Fiche TD avec le logiciel : tdr621

Les analyses en composantes principales inter et intra classes

A.B. Dufour

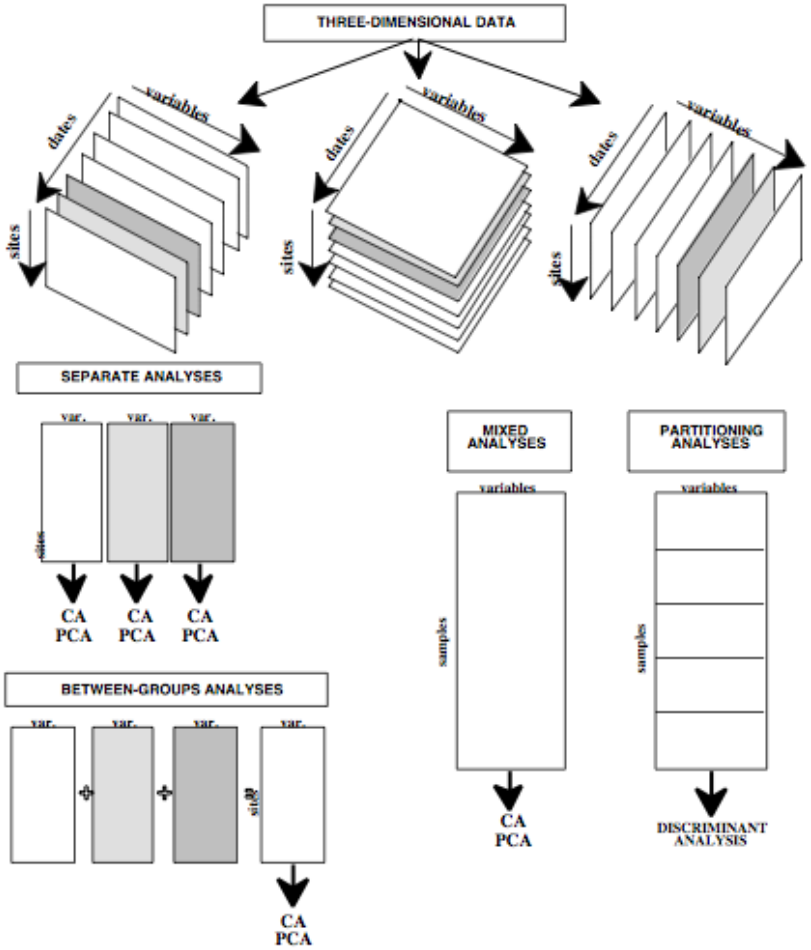
Analyses sur données environnementales recueillies dans 5 stations à 4 dates différentes (une par saison). Cette fiche s'appuie sur la fiche thématique 2.6. d'ADE-4 classique réalisée par S. Dolédec et D. Chessel.

Table des matières

1 Introduction

Les analyses de données doivent tenir compte des objectifs écologiques comme les conditions expérimentales (temps, espace...) afin de résoudre des problèmes tels que :

1. qu'est-ce qui, dans un ensemble de tableaux, dépend seulement du temps ? de l'espace ? et qu'est-ce qui peut être expliqué par une interaction entre l'espace et le temps ?
2. qu'est-ce qui dans un espace faunistique ne dépend pas des conditions d'échantillonnage (cf par exemple Usseglio-Polatera et Auda, 1987 [8]) ?



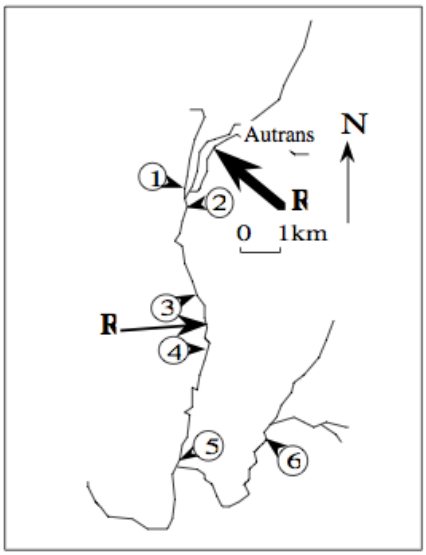
Quatre options d'ordination peuvent être trouvées dans la littérature (Dolédec et Chessel, 1991 [3]). Elles sont appelées respectivement : analyses séparées (1), analyses intergroupes (2), analyses mixtes (3), analyses partitionnées (4).

2 Approche classique

2.1 Les données

Le méaudret est une petite rivière du Vercors recevant les affluents de deux villages (Autrans et Méaudre). Cinq sites ont été choisis en amont et en aval du Méaudret. Des échantillons physico-chimiques sont prélevés sur chacun des cinq sites) à quatre occasions (Pegaz-Maucet, 1980 [6]). 9 variables sont mesurées :

1. `Temp` Température de l'eau (en degré celcius)
2. `Debit` Débit de l'eau (en litre par seconde)
3. `pH` pH
4. `Condu` Conductivité (en μ S/cm)
5. `Dbo5` Demande biologique en oxygène - 5 jours (en mg/l)
6. `Oxyd` Oxygène (en mg/L oxygène)
7. `Ammo` Ammoniaque (en mg/l NH_4^+)
8. `Nitra` Nitrate (en mg/l NO_3^-)
9. `Phos` Orthophosphate (en mg/l PO_4^{---})




```
library(ade4)
library(xtable)
data(meaudret)
names(meaudret)
```

```
[1] "mil"  "plan" "fau"
```

```
names(meaudret$mil)
```

```
[1] "Temp"  "Debit" "pH"    "Condu" "Dbo5"  "Oxyd"  "Ammo"  "Nitra" "Phos"
```

