Fiches de stage avec le logiciel

Invitation à utiliser la librairie ade4

D. Chessel, A.-B. Dufour & J. Lobry

24 juin 2006

La fiche prépare une visite guidée de la librairie ade4 sous forme de quelques repères théoriques centrés sur des représentations graphiques. Elle a été écrite pour une session d'un atelier de formation statistique organisé au CEREGE (Aix-enProvence) par J. Guiot (Lundi 28 février 2005).

Table des matières

1 La composante graphique **2**
1.1 La représentation triangulaire 2
1.2 Espace concret et cartes 2
1.3 Espace abstrait et cartes factorielles 2

2 Qu'est-ce que l'analyse en composantes principales ? **19**
2.1 Estimer les paramètres d'une loi normale multivariée 19
2.2 La variable cachée des psychométriciens [11] 20
2.3 Faire une typologie de variable 20
2.4 Les valeurs propres : un élément fondamental 20
2.5 Faut-il centrer, normaliser, transformer ? 20
2.6 ACM : l'équivalent pour des variables qualitatives 20

3 L'Analyse des correspondances **20**
3.1 Les individus sont dans les cases du tableau 25
3.2 Les occurrences sont des individus statistiques 25
3.3 l'AFC est une double analyse discriminante 25

4 Les stratégies de couplage de tableaux **25**
4.1 Assemblages de tableaux 25
4.2 Analyses canoniques 25
4.3 Variables instrumentales 25
4.4 La co-inertie 32

5 Utiliser des matrices de distance **32**
5.1 De distances à dendrogrammes : les CAH 32
5.2 De distances à tableaux : les coordonnées principales 32
5.3 ktab et kdist : vers de nouvelles classe d'objets 32

Références **36**

1 La composante graphique

La partie graphique est vue comme langage d'interface entre le monde expérimental et l'espace théorique. Utiliser les données `euro123` et la documentation de `triangle.plot`.

1.1 La représentation triangulaire

La représentation triangulaire (figure 1) est une carte factorielle sans erreur et sera étendue en dimension quelconque par l'ACP [7]. Trajectoires des pays, déplacement du bloc, conservation des positions relatives : toutes les questions de la statistique multi-tableaux sont posées. Analyser c'est d'abord prendre en compte un objectif et résoudre des problèmes de contraintes. Nuage de points, centre de gravité, couplage de nuages de points, définition de l'axe principal sont sur la figure 2. Il est rare de voir les axes représentés dans le nuage et non pas le nuage positionné par les axes ! La figure 3 permet d'apprécier le déplacement d'un centre de gravité et la modification de la dimension du nuage : elle invite à passer d'un discours sur la valeur à un discours sur la typologie. Ceci implique d'utiliser des outils spécifiques.

1.2 Espace concret et cartes

Coordonnées, unités surfaciques, graphes de voisinages sont des éléments qui introduisent l'espace en analyse des données. Voir les données dans l'espace est essentiel. Utiliser la documentation de la fonction `area.plot` (figure 4). Quand c'est possible, voir les données est préalable à toute analyse (figure 5). Ce peut être même préférable à toute analyse (figure 6) ! Pour obtenir la représentation de la migration des Sarcelles d'hiver [10] [14], installer si possible la librairie `pixmap` et utiliser :

```
if (require(pixmap, quietly = TRUE)) {
    bkgnd.pnm <- read.pnm(system.file("pictures/sarcelles.pnm",
        package = "ade4"))
    data(sarcelles)
    par(mfrow = c(4, 3))
    for (i in 1:12) {
        s.distri(sarcelles$xy, sarcelles$tab[, i], pixmap = bkgnd.pnm,
            sub = sarcelles$col.names[i], clab = 0, csub = 2)
        s.value(sarcelles$xy, sarcelles$tab[, i], add.plot = TRUE,
            cleg = 0)
    }
}
```

L'analyse intégrale est formée d'une figure. On trouvera un exemple équivalent dans la documentation de `s.image`.

1.3 Espace abstrait et cartes factorielles

La capacité de lire une structure se perd totalement au delà de la dimension 3. C'est pourquoi on utilise l'analyse des données. Pour voir ce qui est de dimension multiple, on utilise la projection euclidienne sur des espaces de dimension réduite. Il s'agit malheureusement de mathématiques.

```
library(ade4)
data(euro123)
tot <- rbind.data.frame(euro123$in78, euro123$in86, euro123$in97)
row.names(tot) <- paste(row.names(euro123$in78), rep(c(1,
    2, 3), rep(12, 3)), sep = "")
triangle.plot(tot, label = row.names(tot), clab = 0.75)
```

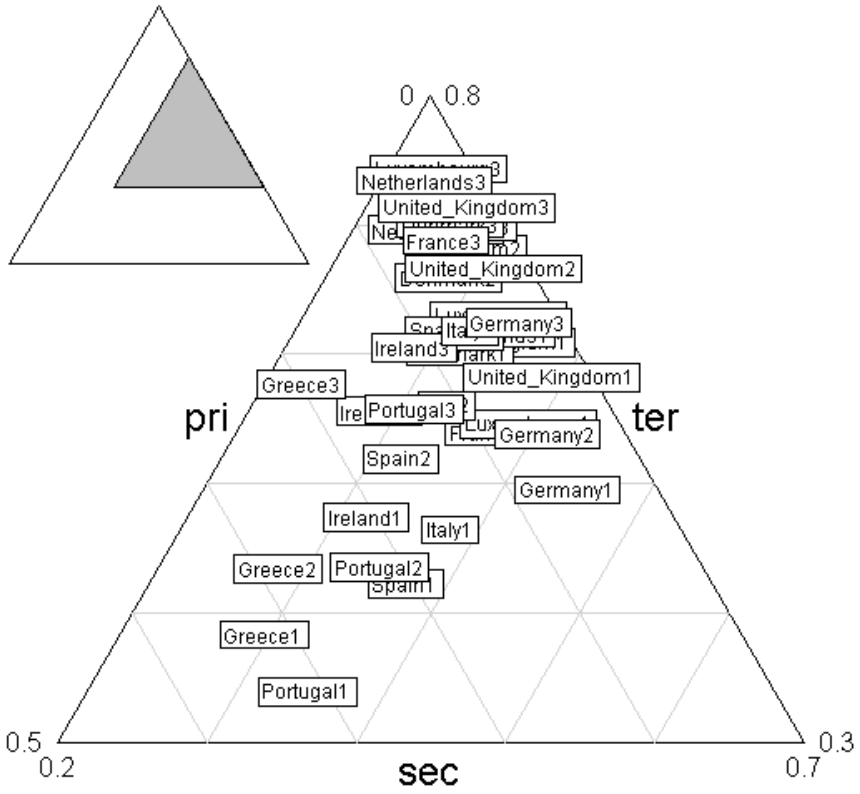


FIG. 1 – La représentation triangulaire exprime toute l'information qui concerne la variabilité d'une distribution de fréquence à trois catégories.

```
par(mfrow = c(2, 2))
triangle.plot(euro123$in78, clab = 0, cpoi = 2, addmean = TRUE,
    show = FALSE)
triangle.plot(euro123$in86, label = row.names(euro123$in78),
    clab = 0.8)
triangle.biplot(euro123$in78, euro123$in86)
triangle.plot(rbind.data.frame(euro123$in78, euro123$in86),
    clab = 1, addaxes = TRUE, sub = "Principal axis",
    csub = 2, possub = "topright")
```

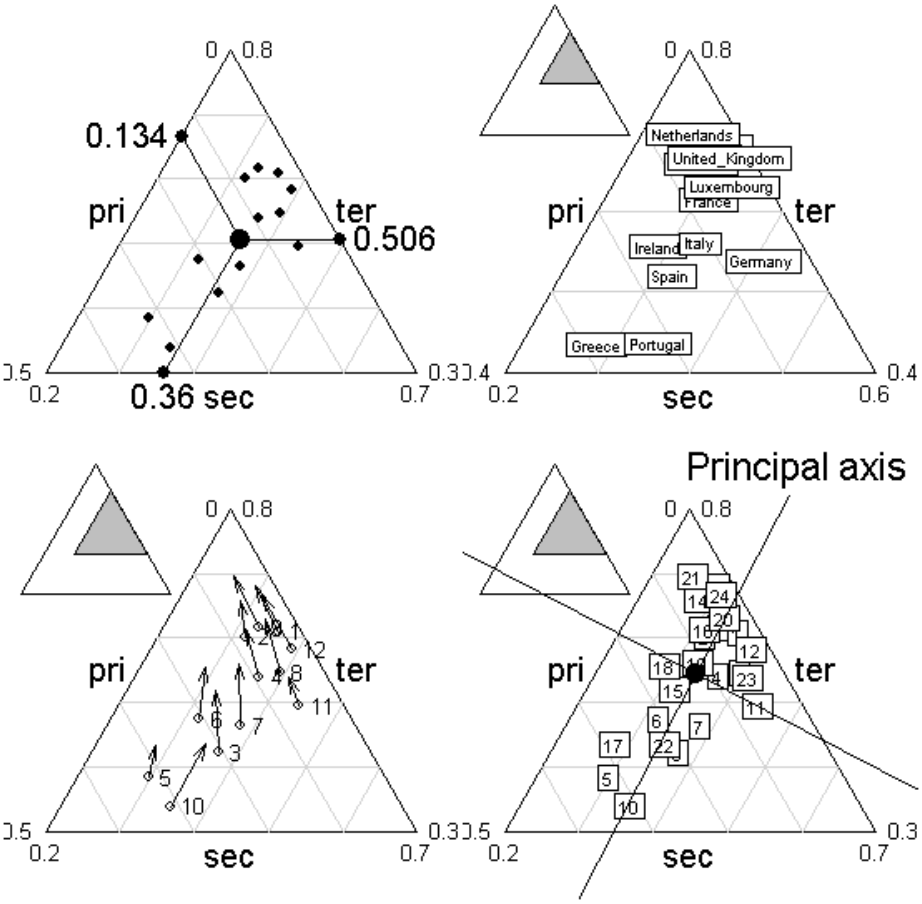


FIG. 2 – La représentation des points dans une représentation triangulaire est l'introduction naturelle à la carte factorielle. La représentation est purement technique et ne préjuge en rien du discours qu'elle supporte. Parler d'une réalité en face d'une image euclidienne est ce qu'on appelle l'interprétation. C'est un lieu charnière où l'origine technique de la figure et la nature expérimentale de l'approche de la réalité s'enrichissent réciproquement.

```
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 = TRUE)
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 = TRUE)
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 = TRUE)
triangle.plot(rbind.data.frame(euro123[[1]], euro123[[2]],
    euro123[[3]]))
```

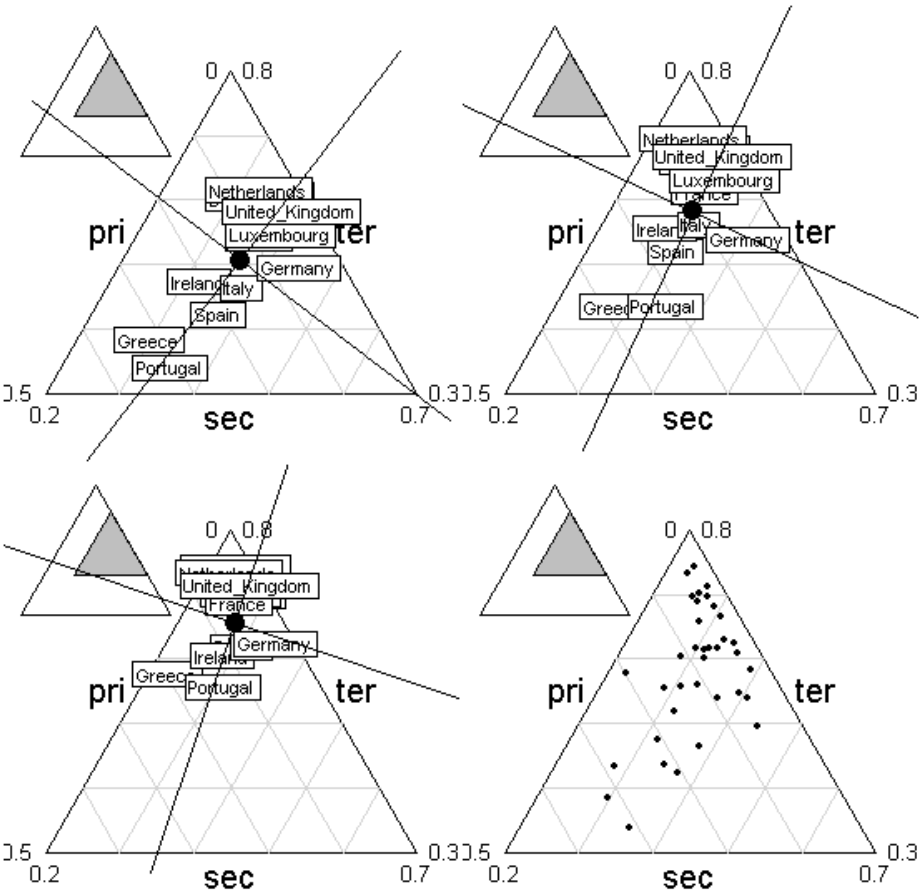


FIG. 3 – La question du choix entre typologie d'évolutions, ce qui distingue les trajectoires de chaque unité, ou évolution d'une typologie, ce qui est modifiée dans l'ensemble des positions relatives des points entre eux, est posée dans cette exemple très simple. Elle justifie la logique de la statistique multi-tableaux.

```
data(elec88)
par(mfrow = c(2, 2))
area.plot(elec88$area, cpoint = 1)
area.plot(elec88$area, lab = elec88$lab[, 1], clab = 0.75)
area.plot(elec88$area, clab = 0.75)
area.plot(elec88$area, graph = elec88$neig, sub = "Neighbourhood graph",
    possub = "topright")
```

