in factorial ecology. iii three-mode principal component analysis :"analyse triadique complète". *Acta Œcologica, Œcologia Generalis*, 10 :245–256, 1989.

[15] Ch. Lavit. *Analyse conjointe de tableaux quantitatifs*. Masson, Paris, 1988.

[16] B. Leibovici. *Facteurs à mesures répétées et analyses factorielles : application à un suivi épidémiologique*. PhD thesis, 1993.

[17] J.R. Lobry and D. Chessel. Internal correpondence analysis of codon and amino acid usage in thermophilic bacteria. *Journal of Applied Genetics*, 44 :235–261, 2003.

[18] B. Statzner, K. Hoppenhaus, M.-F. Arens, and Ph. Richoux. Reproductive traits, habitat use and templet theory : a synthesis of world-wide data on aquatic insects. *Freshwater Biology*, 38 :109–135, 1997.

[19] M.D. Swaine and P. Greig-Smith. An application of principal components analysis to vegetation changes in permanent plot. *Journal of Ecology*, 68 :33–41, 1980.

[20] J. Thioulouse and D. Chessel. 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, 1987.

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

[22] J. Verneaux. Cours d'eau de franche-comté (massif du jura). recherches écologiques sur le réseau hydrographique du doubs. essai de biotypologie. Thèse d'état, besançon, 1973.