Nous avons 5 stations mesurées pendant les 4 saisons. Les variables qui définissent le site échantillonné et la date d'échantillonnage sont données dans `meaudret$plan`.

```
summary(meaudret$plan$sta)
```

```
S1 S2 S3 S4 S5
 4  4  4  4  4
```

```
summary(meaudret$plan$dat)
```

```
autumn spring summer winter
     5      5      5      5
```

2.2 L'analyse en composantes principales normée

Dans un premier temps, on réalise une analyse en composantes principales normée sur les 9 mesures physico-chimiques.

```
acp1 <- dudi.pca(meaudret$mil, scann = F, nf = 3)
acp1$eig
```

```
[1] 5.174736624 1.320418552 1.093376100 0.732113258 0.490213700 0.109834881
[7] 0.052960338 0.020030611 0.006315936
```

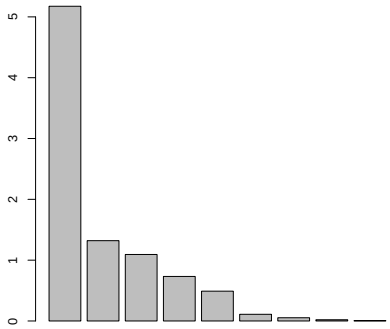
```
sum(acp1$eig)
```

```
[1] 9
```

```
cumsum(acp1$eig)/sum(acp1$eig)
```

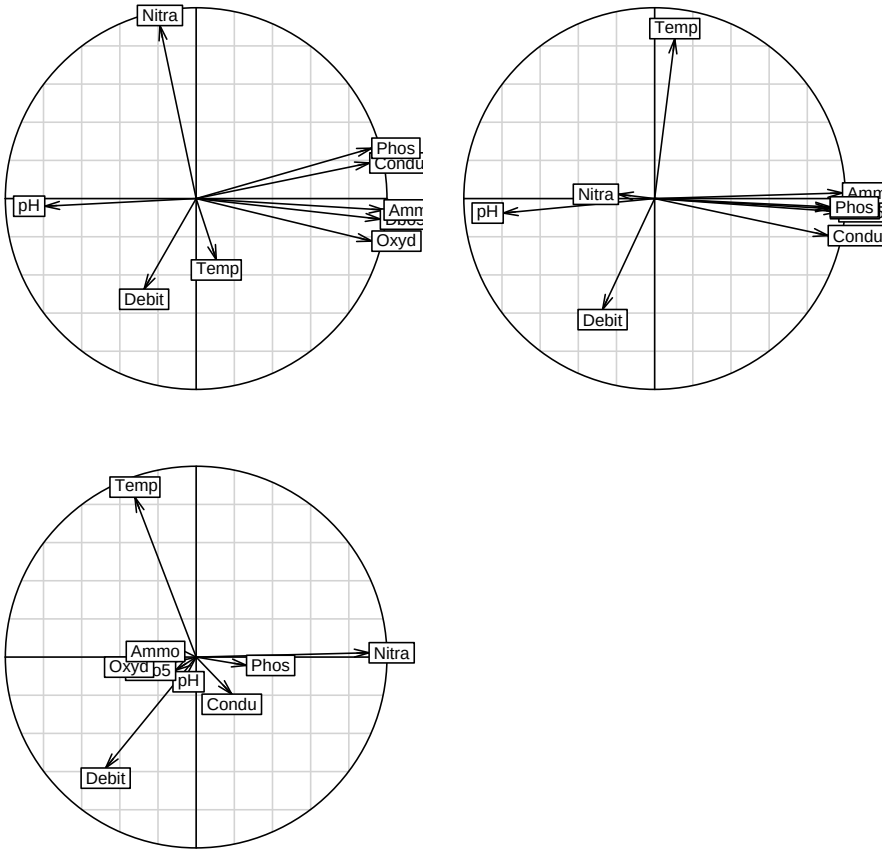
```
[1] 0.5749707 0.7216839 0.8431701 0.9245161 0.9789842 0.9911881 0.9970726 0.9992982
[9] 1.0000000
```

```
inertie <- cumsum(acp1$eig)/sum(acp1$eig)
barplot(acp1$eig)
```



On conserve les trois premières valeurs propres (celles qui sont ici supérieures à 1) et on représente la géométrie des 9 points (les variables) dans l'espace multidimensionnel $\mathbb{R}^{20}$. Les cercles des corrélations montrent une nette redondance entre les variables conductivité (`Condu`), demande biologique en oxygène (`Dbo5`), Oxygène (`Oxyd`), Ammoniaque(`Ammo`) et Orthophosphate (`Phos`) qui sont toutes des descripteurs de pollution organique.