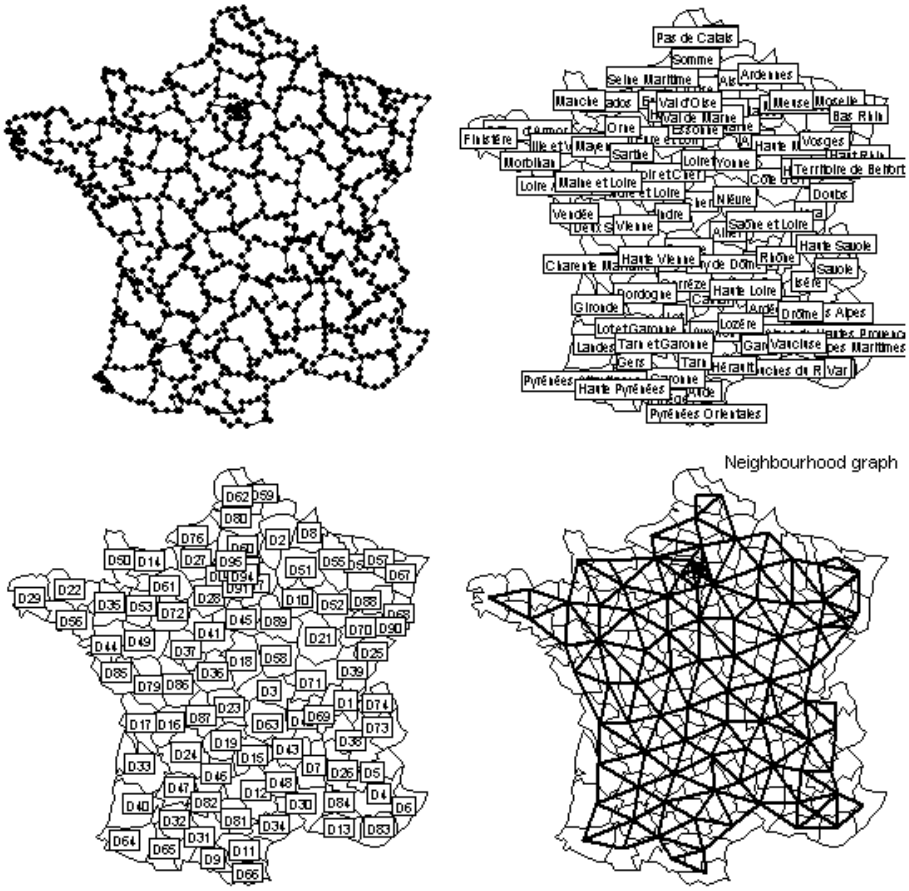


FIG. 4 – Les unités surfaciques : définition comme polygone de points, étiquetage en clair ou en mode réduit, graphe de voisinage.

```
par(mfrow = c(3, 3))
for (i in 1:3) {
    x <- elec88$tab[, i]
    area.plot(elec88$area, val = x, clab = 0, sub = names(elec88$tab)[i],
        csub = 3)
}
for (i in 4:6) {
    x <- elec88$tab[, i]
    s.value(elec88$xy, x, contour = elec88$contour, meth = "greylevel",
        sub = names(elec88$tab)[i], csub = 3, cleg = 1.5,
        incl = FALSE)
}
for (i in 7:9) {
    x <- scale(elec88$tab[, i])
    s.value(elec88$xy, x, contour = elec88$contour, meth = "squaresize",
        sub = names(elec88$tab)[i], csi = 1.5, csub = 3,
        cleg = 1.5, incl = FALSE)
}
```

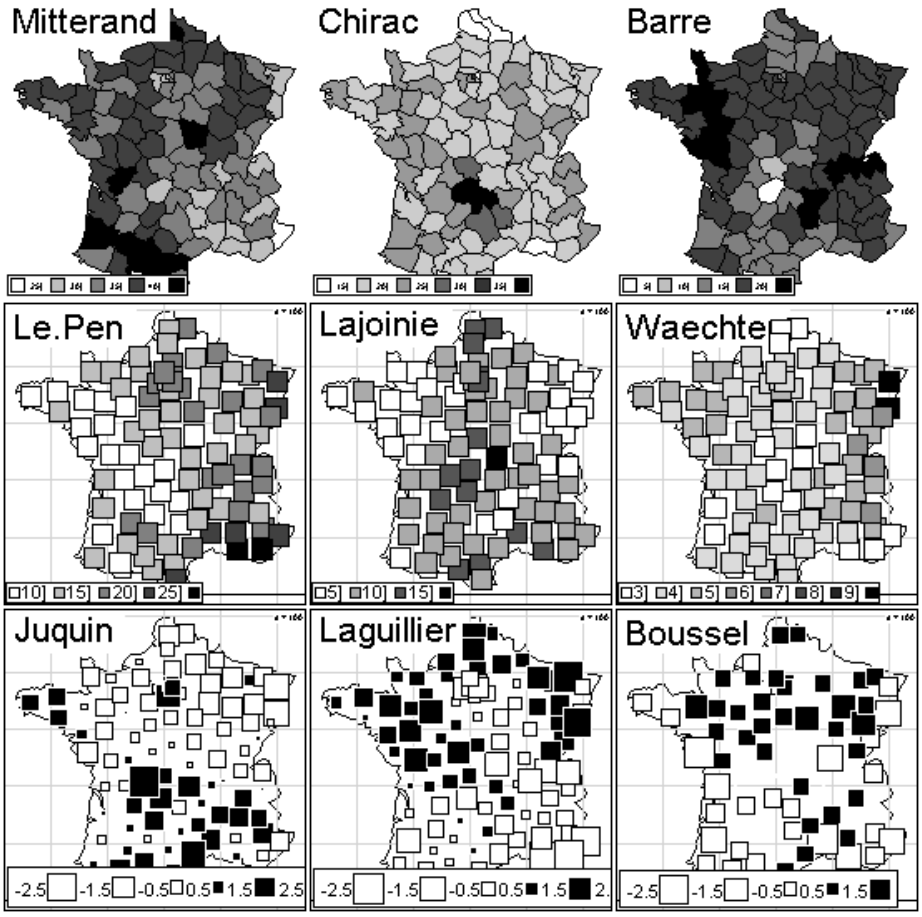


FIG. 5 – Cartographie et coordonnées : par unités surfaciques, par niveaux de gris, par taille de carrés. Noter que chaque carte se lit indépendamment des autres et que l'information ne porte que sur la structure spatiale de chacun des résultats d'un candidat. On obtient un résultat radicalement différent si l'échelle est la même pour chaque fenêtre. Encore une question d'objectifs.

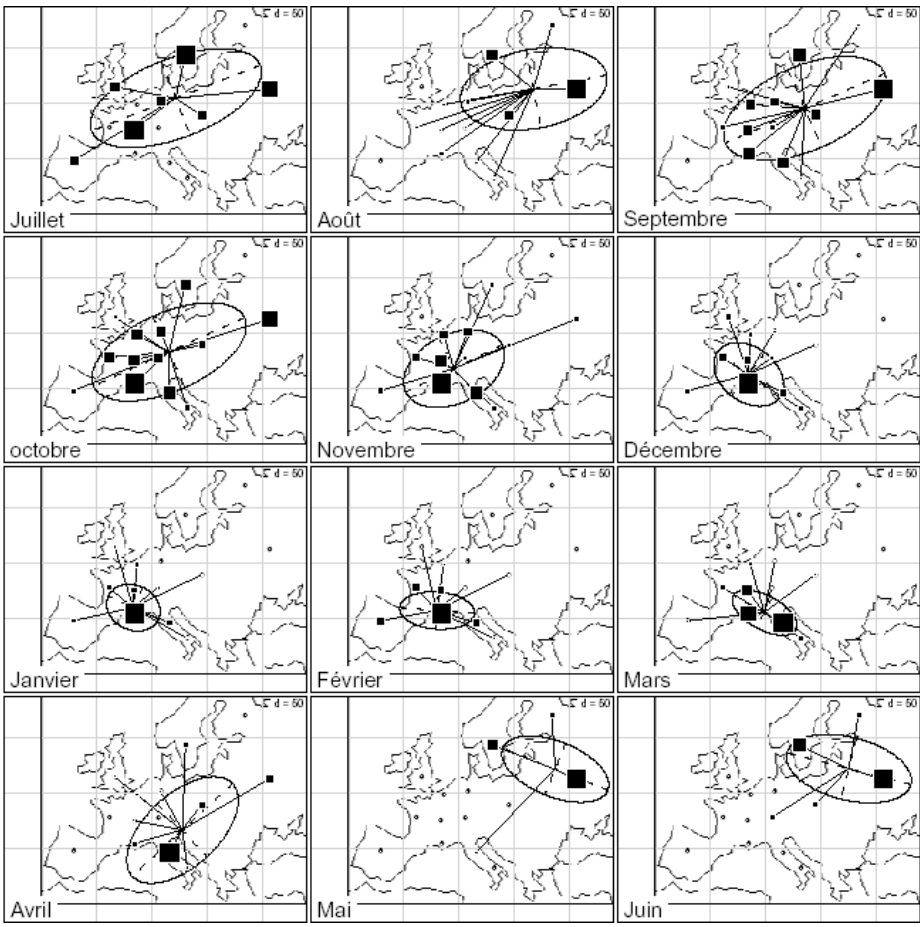


FIG. 6 – Migration des Sarcelles d'hiver. Pour chaque mois, la distribution des bagues renvoyées est schématisée par une ellipses d'inertie.

1. *Représenter des classes : figure 7*

```
xy <- cbind.data.frame(x = runif(200, -1, 1), y = runif(200,
    -1, 1))
posi <- factor(xy$x > 0):factor(xy$y > 0)
coul <- c("black", "red", "green", "blue")
par(mfrow = c(2, 2))
s.class(xy, posi, cpoi = 2)
s.class(xy, posi, cell = 0, cstar = 0.5)
s.class(xy, posi, cell = 2, axesell = FALSE, csta = 0,
    col = coul)
s.chull(xy, posi, cpoi = 1)
par(mfrow = c(1, 1))
```

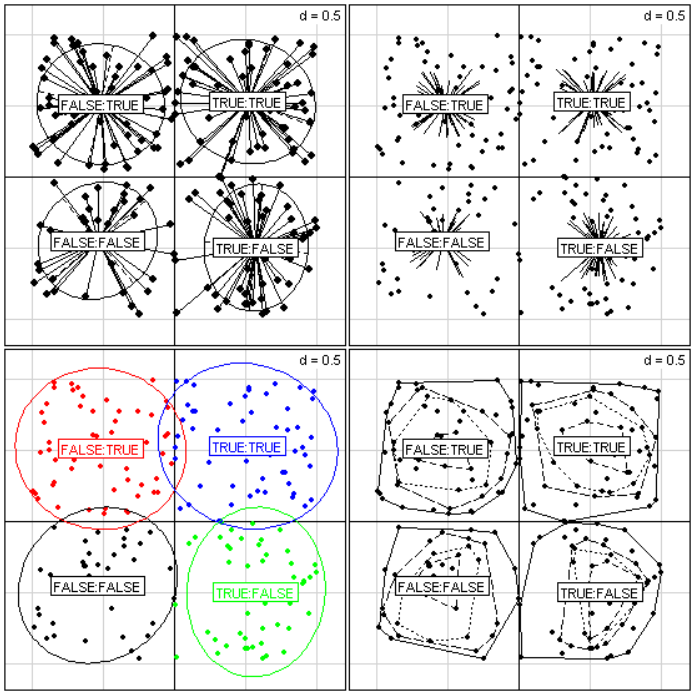


FIG. 7 – Représentation de classes de points sur une carte factorielle, par étoiles, par ellipses de dispersion, par polygones de contour.

2. *Représenter des valeurs : figure 8*

3. *Associer des couples de points : figure 9*

4. *Représenter des distributions de fréquences : figure 10*

5. *Représenter des trajectoires : figure 11*

6. *Voir les tableaux : figure 12*

7. *Voir les couples de tableaux : figure 13*

```
data(tarentaise)
w1 = dudi.acm(tarentaise$envir, scannf = F, nf = 3)
s.value(w1$li[, 1:2], w1$li[, 3], csi = 0.75)
```

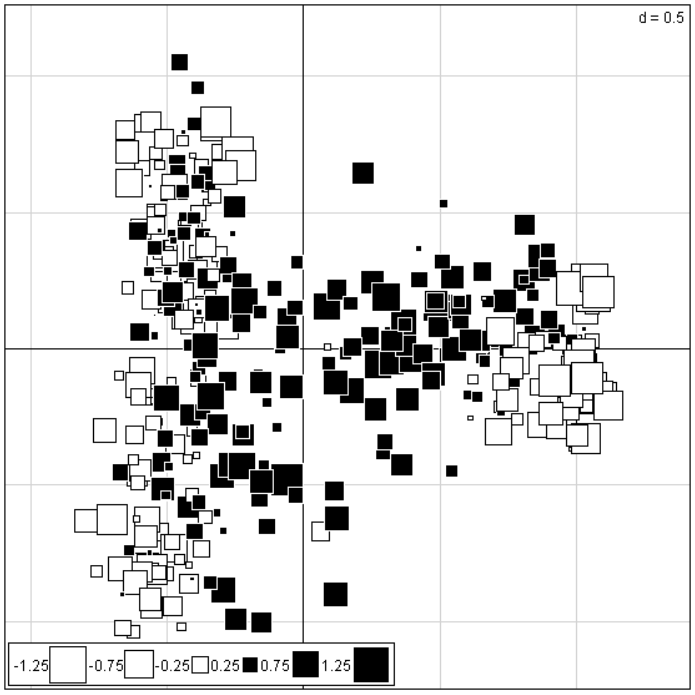


FIG. 8 – Représentation de valeurs sur une carte factorielle : la troisième coordonnée d'une analyse des correspondances multiples est éditée dans le plan des deux premières.

```
X <- data.frame(x = runif(50, -1, 2), y = runif(50, -1,
    2))
Y <- X + rnorm(100, sd = 0.3)
par(mfrow = c(2, 2))
s.match(X, Y)
s.match(X, Y, edge = FALSE, clab = 0)
s.match(X, Y, edge = FALSE, clab = 0)
s.label(X, clab = 1, add.plot = TRUE)
s.label(Y, clab = 0.75, add.plot = TRUE)
s.match(Y, X, clab = 0)
```

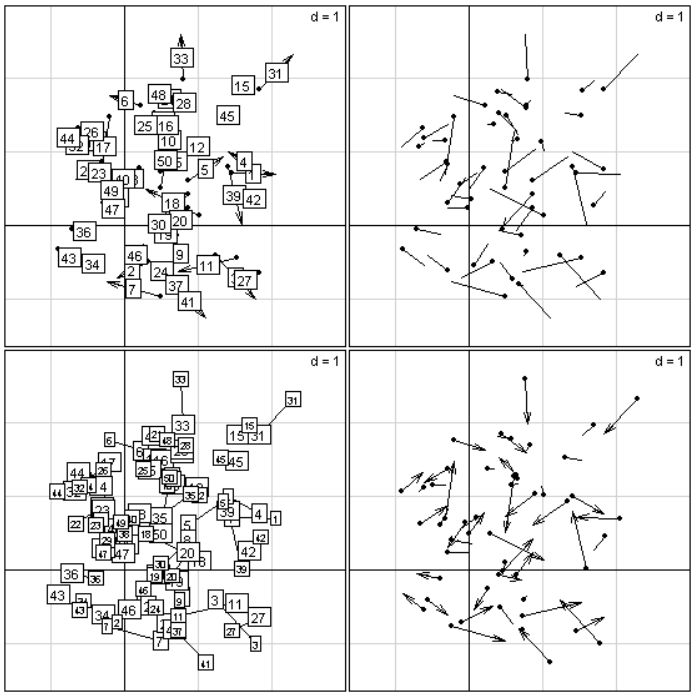


FIG. 9 – Représentation de nuages appariés : une entité présente deux fois sur la carte donne un segments. C'est la représentation de base dans les méthodes de couplages de tableaux.

```
xy <- cbind.data.frame(x = runif(200, -1, 1), y = runif(200,
    -1, 1))
distri <- data.frame(w1 = rpois(200, xy$x * (xy$x > 0)))
s.value(xy, distri$w1, cpoi = 1)
s.distri(xy, distri, add.p = TRUE)
```

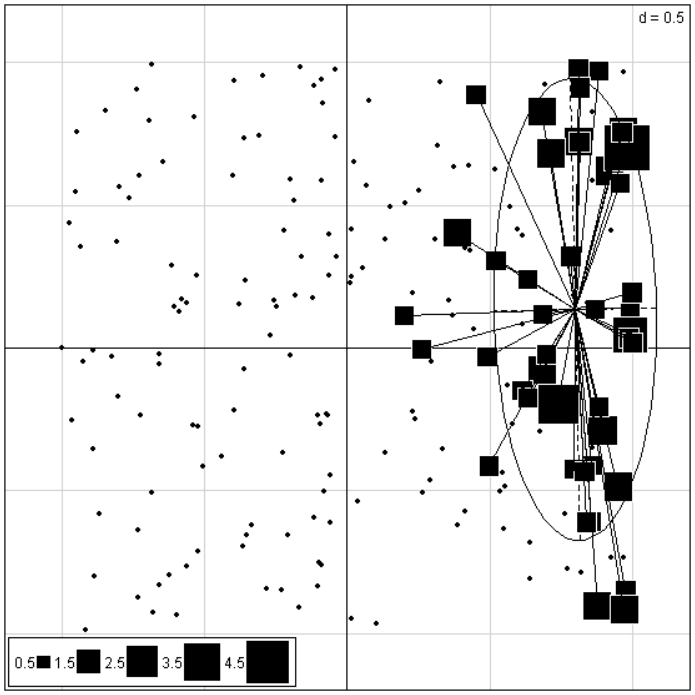


FIG. 10 – Représentation de distributions de fréquence. C'est le graphe de base dans les méthodes qui utilisent l'*averaging*, c'est-à-dire qui positionne les espèces comme distribution de fréquences à la moyenne des relevés qu'elles occupent [12].

```
rw <- function(a) {
    x <- 0
    for (i in 1:49) x <- c(x, x[length(x)] + runif(1,
        -1, 1))
    x
}
y <- unlist(lapply(1:5, rw))
x <- unlist(lapply(1:5, rw))
z <- gl(5, 50)
s.traject(data.frame(x, y), z, edge = FALSE)
```

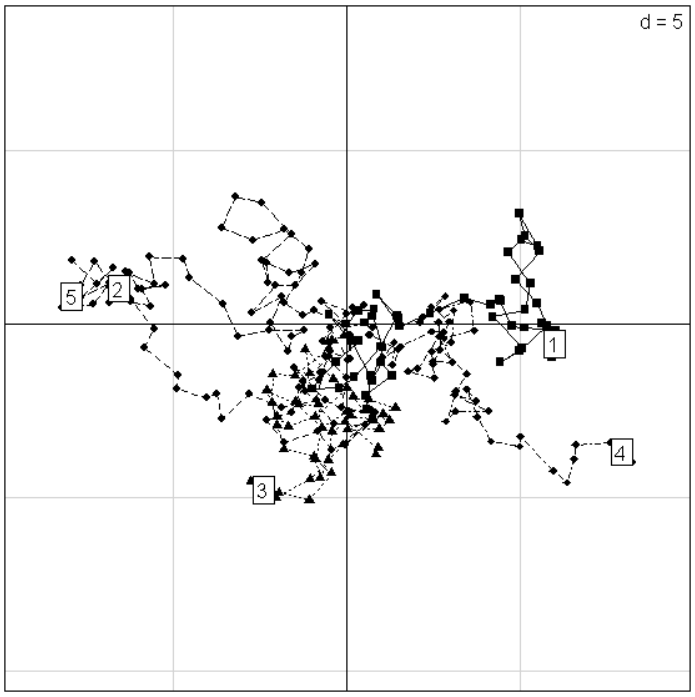


FIG. 11 – Représentation de trajectoires : une autre manière de placer une classe sur une carte factorielle. Elle sert quand l'ordre des points a un sens, spatial, temporel ou autre.

```
data(olympic)
w <- olympic$tab
w <- data.frame(scale(w))
wpca <- dudi.pca(w, scann = FALSE)
par(mfrow = c(1, 3))
table.value(w, csi = 2, clabel.r = 2, clabel.c = 2)
table.value(w, y = rank(wpca$li[, 1]), x = rank(wpca$co[,
    1]), csi = 2, clabel.r = 2, clabel.c = 2)
table.value(w, y = wpca$li[, 1], x = wpca$co[, 1], csi = 2,
    clabel.r = 2, clabel.c = 2)
par(mfrow = c(1, 1))
```

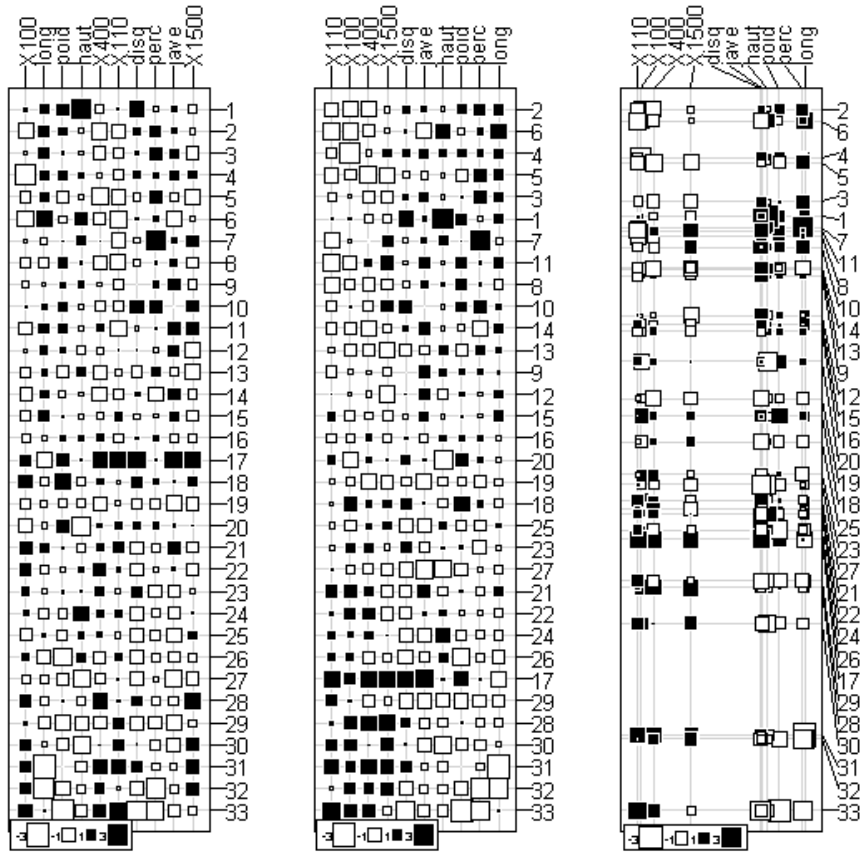


FIG. 12 – Représentation des valeurs d'un tableau individus-lignes, variables-colonnes. La représentation de gauche est celle du tableau centré-réduit. Au milieu, les lignes et les colonnes sont ordonnées par la première coordonnée de l'ACP. A droite les valeurs des coordonnées positionnent les objets. Cette technique est très utilisée en écologie.

```
data(rpjdl)
X <- data.frame(t(rpjdl$fau))
Y <- data.frame(t(rpjdl$mil))
layout(matrix(c(1, 2, 2, 2, 1, 2, 2, 2, 1, 2, 2, 2, 1,
    2, 2, 2), 4, 4))
coa1 <- dudi.coa(X, scan = FALSE)
x <- rank(coa1$co[, 1])
y <- rank(coa1$li[, 1])
table.paint(Y, x = x, y = 1:8, clabel.c = 0, cleg = 0)
abline(v = 114.9, lwd = 3, col = "red")
abline(v = 66.4, lwd = 3, col = "red")
table.paint(X, x = x, y = y, clabel.c = 0, cleg = 0,
    row.lab = paste("  ", row.names(X), sep = ""))
abline(v = 114.9, lwd = 3, col = "red")
abline(v = 66.4, lwd = 3, col = "red")
```