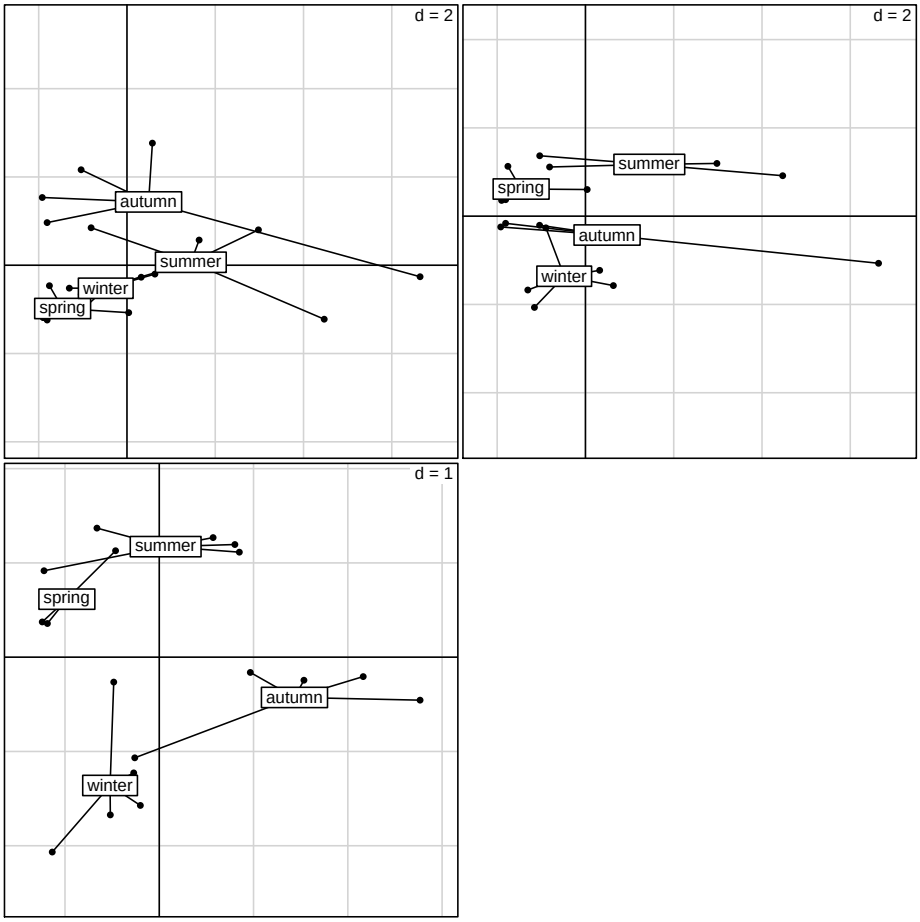
```
par(mfrow = c(2, 2))
s.corcircle(acp1$co, xax = 1, yax = 2)
s.corcircle(acp1$co, xax = 1, yax = 3)
s.corcircle(acp1$co, xax = 2, yax = 3)
```



Les cartes factorielles résument l'ACP normée. La fonction `s.class` permet de représenter les centres de gravité de chaque groupe et le lien entre un échantillon et son groupe d'appartenance.

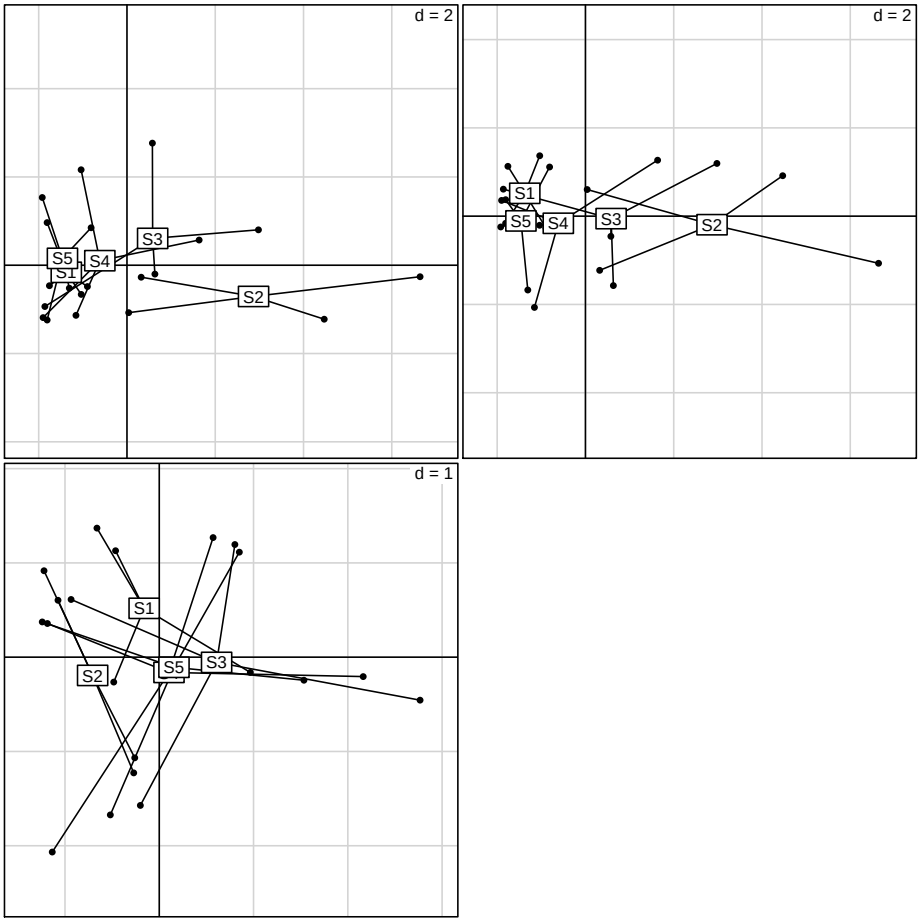
Pour les saisons

```
par(mfrow = c(2, 2))
s.class(acp1$li, meaudret$plan$dat, xax = 1, yax = 2, cellipse = 0)
s.class(acp1$li, meaudret$plan$dat, xax = 1, yax = 3, cellipse = 0)
s.class(acp1$li, meaudret$plan$dat, xax = 2, yax = 3, cellipse = 0)
```



Pour les stations

```
par(mfrow = c(2, 2))
s.class(acp1$li, meaudret$plan$sta, xax = 1, yax = 2, cellipse = 0)
s.class(acp1$li, meaudret$plan$sta, xax = 1, yax = 3, cellipse = 0)
s.class(acp1$li, meaudret$plan$sta, xax = 2, yax = 3, cellipse = 0)
```



Les trois premiers axes de l'ACP normée des données physico-chimiques sont utilisés pour décrire les corrélations entre les variables qui sont liées à la structure spatio-temporelle. Le premier axe (57.5%) prend en compte le pH, la conductivité (`Condu`), la demande biologique en oxygène (`Dbo5`), l'oxygène (`Oxyd`), l'ammoniaque (`Ammo`) et l'orthophosphate (`Phos`). Cela peut être interprété comme un gradient de minéralisation et indiquer également un taux élevé de pollution pour le site 2 durant l'automne.

```
rownames(meaudret$mil)[which.max(acp1$li[, 1])]
```

```
[1] "au_2"
```

Une telle pollution induit une acidité (faible pH), une concentration en oxygène faible, des valeurs élevées de demande biologique en oxygène et d'oxydabilité. Les fortes concentrations en ammoniaque et phosphate sont aussi caractéristiques d'une pollution organique forte. Une restauration de la rivière peut être observée sur les sites 3, 4 et 5. Le site 1 représente un site non pollué. L'évolution temporelle de la pollution est différente selon le cycle saisonnier défini par la température de l'eau (sur l'axe 3).