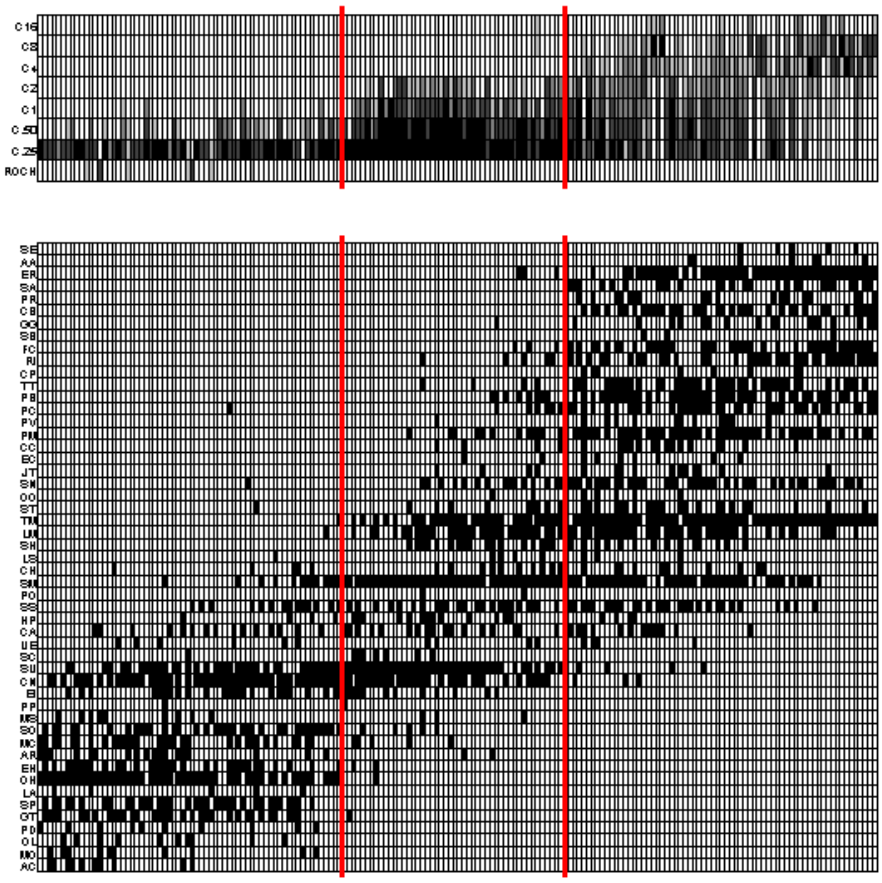


FIG. 13 – Représentation de tableaux[20] lignes=variables, colonnes=individus. Le tableau faunistique est ordonné, pour les lignes comme pour les colonnes, par la première coordonnée de l'AFC (en bas). Le tableau sites-variables au-dessus est permuté pour que les deux tableaux soient appariés.

8. *Voir une variable et une phylogénie : figure 14*

9. *Voir un tableau et une phylogénie : figure 15*

Il convient d'abandonner l'idée que l'analyse des données est une question de cartes factorielles avec des étiquettes.

```
data(mjrochet)
mjrochet.phy <- newick2phylog(mjrochet$tre)
tab <- log((mjrochet$tab))
tab0 <- data.frame(scalewt(tab))
par(mfrow = c(2, 3))
for (j in 1:6) {
    w <- tab0[, j]
    dotchart.phylog(mjrochet.phy, w, cdot = 1.5, sub = names(tab0)[j],
        csub = 3, cnodes = 2, ceti = 1.5)
}
par(mfrow = c(1, 1))
```

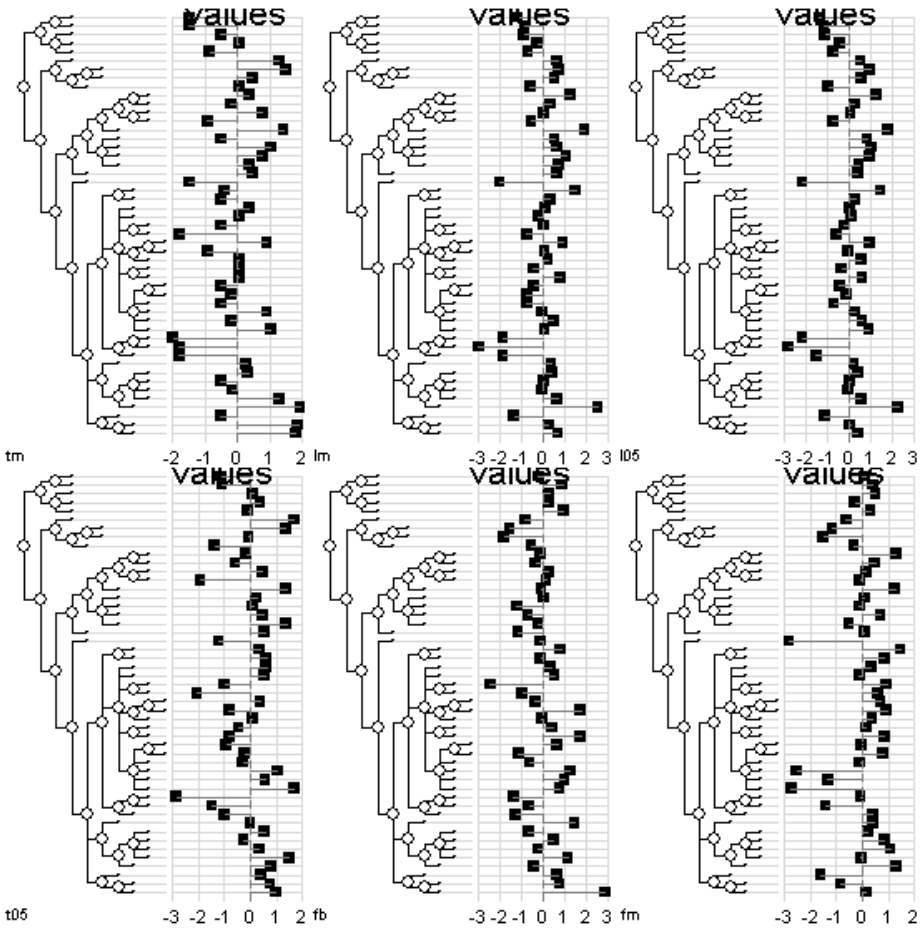


FIG. 14 – Le graphique par points représente une valeur qui a une étiquette[3][4]. C'est particulièrement adapté à la représentation de la valeur d'un trait biologique portée par une espèce.

```
data(mjrochet)
table.phylog(tab0, mjrochet.phy, csi = 2, clabel.r = 0.75)
```

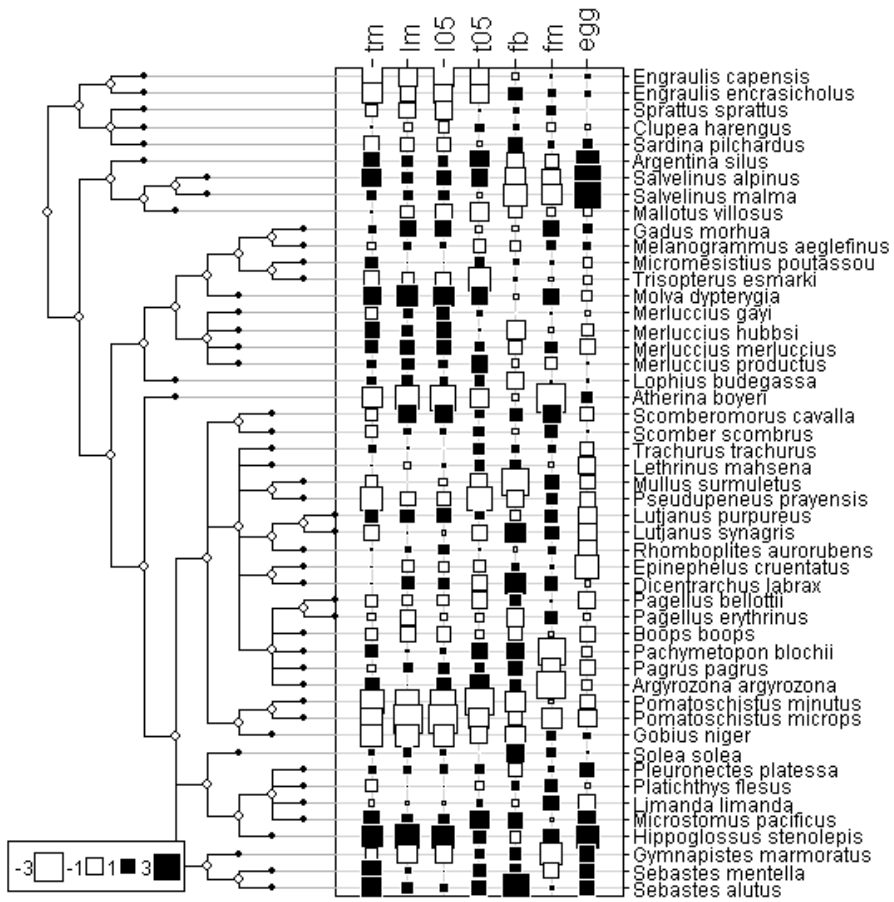


FIG. 15 – La représentation d'un tableau [22] avec une phylogénie sur une des marges. La biologie évolutive pose des problèmes ouverts pour l'analyse de ses données.

2 Qu'est-ce que l'analyse en composantes principales ?

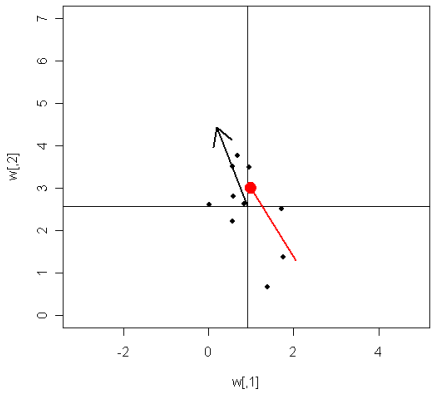
Depuis [19], tout le monde est d'accord sur comment faire le calcul, mais pas du tout sur ce qu'il signifie.

2.1 Estimer les paramètres d'une loi normale multivariée

Traçons l'axe principal d'une matrice de covariances à deux dimensions et son estimation par un échantillon aléatoire de n points. A ce jeu, il vaut mieux beaucoup plus d'individus que de variables.

```
library(MASS)
f1 <- function(n) {
    Sigma <- matrix(c(1, -1, -1, 2), 2, 2)
    w = mvrnorm(n, c(1, 3), Sigma)
    mx <- mean(w[, 1])
    my <- mean(w[, 2])
    plot(w, asp = 1, ylim = c(0, 7), pch = 19)
    abline(v = mx, h = my)
    points(1, 3, col = "red", pch = 19, cex = 2)
    print(vrai <- eigen(Sigma))
    print(esti <- eigen(var(w)))
    print(prcomp(w))
    print(princomp(w))
    arrows(mx, my, mx + 2 * esti$vectors[1, 1], my +
        2 * esti$vectors[2, 1], lwd = 2)
    segments(1, 3, 1 + 2 * vrai$vectors[1, 1], 3 + 2 *
        vrai$vectors[2, 1], lwd = 2, col = "red")
}
f1(10)
```

```
$values
[1] 2.618034 0.381966

$vectors
           [,1]      [,2]
[1,]  0.5257311 0.8506508
[2,] -0.8506508 0.5257311

$values
[1] 1.4793490 0.4996972

$vectors
           [,1]      [,2]
[1,] -0.4286789 0.9034569
[2,]  0.9034569 0.4286789

Standard deviations:
[1] 1.2162849 0.7068926

Rotation:
           PC1       PC2
[1,] -0.4286789 0.9034569
[2,]  0.9034569 0.4286789
Call:
princomp(x = w)

Standard deviations:
   Comp.1    Comp.2
1.1538692 0.6706172

 2  variables and  10 observations.
```



2.2 La variable cachée des psychométriciens [11]

Trouver une variable cachée qui prédit les variables mesurées : la composante principale (figure16). Les composantes d'une acp sont `tab`, `cw`, `lw` ...

2.3 Faire une typologie de variable

Les individus sont des échantillons sans personnalité : ils sont là pour mesurer la cohérence des variables. Il s'agit de résumer une matrice de corrélation (figure17).

2.4 Les valeurs propres : un élément fondamental

Il n'y a pas de bonnes ou de mauvaises analyses. Il n'y a que des chercheurs qui acceptent ou qui refusent de se laisser faire par les résultats (figure 18).

2.5 Faut-il centrer, normaliser, transformer ?

La question est très difficile. Le *devoir faire* est une erreur. Le relation entre l'algorithme et la donnée est extrêmement riche [18]. On peut faire de l'ACP décentrée (figure 19), non centrée, doublement centrée, à centrage additif ou multiplicatif.

2.6 ACM : l'équivalent pour des variables qualitatives

L'Analyse des Correspondances multiples [24] étend ces pratiques aux variables qualitative (`dudi.acm`) et le mélange des variables est autorisé dans `dudi.mix` ou `dudi.hillsmith`. Les variables peuvent être encore distributionnelles ou floues (`dudi.fca`).

3 L'Analyse des correspondances

L'analyse des correspondances n'est voisine qu'au plan algébrique. Au plan statistique la rupture est radicale.

```
data(lascaux)
score(dudi.pca(na.omit(lascaux$morpho), scan = F), csub = 1)
```

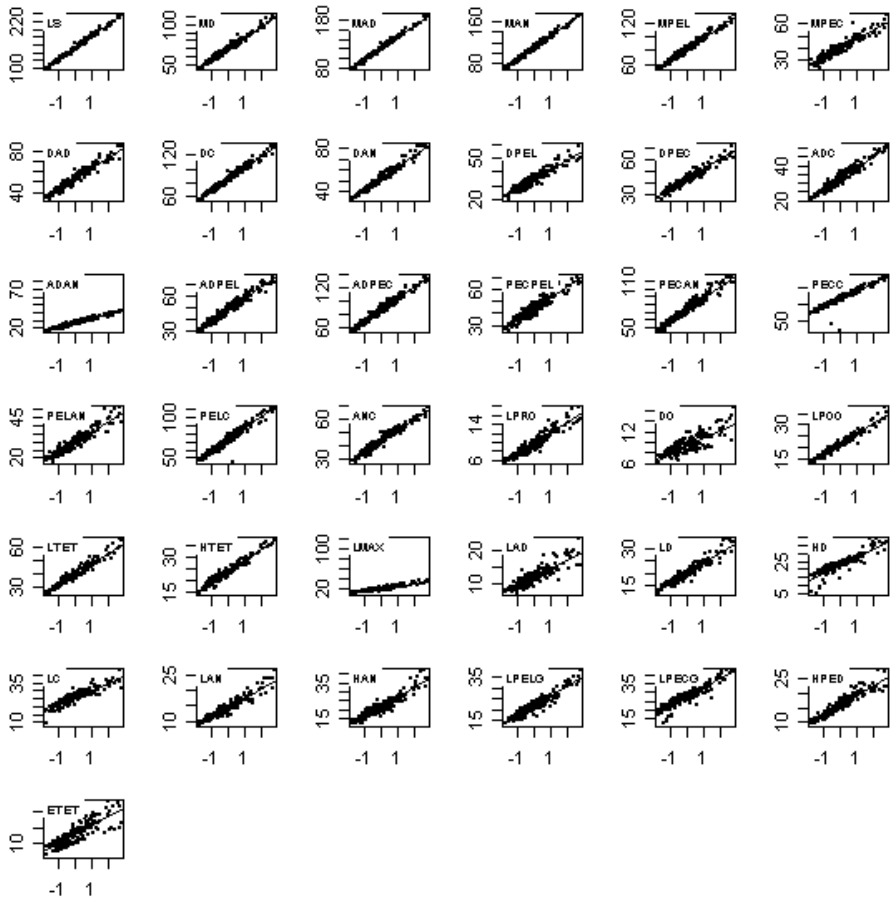


FIG. 16 – Une même variable (la première composante principale) maximise la somme des carrés des corrélations avec un ensemble de variables. Ici la taille synthétique prédit les mensurations (effet taille). On repère des données aberrantes.

```
par(mfrow = c(2, 2))
data(fruits)
s.corcircle(dudi.pca(fruits$jug, scan = FALSE)$co)
data(macon)
s.corcircle(dudi.pca(macon, scan = FALSE)$co)
data(rankrock)
s.corcircle(dudi.pca(rankrock, scan = FALSE)$co)
data(olympic)
s.corcircle(dudi.pca(olympic$tab, scan = FALSE)$co)
```

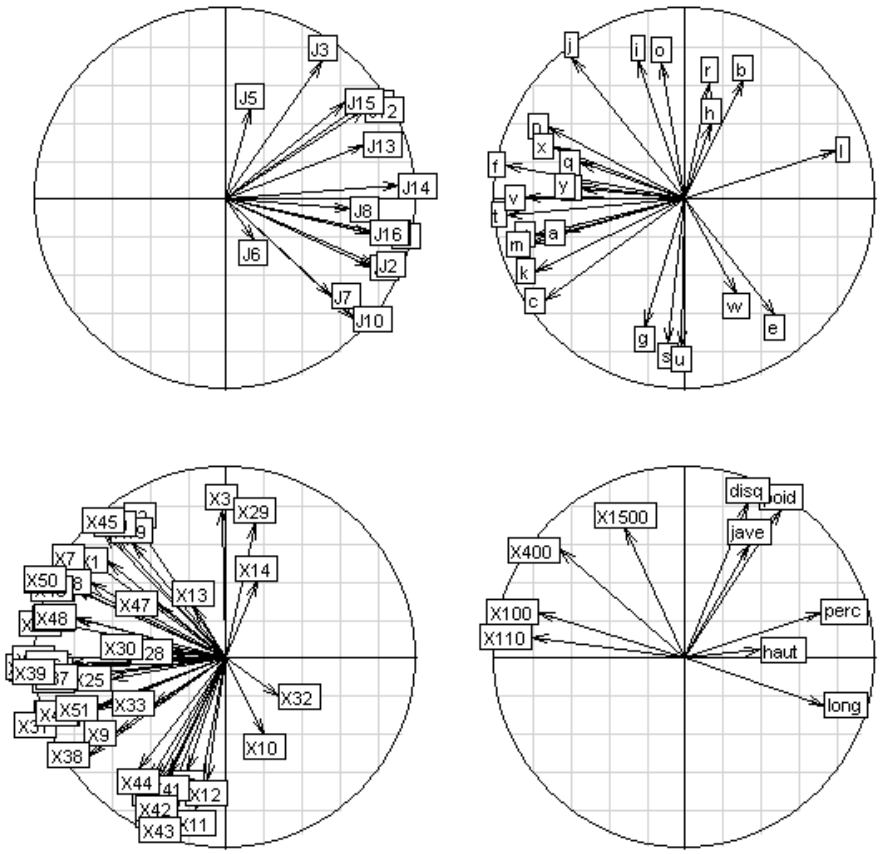


FIG. 17 – Cercles de correlation en ACP normée. Relier la logique de la représentation (projection euclidienne de points sur une sphère) et la logique de la réalité (performances sportives, concours de dégustation, cohérence des jurys, différences des gouts) n'est pas toujours simple.

```
data(lascaux)
par(mfrow = c(2, 2))
barplot(dudi.pca(na.omit(lascaux$morpho), scan = F)$eig)
barplot(dudi.pca(na.omit(lascaux$meris), scan = F)$eig)
barplot(dudi.pca(na.omit(lascaux$colo), scan = F)$eig)
data(olympic)
barplot(dudi.pca(olympic$tab, scan = F)$eig)
```

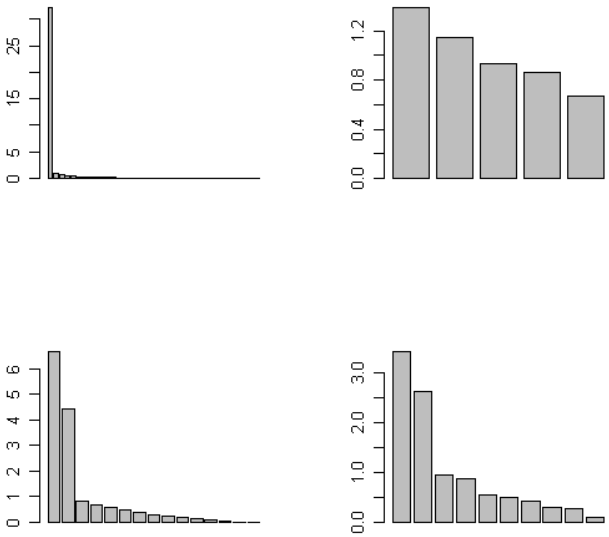


FIG. 18 – Graphes de valeurs propres. Différents types de variables sont étudiés par J.M. Lascaux [13]. En morphométrie, l'effet taille est écrasant et doit être isolé. Les variables méristiques sont sans structure marquée. Les variables de coloration forment deux paquets. Les variables sportives sont plus subtiles : les temps des courses diminuent avec la valeur des athlètes, les distances des sauts ou des lancers augmentent.

```
data(deug)
pca1 <- dudi.pca(deug$tab, scal = FALSE, center = deug$cent,
    scan = FALSE)
s.class(pca1$li, deug$result)
s.arrow(40 * pca1$c1, add.plot = TRUE)
```

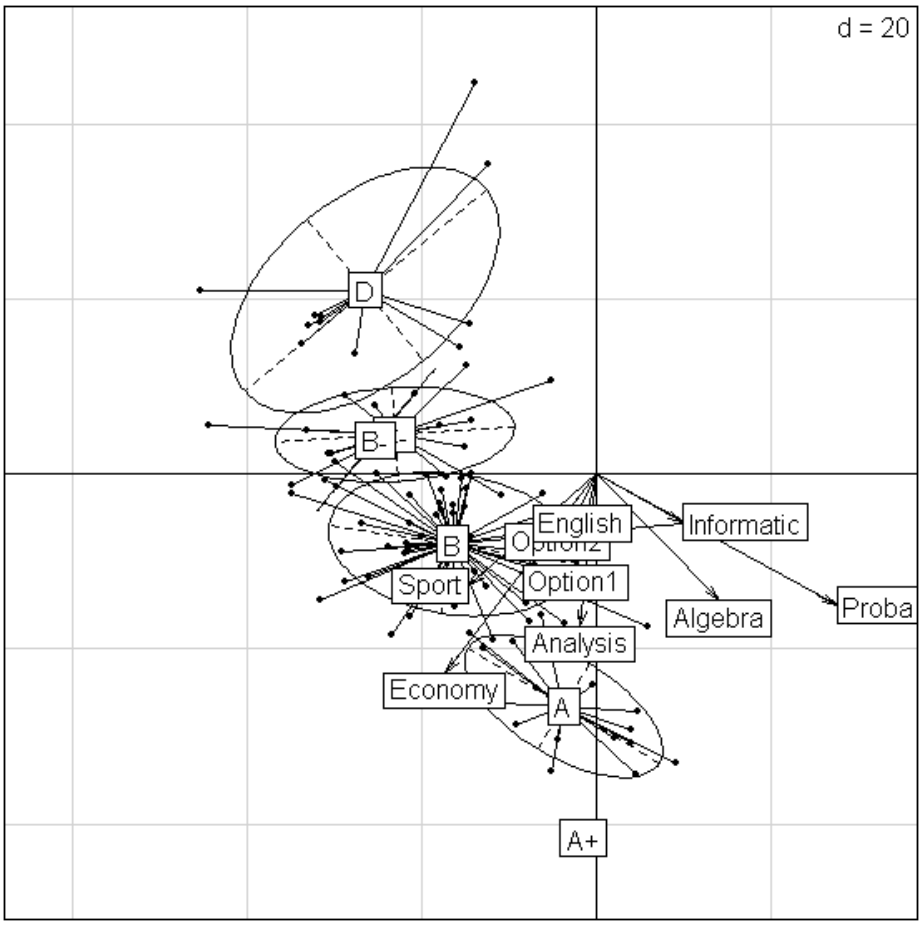


FIG. 19 – ACP décentrée. L'origine est l'étudiant mythique qui a exactement la moyenne dans chaque matière. Les données sont les points d'avance et de retard que comptent les étudiants. Les professeurs sympathiques décalent le nuage, les autres l'étalent. C'est un point de vue sur un tableaux de notes d'examen !