Par conséquent, cette analyse mélange à la fois une typologie selon les saisons et une typologie spatiale qui contrôlent le processus spatio-temporel produit par l'eau qui coule et l'évolution de la température de l'air. Ce processus peut se décomposer (au sens de la géométrie) c'est-à-dire que l'on peut choisir de se

focaliser sur un composant donné (espace ou temps) du plan d'échantillonnage ou alors choisir d'éliminer ce composant.

2.3 Retour aux données - Etude des lignes

Afin de tester l'effet spatial ou l'effet temporel, on peut réaliser une analyse de la variance à un facteur, les variables physico-chimiques prises une à une.

Pour l'effet spatial (station) et la variable température de l'eau (1), on obtient la table de décomposition de la variation :

```
options(show.signif.stars = F)
ressta <- anova(lm(meaudret$mil[, 1] ~ meaudret$plan$sta))
xtable(ressta, dig = 4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
meaudret$plan$sta	4.0000	6.7000	1.6750	0.0438	0.9960
Residuals	15.0000	573.5000	38.2333		

et le résultat des probabilités critiques sur l'ensemble des 9 variables physico-chimiques :

```
probasta <- rep(0, 9)
for (i in 1:9) {
    ressta <- anova(lm(meaudret$mil[, i] ~ meaudret$plan$sta))
    probasta[i] <- ressta[[1, 5]]
}
xtable(rbind(colnames(meaudret$mil), round(probasta, dig = 4)))
```

	1	2	3	4	5	6	7	8	9
1	Temp	Debit	pH	Condu	Dbo5	Oxyd	Ammo	Nitra	Phos
2	0.996	0.2366	0.3464	0.1464	0.0161	0.0022	0.0221	0.1012	0.0528

Pour l'effet temporel (saison) et la variable température de l'eau (1), on obtient la table de décomposition de la variation :

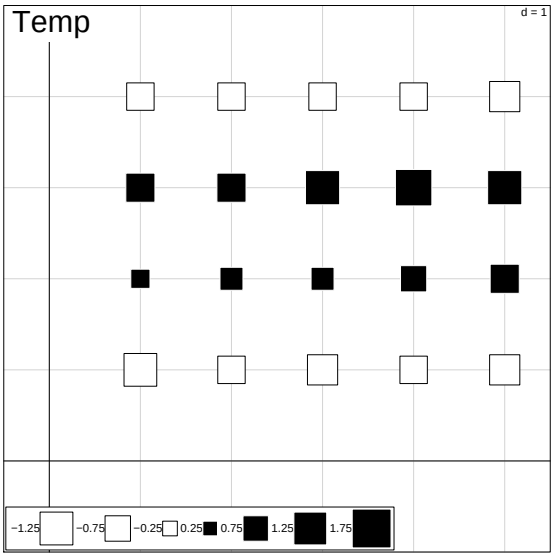
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
meaudret$plan$dat	3.0000	564.2000	188.0667	188.0667	0.0000
Residuals	16.0000	16.0000	1.0000		

et le résultat des probabilités critiques sur l'ensemble des 9 variables physico-chimiques :

	1	2	3	4	5	6	7	8	9
1	Temp	Debit	pH	Condu	Dbo5	Oxyd	Ammo	Nitra	Phos
2	0	0.0032	0.0361	0.0179	0.5991	0.7795	0.3621	0.0795	0.1708

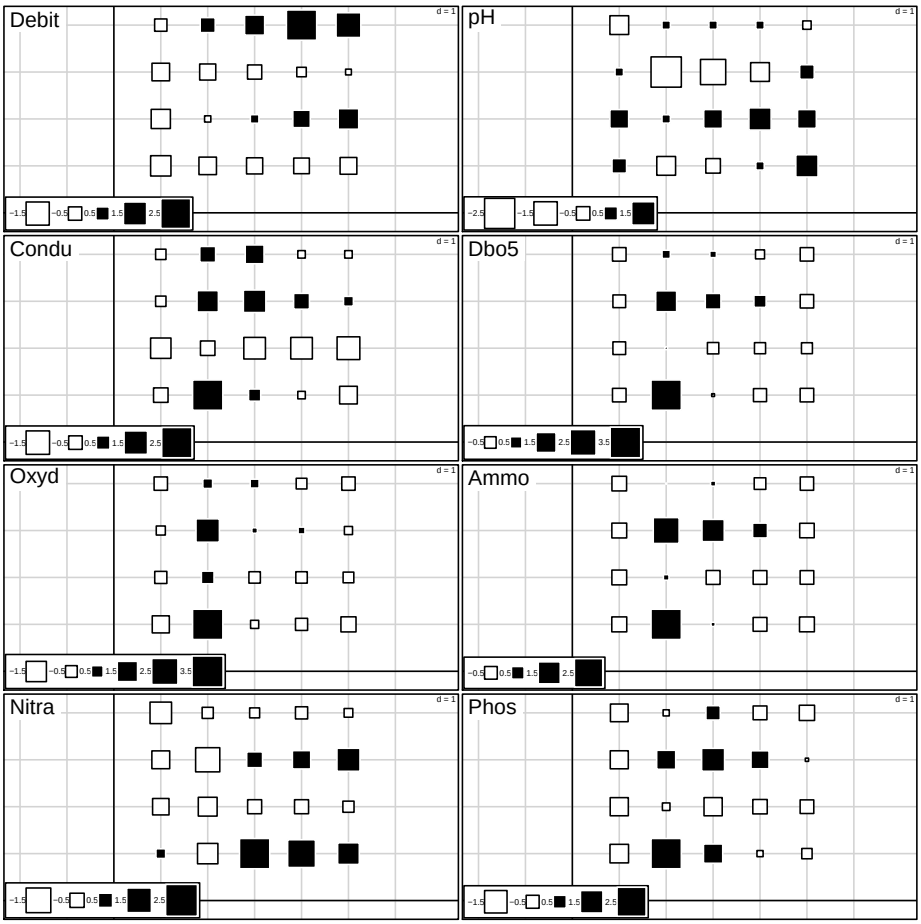
Tous ces résultats peuvent être visualisés à l'aide de la fonction `s.value` comme le montre la représentation ci-dessous de la température avec en colonnes, les 5 stations et en lignes les saisons (de bas en haut : du printemps à l'hiver).