3.1 Les individus sont dans les cases du tableau

L'AFC est d'abord conçue pour des tables de contingence (figure 20)

3.2 Les occurrences sont des individus statistiques

La case non vide d'un tableaux de nombres positifs ou nuls est un individu : la valeur de la case est son poids. Écologiquement, le déplacement est très signifiant (figure 21). Disposer les espèces pour maximiser la variance des moyennes conditionnelles par sites, disposer les sites pour maximiser la variance des moyennes conditionnelles par espèces, disposer les occurrences pour optimiser ensemble et également la variance des moyennes par site (maximiser la diversité beta) et la variance des moyennes conditionnelles par espèces (minimiser le chevauchement de niche).

Avantage : on ne dit pas ce qu'on a à faire (typologie des sites par les espèces, typologie des espèces par les sites). Inconvénient : l'objectif qu'on veut ignorer est à l'oeuvre sans qu'on le sache (et on paye le prix).

3.3 l'AFC est une double analyse discriminante

L'objectif est de caractériser des structures (figure 22). L'analyse des données est seule à faire cette opération [**?**].

4 Les stratégies de couplage de tableaux

4.1 Assemblages de tableaux

On peut faire l'analyse des deux tableaux accolés[5], c'est le premier pas vers l'analyse factorielle multiple (figure 23). On peut faire l'analyse du premier et illustrer par le second ou inversement. Voir stage5.

4.2 Analyses canoniques

L'analyse canonique[6]ne peut se pratiquer qu'en dimensions limitées mais justifie les stratégies sur composantes (figure24). Elle considère chaque tableau comme un paquet de variables. L'AFC est une analyse canonique dont les individus sont les ocurrences. L'analyse discriminante est une analyse canonique.

4.3 Variables instrumentales

Les ACPVI (figure 25) considèrent un tableau comme un paquet de variables (les variables instrumentales) et l'autre comme un paquet d'individus (dont on cherche à résumer la typologie)[21]. La plus connue, la plus utilisée (CANOCO) et la plus difficile des méthodes de couplage est la CCA [25] ou AFCVI[15]. Les inter-classes (`between` et les intra-classes (`within`) font partie de cette famille. Voir `avimedi` pour l'AFCVI. La librairie spécialisée est `vegan` de J. Oksanen.

```
data(chats)
chatsw <- data.frame(t(chats))
chatscoa <- dudi.coa(chatsw, scann = FALSE)
par(mfrow = c(2, 2))
table.cont(chatsw, abmean.x = TRUE, csi = 2, abline.x = TRUE,
    clabel.r = 1.5, clabel.c = 1.5)
table.cont(chatsw, abmean.y = TRUE, csi = 2, abline.y = TRUE,
    clabel.r = 1.5, clabel.c = 1.5)
table.cont(chatsw, x = chatscoa$c1[, 1], y = chatscoa$l1[,
    1], abmean.x = TRUE, csi = 2, abline.x = TRUE, clabel.r = 1.5,
    clabel.c = 1.5)
table.cont(chatsw, , x = chatscoa$c1[, 1], y = chatscoa$l1[,
    1], abmean.y = TRUE, csi = 2, abline.y = TRUE, clabel.r = 1.5,
    clabel.c = 1.5)
par(mfrow = c(1, 1))
```

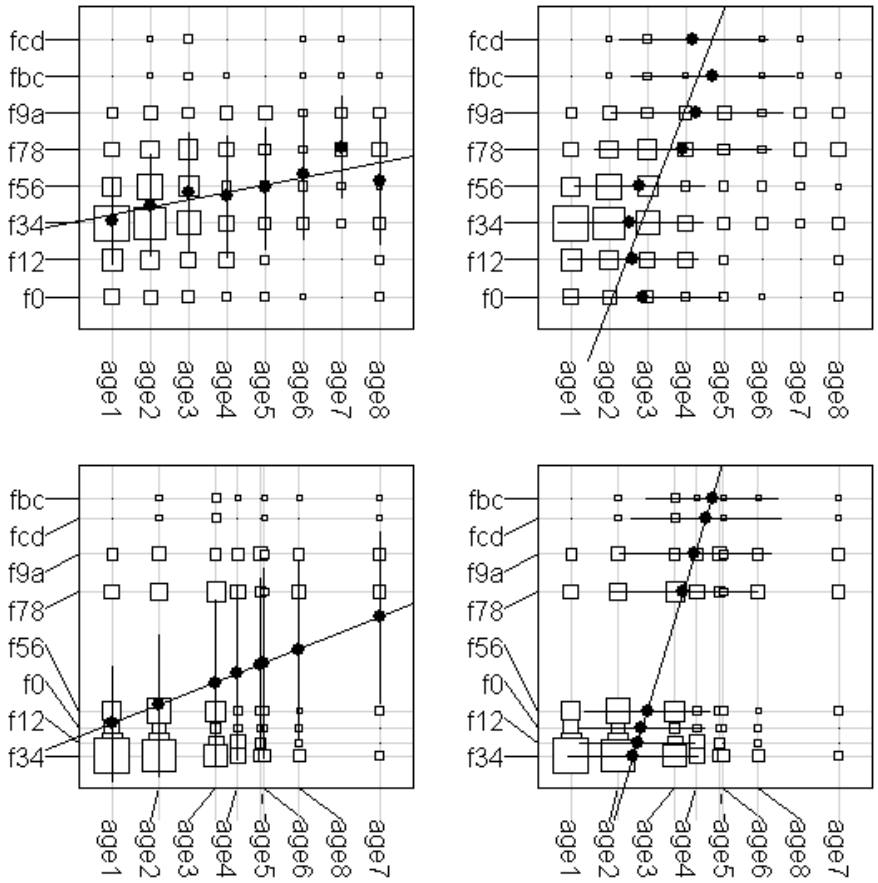


FIG. 20 – Une table de contingence croise des classes d'âge et des classes de fécondité annuelle pour des chattes domestiques [16]. En haut, droites et courbes de régression semblent résumer les données. En bas, les lignes et les colonnes sont positionnées par la première coordonnée de l'AFC. Le groupement des classes de fécondité indique la présence de la variable cachée *nombre de portées*. Les deux régressions sont linéaires [9].

```
layout(matrix(c(1, 1, 2, 3), 2, 2), resp = FALSE)
data(aviurba)
dd1 <- dudi.coa(aviurba$fau, scan = FALSE)
score(dd1, clab.r = 0, clab.c = 0.75)
abline(v = 1, lty = 2, lwd = 3)
sco.distri(dd1$l1[, 1], aviurba$fau)
sco.distri(dd1$c1[, 1], data.frame(t(aviurba$fau)))
```

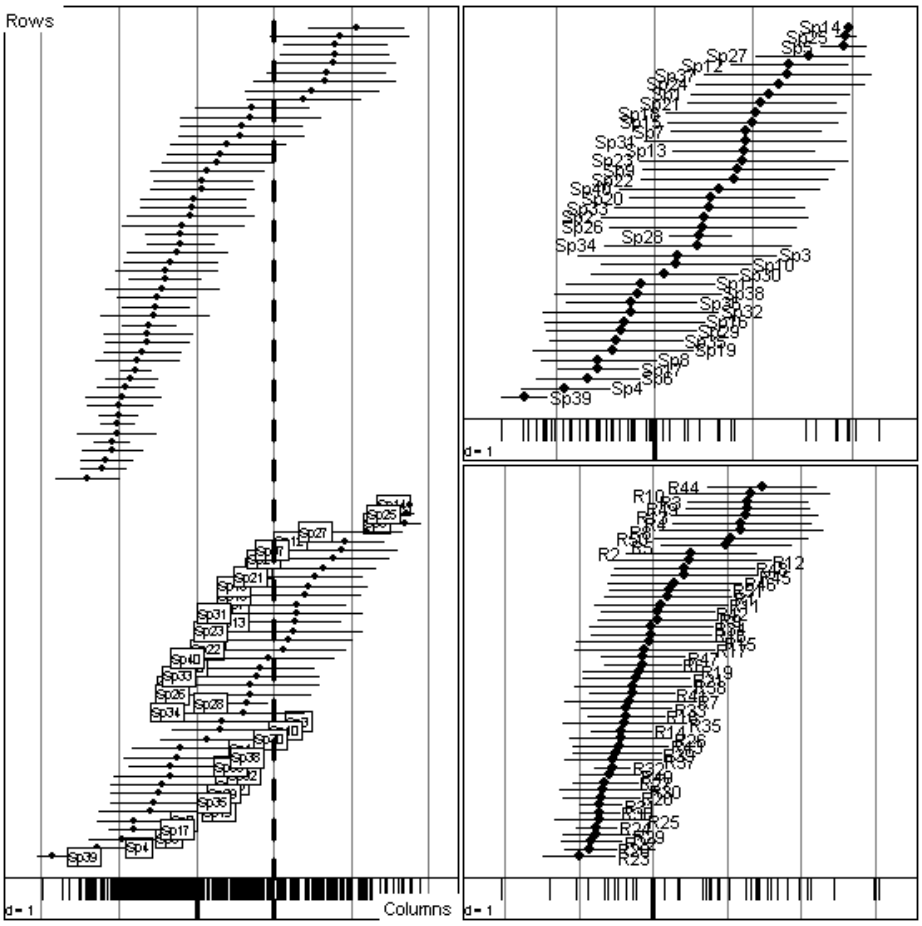


FIG. 21 – A gauche, l'AFC comme analyse canonique[27] : les occurrences (en bas) positionnent les lignes (relevés) et les colonnes (espèces) par des moyennes conditionnelles. A droite, l'AFC comme analyse discriminante des espèces par les relevés ou des relevés par les espèces.

```
data(avimedi)
prov = avimedi$fau[avimedi$plan$reg == "Pr", ]
cors = avimedi$fau[avimedi$plan$reg == "Co", ]
data(rpjdl)
alber = rpjdl$fau
par(mfrow = c(2, 2))
xy <- dudi.coa(alber, scann = FALSE)$l1
s.distri(xy, alber, 2, 1, cstar = 0.3, cell = 0, csub = 2,
    sub = "Albères")
xy <- dudi.coa(prov, scann = FALSE)$l1
s.distri(xy, prov, 2, 1, cstar = 0.3, cell = 0, csub = 2,
    sub = "Provence")
xy <- dudi.coa(cors, scann = FALSE)$l1
s.distri(xy, cors, 2, 1, cstar = 0.3, cell = 0, csub = 2,
    sub = "Corse")
```

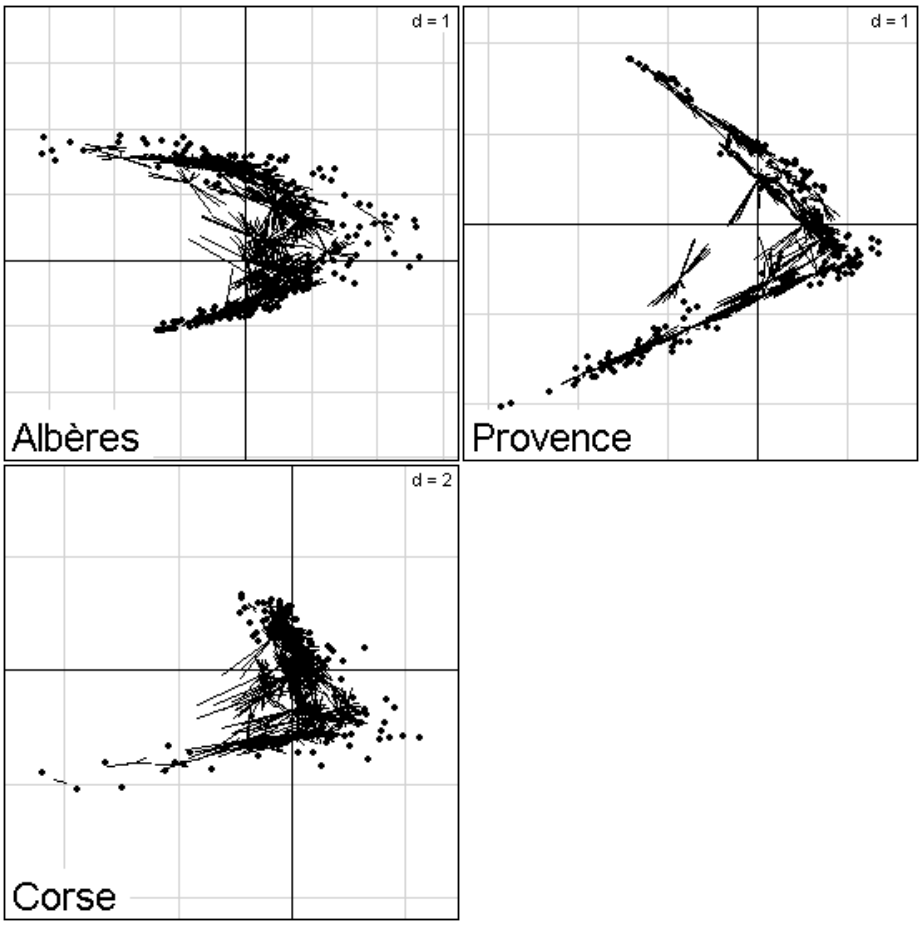


FIG. 22 – Les points sont des relevés, les étoiles sont des espèces. Chaque figure est une analyse de l'avifaune [20] [1]dans sa relation avec l'ouverture de la végétation. Chacun des espaces écologiques (isolé, continental, insulaire) réalise la relation avifaune-végétation avec originalité.

```
data(doubs)
w = cbind.data.frame(doubs$mil, doubs$poi)
pcaw = dudi.pca(w, scan = F)
par(mfrow = c(2, 2))
barplot(pcaw$eig)
s.traject(pcaw$li, clab = 0)
s.label(pcaw$li[c(1, 30), ], add.plot = T)
s.corcircle(pcaw$co[1:11, ])
s.corcircle(pcaw$co[12:38, ])
```

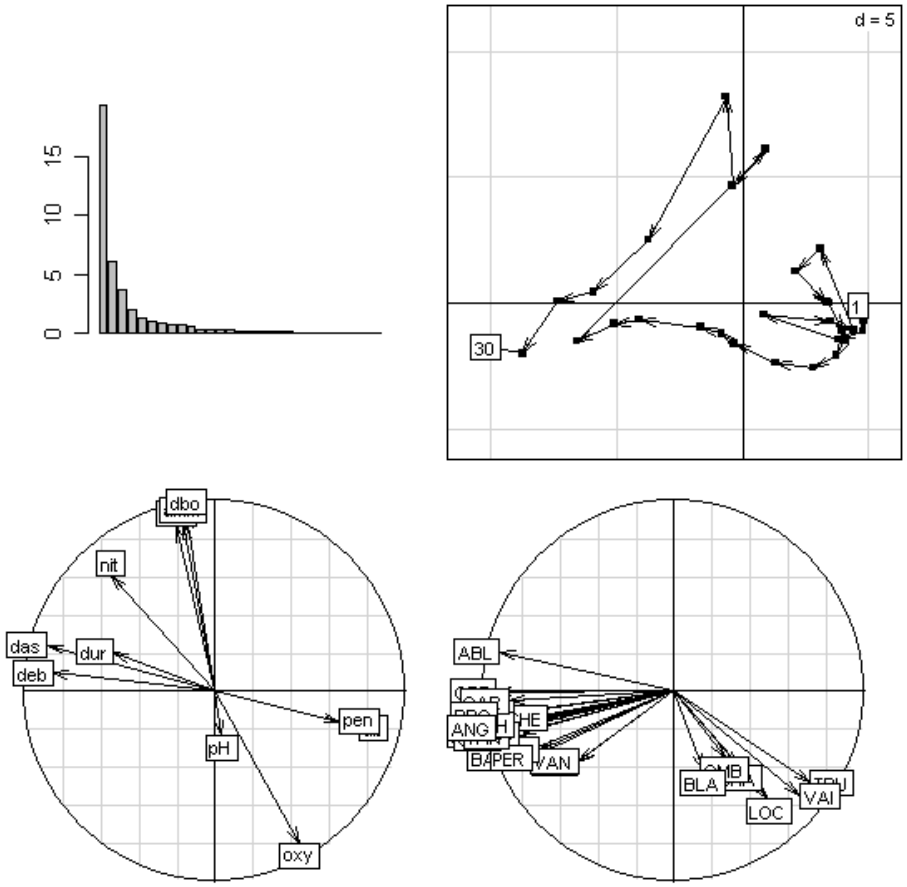


FIG. 23 – Un tableau sites-variables est accolé avec un tableau sites-espèces et le tout est soumis à une ACP normée. Données extraites de [26]

```
pcamil <- dudi.pca(doubs$mil, scannf = FALSE)
pcafau <- dudi.pca(doubs$poi, scal = F, scan = F)
can <- cancor(pcamil$li, pcafau$li)
scormil <- as.data.frame(as.matrix(pcamil$li) %*% can$xcoef)
scorfau <- as.data.frame(as.matrix(pcafau$li) %*% can$ycoef)
par(mfrow = c(2, 2))
s.traject(scormil)
s.label(scormil, add.p = T)
s.traject(scorfau)
s.label(scorfau, add.p = T)
plot(scormil[, 1], scorfau[, 1])
s.match(scormil, scorfau)
```

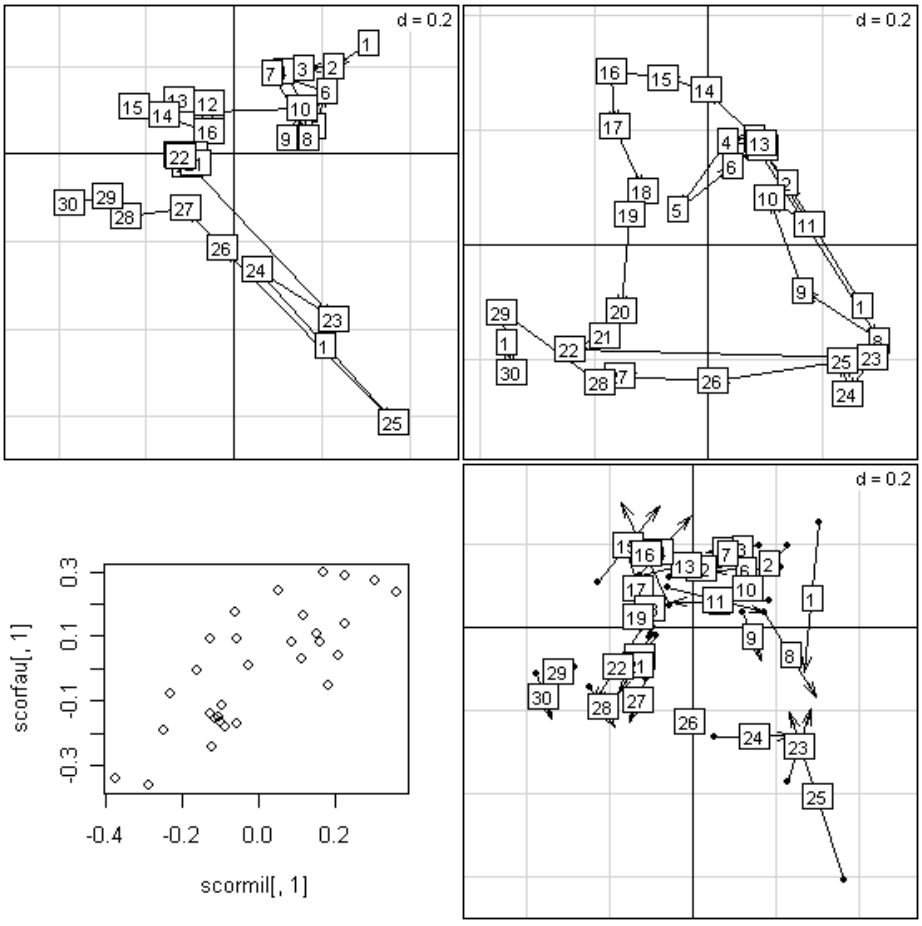


FIG. 24 – Le tableau sites-variables est soumis à une ACP normée. Le tableau sites-espèces est soumis à une ACP centrée. Les coordonnées de chacun des tableaux sont associés dans une analyse canonique.

```
pcaiv1 <- pcaiv(pcafau, doubs$mil, scannf = F)
plot(pcaiv1)
```

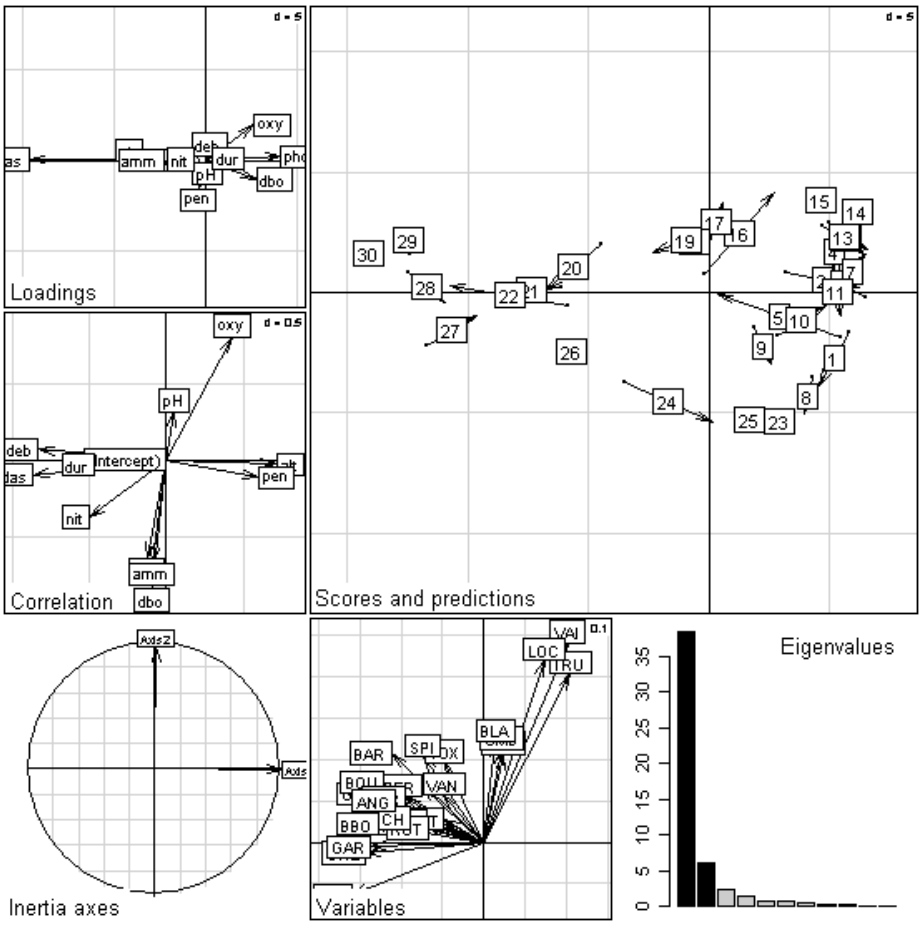


FIG. 25 – Les variables instrumentales sont les variables de milieu (explicatives). On fait l'ACP du tableau sites-espèces centré sous contrainte du tableau sites-variables.

4.4 La co-inertie

L'analyse de co-inertie [2] est la plus facile et la plus sûre des méthodes de couplage (figure 26). Les rotations procustéennes [23] sont de même logique (voir `macaca`). Pour qui demande une p-value, un test de permutations est disponible (figure 26).

5 Utiliser des matrices de distance

C'est particulièrement utile quand on utilise des marqueurs et que seule une typologie induite est en cause[17]. Cette stratégie est particulièrement pratiquée en données sensorielles, génétique et écologie des communautés. Voir les fonctions :

- ⋆ `dist.binary` (dissimilarités sur données binaires)
- ⋆ `dist.prop` (distances entre profils)
- ⋆ `dist.dudi` (distances euclidiennes dérivées des schémas de dualité)
- ⋆ `dist.neig` (distances dérivées des graphes de voisinages)
- ⋆ `dist.genet` (distances génétiques multi-loci)
- ⋆ `dist.quant` (distances sur variables quantitatives, morphométrie)

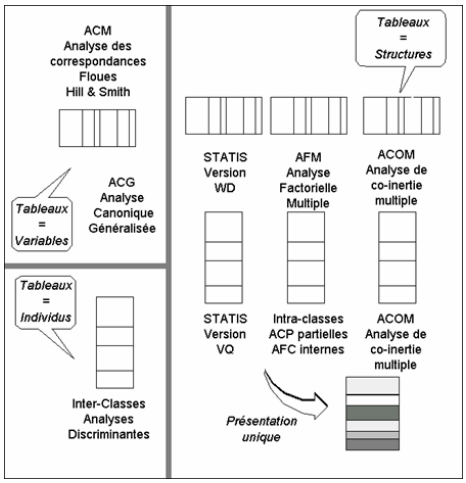
5.1 De distances à dendrogrammes : les CAH

5.2 De distances à tableaux : les coordonnées principales

Que veut-dire distance euclidienne ? Comment avoir des distances euclidiennes ? Quel en est l'intérêt ? [8]. On peut coupler par une co-inertie sur les représentations euclidiennes deux matrices de distances euclidiennes (figure 27).