```
s.value(meaudret$plan, acp1$tab[, 1], xax = 2, yax = 1, sub = colnames(meaudret$mil)[1],
    possub = "topleft", csub = 2)
```



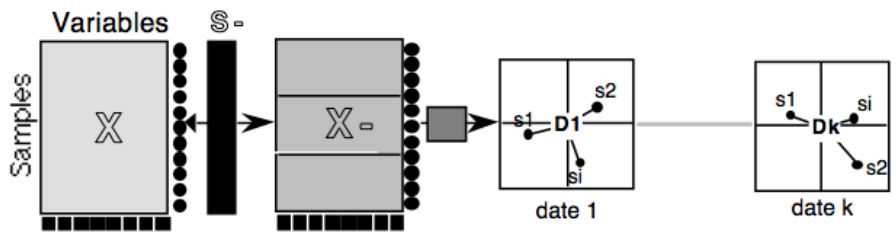
et les huit autres variables physico-chimiques :

```
par(mfrow = c(4, 2))
for (i in 2:9) s.value(meaudret$plan, acp1$tab[, i], xax = 2, yax = 1,
    sub = colnames(meaudret$mil)[i], possub = "topleft", csub = 2)
```



3 Enlever un effet : l'ACP intragroupe

En analyse en composantes principales intragroupe, tous les centres des classes sont placés à l'origine des cartes factorielles et les individus sont représentés avec une variance maximale autour de l'origine. Ainsi, l'objectif principal de l'analyse est de permettre l'étude simultanée des typologies spatiales ou de faire une collection de typologies spatiales.



Pour réaliser une ACP intragroupe sous `ade4`, le plus simple consiste à utiliser la fonction `within` qui permet d'étudier le lien entre une table et une variable

qualitative identifiant les groupes. On cherche par exemple à éliminer l'effet saison `meaudret$plan$dat`.

```
wit1 <- within(acp1, meaudret$plan$dat, scan = FALSE)
barplot(wit1$eig)
eig1 <- wit1$eig[1]/9
eig1
```

```
[1] 0.4619491
```

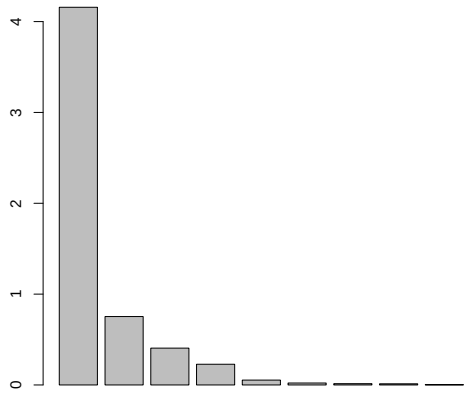
```
inertia.dudi(wit1)
```

```
$TOT
      inertia      cum     ratio
1 4.157541744 4.157542 0.7359025
2 0.753075168 4.910617 0.8692000
3 0.405374843 5.315992 0.9409530
4 0.228024043 5.544016 0.9813143
5 0.053611513 5.597627 0.9908037
6 0.021426655 5.619054 0.9945963
7 0.014152646 5.633207 0.9971014
8 0.012806219 5.646013 0.9993682
9 0.003569641 5.649582 1.0000000
```

```
wit1$ratio
```

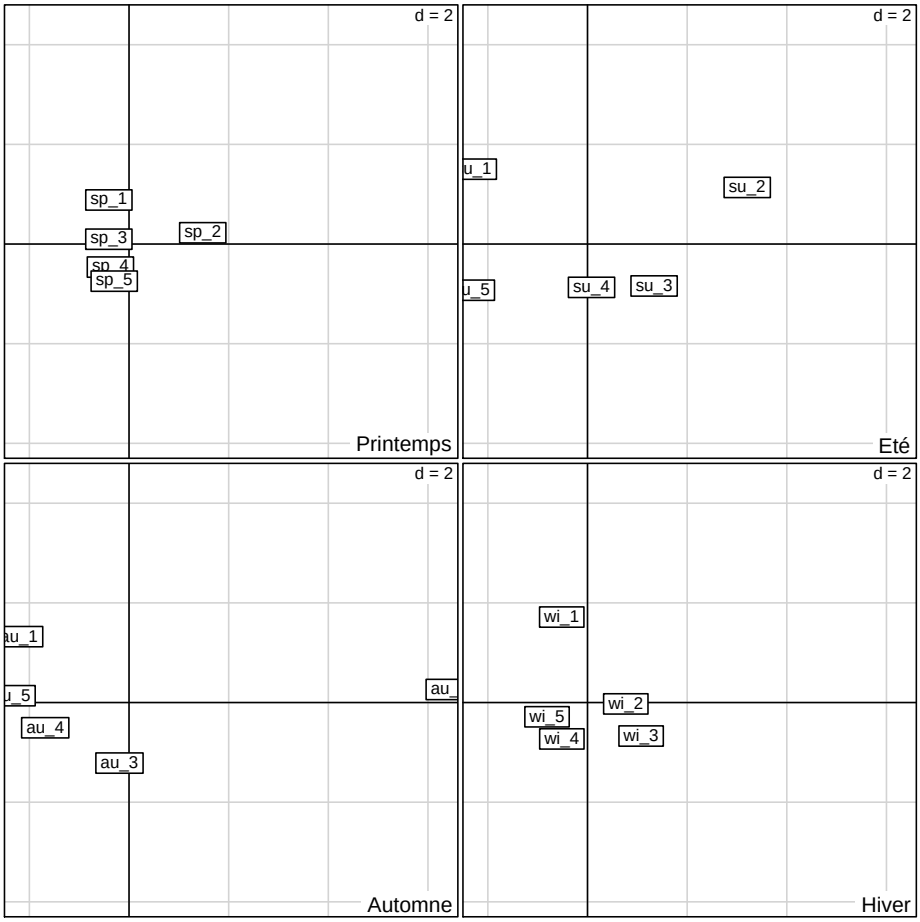
```
[1] 0.6277314
```

```
inerwit <- wit1$ratio
```



En terme d'inertie, l'ACP globale des données du milieu est égale à 9 (nombre total de variables dans une ACP normée). L'inertie intraclasse est égale 0.6277 c'est-à-dire que 62.77 % de l'inertie totale est attribué à l'ACP intragroupe. De plus, 46.19% de l'inertie intraclasse est donné par le premier axe. Réaliser une ACP intraclasse est presque la même chose que réaliser simultanément les ACP des 4 tableaux "sites × variables" définis par les 4 saisons. Il serait donc possible de rechercher une représentation graphique liant quatre cartes factorielles différentes à l'ACP intragroupe (cf analyses séparées `sepan`).

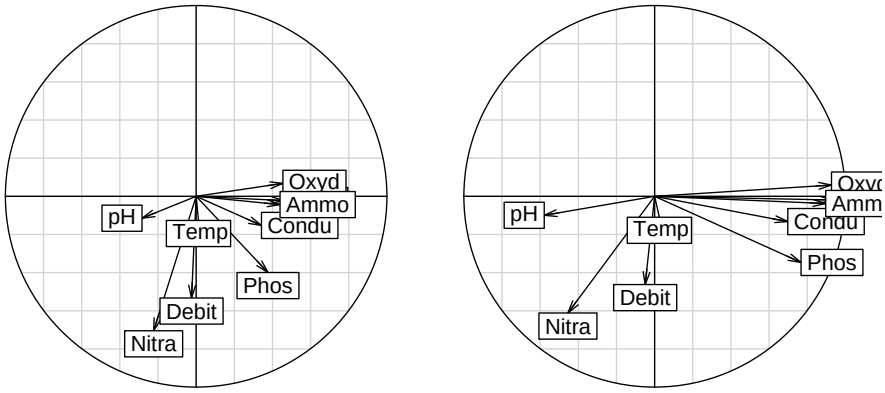
```
par(mfrow = c(2, 2))
s.label(wit1$li[1:5, ], label = rownames(meaudret$mil)[1:5], xlim = c(-2.5,
    6.6), ylim = c(-1.3, 1.8), sub = "Printemps", possub = "bottomright")
s.label(wit1$li[6:10, ], label = rownames(meaudret$mil)[6:10], xlim = c(-2.5,
    6.6), ylim = c(-1.3, 1.8), sub = "Eté", possub = "bottomright")
s.label(wit1$li[11:15, ], label = rownames(meaudret$mil)[11:15],
    xlim = c(-2.5, 6.6), ylim = c(-1.3, 1.8), sub = "Automne", possub = "bottomright")
s.label(wit1$li[16:20, ], label = rownames(meaudret$mil)[16:20],
    xlim = c(-2.5, 6.6), ylim = c(-1.3, 1.8), sub = "Hiver", possub = "bottomright")
```



La typologie spatiale n'est pas similaire d'une saison à l'autre. Les axes de l'inertie intragroupe représentent les axes produits par la superposition des 4 groupes (saisons) des 5 sites centrés par saison.

La représentation des variables, donnée par le cercle des corrélation speut être réalisée à partir de `c1` (les vecteurs sont normés à 1), à partir de `co` (les vecteurs sont normés à la valeur propre). Dans ce dernier cas, ils ne représentent pas les corrélations entre les axes et les variables mais les covariances.