5.3 ktab et kdist : vers de nouvelles classe d'objets

La multiplicité des tableaux a des origines diverse. Voir stage 6.



```
coi1 <- coinertia(pcafau, pcamil, scan = F)
plot(coi1)
```

```
w = randtest(coi1, 9999)
plot(w)
print(w)
```

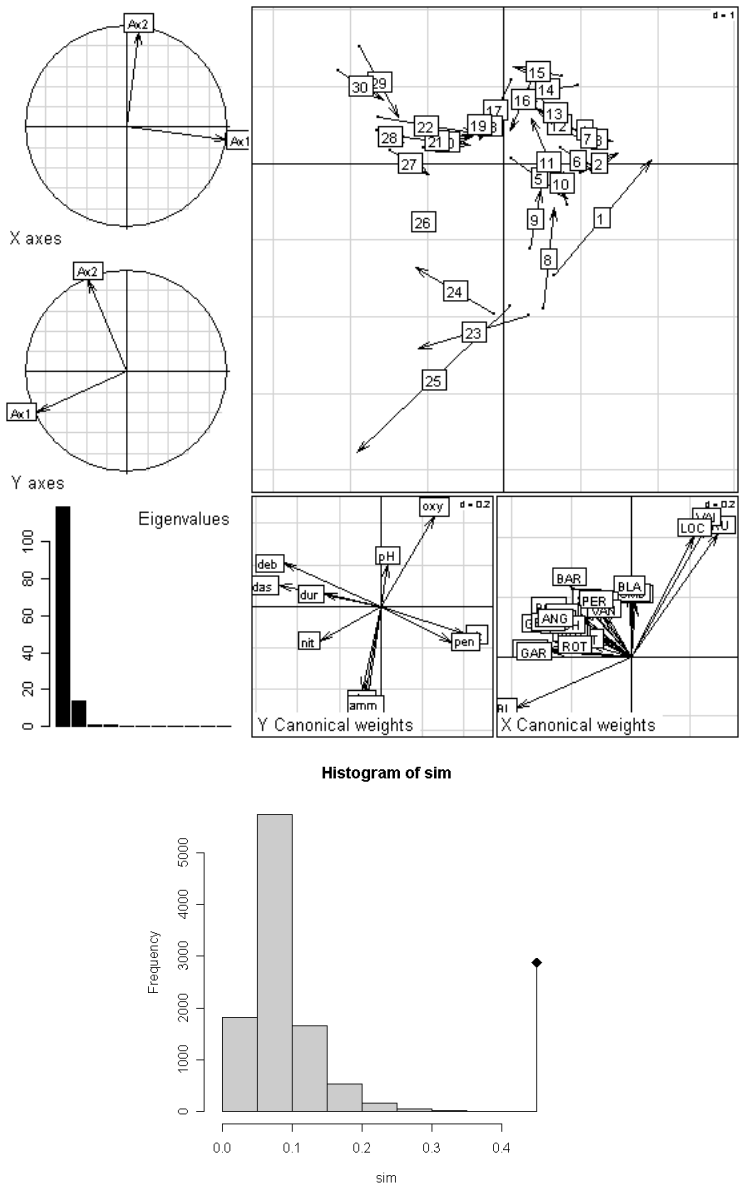


FIG. 26 – Analyse de co-inertie de deux nuages centrés issus de l'ACP centrée du tableau sites-espèces et du tableau centré-réduit sites-variables.

```
data(yanomama)
gen <- quasieuclid(as.dist(yanomama$gen))
geo <- quasieuclid(as.dist(yanomama$geo))
ant <- quasieuclid(as.dist(yanomama$ant))
geo1 <- dudi.pco(geo, scann = FALSE, nf = 3)
gen1 <- dudi.pco(gen, scann = FALSE, nf = 3)
ant1 <- dudi.pco(ant, scann = FALSE, nf = 3)
plot(coinertia(ant1, gen1, scann = FALSE))
```

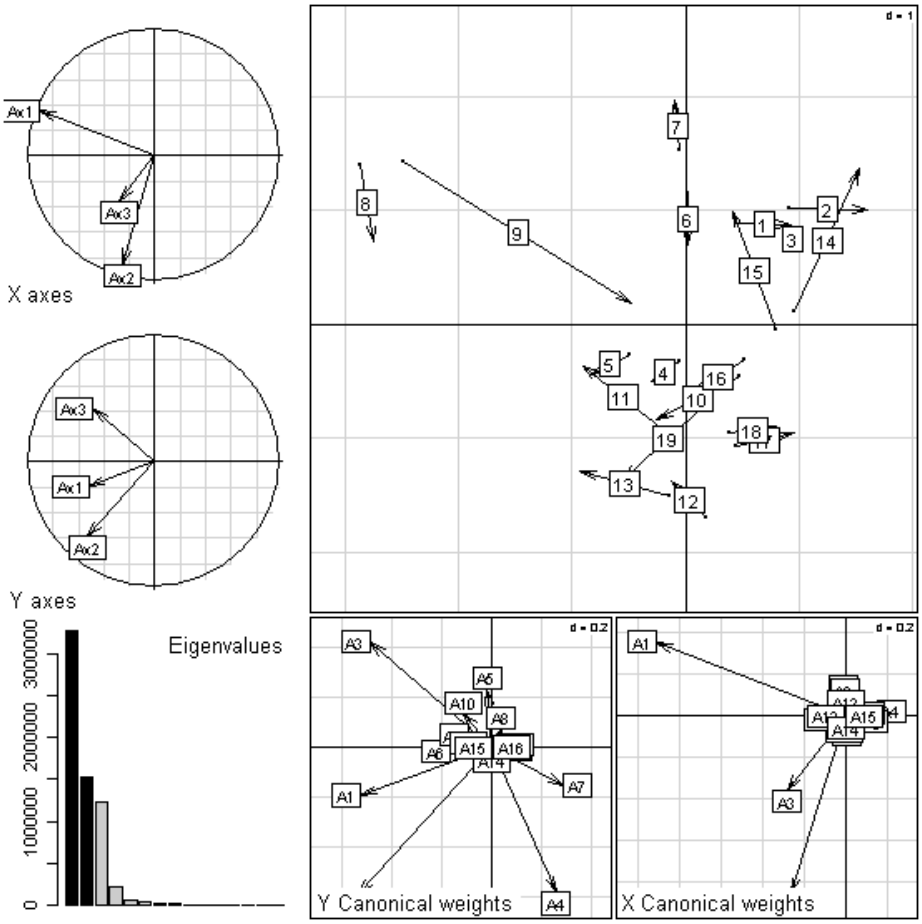


FIG. 27 – Des matrices de distances deviennent euclidiennes de plusieurs manières (voir `quasieuclid`, `lingoes` ou `cailliez`). Les matrices de distances euclidiennes deviennent des tableaux et peuvent se coupler.

```
data(ecomor)
d1 <- dist.binary(ecomor$habitat, 1)
d2 <- dist.prop(ecomor$forsub, 5)
d3 <- dist.prop(ecomor$diet, 5)
d4 <- dist.quant(ecomor$morpho, 3)
d5 <- dist.taxo(ecomor$taxo)
ecomor.kd <- kdist(d1, d2, d3, d4, d5)
names(ecomor.kd) = c("habitat", "forsub", "diet", "morpho",
    "taxo")
s.corcircle(dudi.pca(as.data.frame(ecomor.kd), scan = FALSE)$co)
```

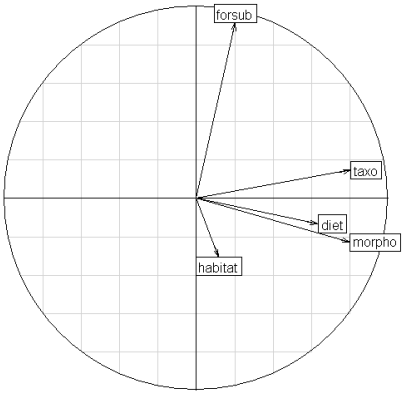


FIG. 28 – Une ACP normée sur des demi-matrices de distances vectorisées est la première approche simple des méthodes de compromis entre tableaux. Voir `statis`, `afm` et `mcoa`.

Distances spatiales, taxonomiques, phylogénétiques, morphométriques, génétiques : entre typologie et diversité. Les classes `ktab` et `kdist` portent de nouvelles problématiques (figure 28). Ceci était une invitation à découvrir `ade4`.

Références

[1] J. Blondel, D. Chessel, and B. Frochot. Bird species impoverishment, niche expansion, and density inflation in mediterranean island habitats. *Ecology*, 69 :1899–1917, 1988.

[2] D. Chessel and P. Mercier. Couplage de triplets statistiques et liaisons espèces-environnement. In J.D. Lebreton and B. Asselain, editors, *Biométrie et Environnement*, pages 15–44. Masson, Paris, 1993.

[3] W.S. Cleveland. *Visualizing data*. Hobart Press, Summit, New Jersey, 1993.

[4] W.S. Cleveland. *The elements of graphing data*. Hobart Press, Summit, New Jersey, 1994.

[5] P. Dagnelie. L'étude des communautés végétales par l'analyse statistique des liaisons entre les espèces et les variables écologiques : principes fonadamentaux, un exemple. *Biometrics*, 21 :345–361, 890–907, 1965.

[6] R. Gittins. *Canonical analysis, a review with applications in ecology*. Springer-Verlag, Berlin, 1985.

[7] J.C. Gower. Multivariate analysis and multivariate geometry. *The statistician*, 17 :13–28, 1967.

[8] J.C. Gower. Distance matrices and their euclidean approximation. In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone, editors, *Data Analysis and Informatics, III*, pages 3–21. Elsevier, North-Holland, 1984.

[9] H.O. Hirschfeld. A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences*, 31 :520–524, 1935.

[10] L. Hoffmann. Untersuchungen an enten in der camargue. *Ornithologischer Beobatcher*, 57 :35–50, 1960.

[11] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 :498–520, 1933.

[12] R.H. Jongman, C.J.F. ter Braak, and O.F.R. van Tongeren. *Data analysis in community and landscape ecology*. Pudoc, Wageningen, 1987.

[13] J.M. Lascaux. *Analyse de la variabilité morphologique de la truite commune (Salmo trutta L.) dans les cours d'eau du bassin pyrénéen méditerranéen*. PhD thesis, 1996.

[14] J.D. Lebreton. Etude des déplacements saisonniers des sarcelles d'hiver, anas c. crecca l., hivernant en camargue à l'aide de l'analyse factorielle des correspondances. *Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris, D*, III :2417–2420, 1973.

[15] J.D. Lebreton, R. Sabatier, G. Banco, and A.M. Bacou. Principal component and correspondence analyses with respect to instrumental variables : an overview of their role in studies of structure-activity and species- environment relationships. In J. Devillers and W. Karcher, editors, *Applied*

Multivariate Analysis in SAR and Environmental Studies, pages 85–114. Kluwer Academic Publishers, 1991.

[16] J.M. Legay and D. Pontier. Relation âge-fécondité dans les populations de chats domestiques, felis catus. *Mammalia*, 49 :395–402, 1985.

[17] P. Legendre and L. Legendre. *Numerical ecology*. Elsevier Science BV, Amsterdam, 2nd english edition edition, 1998.

[18] I. Noy-Meir. Data transformations in ecological ordination. i. some advantages of non-centering. *Journal of Ecology*, 61 :329–341, 1973.

[19] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 :559–572, 1901.

[20] R. Prodon and J.D. Lebreton. Breeding avifauna of a mediterranean succession : the holm oak and cork oak series in the eastern pyrénées. 1 : Analysis and modelling of the structure gradient. *Oïkos*, 37 :21–38, 1981.

[21] C.R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya, A*, 26 :329–359, 1964.

[22] M.J. Rochet, P.-A. Cornillon, R. Sabatier, and D. Pontier. Comparative analysis of phylogenic and fishing effects in life history patterns of teleos fishes. *Oïkos*, 91 :255–270, 2000.

[23] P.H. Schönemann. A generalized solution solution of the orthogonal procustes problem. *Psychometrika*, 31 :1–10, 1966.

[24] M. Tenenhaus and F.W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. *Psychometrika*, 50 :91–119, 1985.

[25] C.J.F. Ter Braak. Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 :1167–1179, 1986.

[26] J. Verneaux. *Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie.* PhD thesis, 1973.

[27] E.J. Williams. Use of scores for the analysis of association in contingency tables. *Biometrika*, 39 :274–289, 1952.