```
par(mfrow = c(1, 2))
s.corcircle(wit1$c1)
s.corcircle(wit1$co)
```



Les deux fichiers (`c1` et `co`) représentent deux points de vue concernant les analyses utilisant les projections.

- Soit $\mathbf{X}$ une matrice de dimension $n \times p$,
- soient deux matrices diagonales $\mathbf{D}$ et $\mathbf{Q}$ associées respectivement aux lignes et aux colonnes de $\mathbf{X}$,
- soit le sous-espace $\mathbf{A}$ défini par la variable qualitative (station ou saison).

L'analyse d'inertie classique du triplet $(\mathbf{P_A}(\mathbf{X}), \mathbf{Q}, \mathbf{D})$ avec $\mathbf{P_A}(\mathbf{X})$ projection du tableau $\mathbf{X}$ sur le sous espace $\mathbf{A}$ conduit aux scores des colonnes (notés `co`) et aux scores des lignes (notés `li`).

Un autre point de vue est de considérer que l'ACP intragroupe a pour objectif de rechercher une combinaison linéaire des variables (notée `li`) à l'aide des coefficients des variables (notés `c1`) telle que l'inertie projetée soit maximale. Ceci introduit l'analyse en composantes principales sur variables instrumentales.

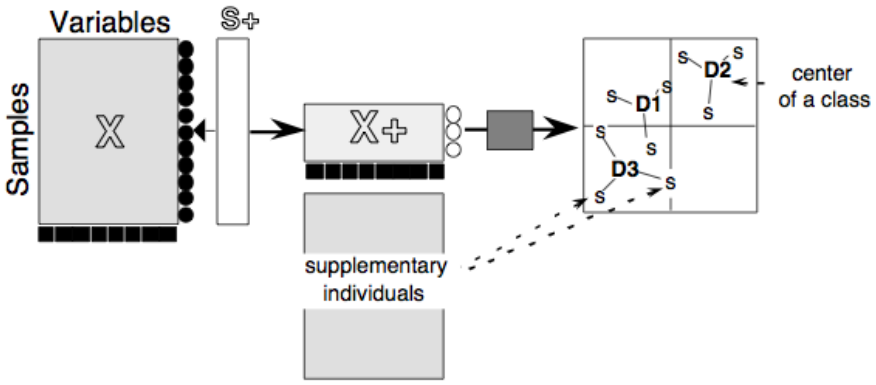
Finalement, l'interprétation de l'ACP intra-saison est la suivante. Durant la période de printemps (faible pollution), les sites 1, 3, 4 et 5 s'opposent au site 2 (pollué). En été, le site 1 se sépare des sites 3, 4 et 5. En automne, la pollution augmente et le site 2 s'éloigne encore plus des autres sites sur l'axe horizontal. En hiver, les sites 2 et 3 sont toujours sous l'influence des effluents du village d'Autrans.

Exercice. Réaliser l'ACP intra-sites recherchant les éléments communs aux tables 'dates × variables'.

4 Mettre l'accent sur un effet : l'ACP intergroupe

Une analyse en composantes principales intergroupes peut être liée à une analyse en composantes principales intragroupes. La seconde recherche les axes partagés par les sous-espaces. La première recherche les axes au centre de gravité

de l'espace et met l'accent sur la différence entre les groupes, dans ce cas, les variations temporelles.



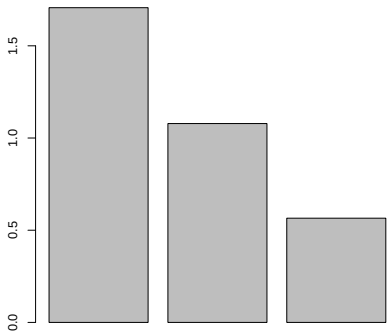
La signification statistique de la dispersion des centres de gravité peut être testée en utilisant la fonction `between`.

```
bet1 <- between(acp1, meaudret$plan$dat, scan = FALSE, nf = 2)
barplot(bet1$eig)
bet1$ratio
```

```
[1] 0.3722686
```

```
inerbet <- bet1$ratio
inertia.dudi(bet1)
```

```
$TOT
     inertia      cum     ratio
1 1.7067496 1.706750 0.5094140
2 1.0784279 2.785177 0.8312926
3 0.5652401 3.350418 1.0000000
```



Dans cette analyse, l'inertie intergroupes est égale à 0.3723 c'est-à-dire que 37.23 % de l'inertie totale est attribuée à l'ACP intergroupes. Comme résultat complémentaire des analyses intergroupes et intragroupes, l'inertie totale du tableau initial peut être décomposé en deux parties. Chaque partie est ainsi décomposée en axes.

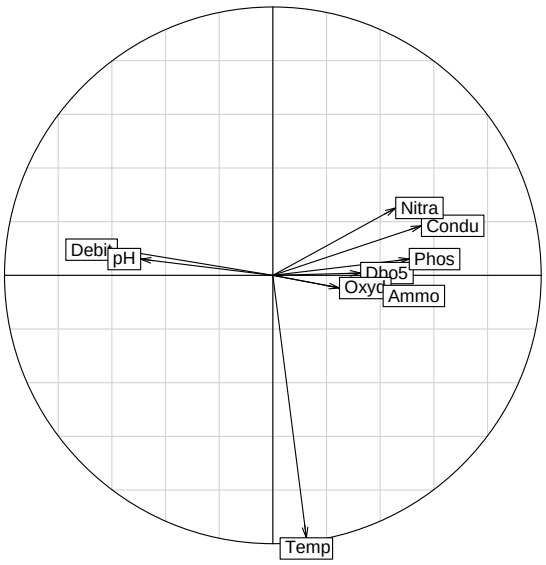
Par conséquent, en considérant l'inertie totale de $\mathbf{X}$ notée I_T, l'inertie de $\mathbf{X}^-$ (modèle intragroupes) notée I_T^-, l'inertie de $\mathbf{X}^+$ (modèle intergroupe) notée I_T^+, on a la relation suivante :

$$I_T = I_T^+ + I_T^-$$

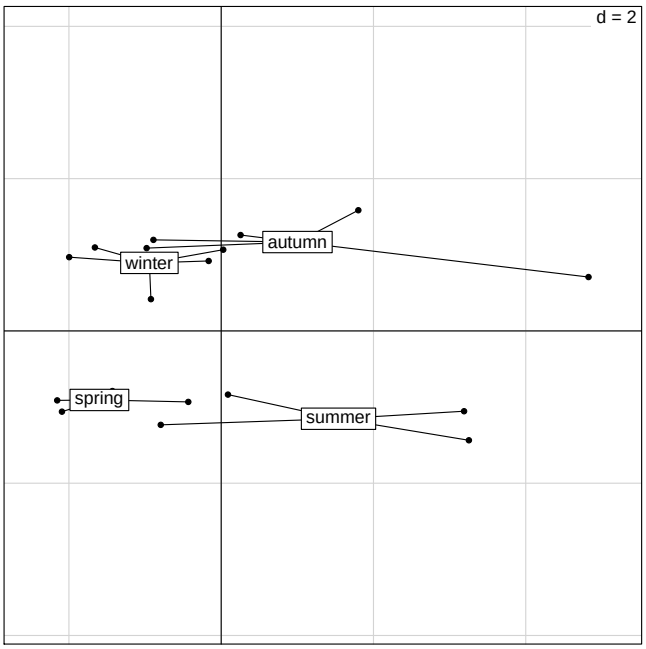
On peut appliquer cette équation à notre exemple :

```
wit1$ratio
[1] 0.6277314
bet1$ratio
[1] 0.3722686
wit1$ratio + bet1$ratio
[1] 1
```

```
s.corcircle(bet1$co)
```




```
s.class(bet1$ls, meaudret$plan$dat, cellipse = 0)
```



En moyenne, sur l'axe 1, on observe une pollution plus importante en automne et en été. Pendant l'hiver et surtout le printemps, la valeur élevée du débit de l'eau conduit à une dilution de la pollution organique dans la rivière. L'axe 2 décrit l'influence du rythme saisonnier avec la température de l'eau. Par conséquent, l'été s'oppose aux trois autres saisons.

Exercice. Réaliser l'ACP inter-sites.

5 Décomposition de la variance

5.1 Utilisation des valeurs propres

Chaque analyse décompose la variabilité totale en variabilité spatiale et en variabilité temporelle. La part la plus importante de la variabilité est prise en compte par la première valeur propre de chaque analyse. Une part de la variabilité totale est perdue lorsqu'on enlève l'effet temporel (ACP intra-dates). Mais une part plus importante de la variabilité est perdue en enlevant l'effet spatial (ACP intra-sites). Une telle dominance de l'effet spatial est également visible avec la première valeur propre de l'ACP inter-sites qui est bien plus grande que la première valeur propre de l'ACP inter-dates.

5.2 Projections sur les sous espaces

La décomposition de la variance est associée au théorème de Pythagore. Soit $\mathbf{z}$ une variable normée, vecteur unitaire de $\mathbb{R}^n$. D'un point de vue géométrique, la longueur (ou le carré de la longueur) de $\mathbf{z}$ est égale à 1. D'un point de vue statistique, la variance de $\mathbf{z}$ est égale à 1. La norme (carré de la longueur du vecteur) du vecteur projeté sur un sous-espace est égale à la variance de ces composantes (si nous sommes dans un sous-espace orthogonal engendré par

$\mathbf{1}_n$ (centrage)). Le rapport de cette deuxième variance sur la première est le pourcentage de variance expliqué par la projection. Par conséquent, la procédure intègre deux étapes dans la projection : (1) projection sur un sous espace, (2) projection sur les axes factoriels de l'ACP (réduction des dimensions de l'espace initial).

5.3 Centrage alternatif

D'une manière générale, une expérience intègre des facteurs qui interfèrent entre eux. Par exemples, les répétitions temporelles sont enregistrées mais la chronologie est sans intérêt pour l'expérimentateur ; un grand nombre de sites sont présents mais le rôle de la distribution spatiale est sans intérêt. Le facteur d'interférence alors présent (temps, site) peut être enlevé en moyenne. C'est le cas de l'ACP intraclasse classique (analysé ici). De plus, l'interférence donnée peut être enlevée de l'effet moyen et de la variance. C'est le cas lorsqu'une ACP intragroupes est réalisée sur un tableau dont les variables sont normées par groupe d'individus (Dolédec et Chessel, 1987 [1]). Dans ce cas, les valeurs moyennes pour une variable et pour chaque groupe sont nulles. Comme l'inertie totale se décompose en inertie interclasse et inertie intraclasse, l'inertie interclasse est nulle. L'inertie totale est donc égale à l'inertie intraclasse c'est-à-dire la moyenne des variances des groupes, soit 1. Ainsi, les valeurs contenues dans le tableau analysé sont égales à :

$$x_{ijk} = \frac{z_{ijk} - z_{i.k}}{s_{i.k}}$$

où z_{ijk} est la valeur en ligne, $z_{i.k}$ est la moyenne et $s_{i.k}$ l'écart-type de la $k$ième variable pour le site i.

Ce dernier exemple montre la souplesse de la librairie `ade4` et surtout le besoin de définir clairement les objectifs lorsqu'on utilise les projections sur des sous-espaces parce qu'un grand nombre d'options sont disponibles. De plus, les analyses intergroupes et intragroupe sont également disponibles dans le cas de l'analyse des correspondances (cf fiche `tdr623`, Dolédec et Chessel 1989 [2]). D'autres méthodes utilisent ce type de démarche comme l'analyse triadique partielle [7] ou STATIS ([5], [4]).

Références

[1] S. Dolédec and D. Chessel. Rythmes saisonniers et composantes stationnelles en milieu aquatique i- description d'un plan d'observations complet par projection de variables. *Acta Œcologica, Œcologia Generalis*, 8 :403–426, 1987.

[2] S. Dolédec and D. Chessel. Rythmes saisonniers et composantes stationnelles en milieu aquatique ii- prise en compte et élimination d'effets dans un tableau faunistique. *Acta Œcologica, Œcologia Generalis*, 10 :207–232, 1989.

[3] S. Dolédec and D. Chessel. Recent developments in linear ordination methods for environmental sciences. *Advances in Ecology, India*, 1 :133–155, 1991.

[4] Y. Escoufier. L'analyse des tableaux de contingence simples et multiples. *Metron*, 40 :53–77, 1982.

[5] H. L'Hermier des Plantes. *Structuration des tableaux à trois indices de la statistique. Théorie et applications d'une méthode d'analyse conjointe.* PhD thesis, 1976.

[6] D. Pegaz-Maucet. *Impact d'une perturbation d'origine organique sur la dérive des macro-invertébérés benthiques d'un cours d'eau. Comparaison avec le benthos.* PhD thesis, 1980.

[7] 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.

[8] P. Usseglio-Polatera and Y. Auda. Influence des facteurs météorologiques sur les résultats de piégeage lumineux. *Annales de Limnologie*, 23 :65–79, 1987.