Fiche TD avec le logiciel R : ter6

# Un tableau et une carte

S. Dray

La fiche présente les principes d'une méthode permettant la prise en compte de l'espace en analyse multivariée. La mise en oeuvre de la fonction `multispati` de la librairie `ade4` est détaillée.

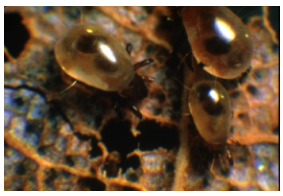
## Table des matières

# 1 Introduction

Il existe un grand nombre de méthodes pour analyser une structure spatiale univariée (voir `http://pbil.univ-lyon1.fr/R/fichestd/ter4.pdf`). De nombreuses tentatives d'extensions au cas multivarié ont été publiées dans la littérature. On peut distinguer deux grands types d'approches. Certaines méthodes visent à représenter l'espace sous la forme d'un tableau contenant des polynomes des coordonnées [Borcard and Legendre, 1994, Borcard et al., 1992] ou des vecteurs propres de voisinage [Méot et al., 1998, Borcard and Legendre, 2002, Borcard et al., 2004, Dray et al., 2006]. La mise en évidence de structures spatiales est obtenue en liant le tableau faunistique au tableau représentant l'espace par une méthode de couplage comme l'analyse canonique des correspondances ou l'analyse des redondances. La prise en compte de variables environmentales par des méthodes dites partielles [ter Braak, 1988] conduit souvent au partitionnement de la *variation* [Borcard et al., 1992, Peres-Neto et al., 2006]. Le deuxième type d'approche consiste à introduire la contrainte spatiale directement dans l'analyse d'un tableau. On étend alors les mesures d'autocorrélation univariée au cas multivarié. Dans ce document, on s'intéresse uniquement à ce deuxième type de méthodes.

# 2 Une carte et $p$ variables

On considère les données mise à disposition à `http://www.fas.umontreal.ca/biol/casgrain/fr/labo/oribates.html`. On a une description complète dans [Borcard and Legendre, 1994] et [Borcard et al., 1992].



```
library(ade4)
data(oribatid)
names(oribatid)
```

```
[1] "fau"   "envir" "xy"
```

```
dim(oribatid$fau)
```

```
[1] 70 35
```

On a un tableau faunistique avec 35 espèces (colonnes) x 70 éléments d'échantillonnage (lignes). Les éléments d'échantillonnage sont des carottes de sol de 5 cm de diamètre et 10 cm de profondeur. On a les coordonnées dans l'espace (en xy) des carottes.

On a un tableau de 70 carottes (lignes) et 5 variables environnementales (colonnes) :

```
names(oribatid$envir)
```

```
[1] "substrate" "shrubs"    "topo"      "density"   "water"
```

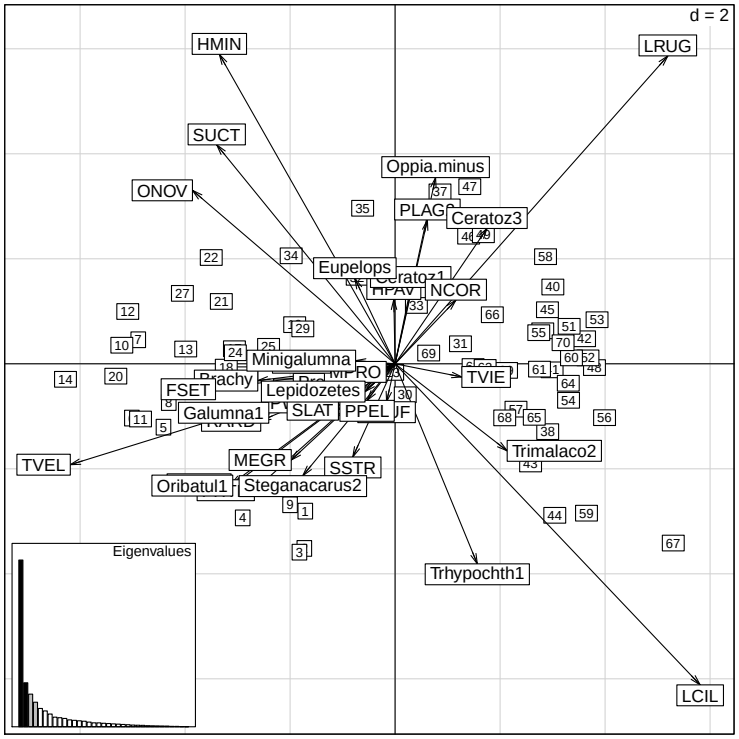
```
summary(oribatid$envir)
```

```
 substrate   shrubs        topo       density          water
inter :27   few :26   blanket:44   Min.   :21.17   Min.   :134.1
litter: 2   many:25   hummock:26   1st Qu.:30.01   1st Qu.:314.1
peat  : 2   none:19                Median :36.38   Median :398.5
sph1  :25                          Mean   :39.28   Mean   :410.6
sph2  :11                          3rd Qu.:46.81   3rd Qu.:492.8
sph3  : 1                          Max.   :80.59   Max.   :827.0
sph4  : 2
```

## 2.1 Analyse multivariée

La synthèse d'un ensemble de variables est assurée par une analyse multivariée. $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ est un schéma de dualité ou analyse de premier niveau (en cas de besoin, voir la fiche `http://pbil.univ-lyon1.fr/R/stage/stage3.pdf` ou Chessel et al. [2004]). $\mathbf{X}$ est un tableau, $\mathbf{Q}$ une pondération de ses colonnes et $\mathbf{D}$ une pondération de ses lignes.
On obtient un résumé des données faunistiques par une ACP :

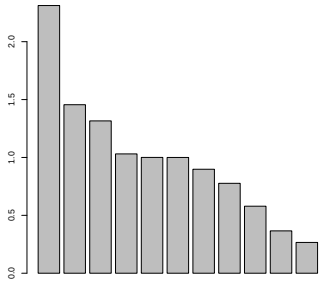
```
ori.log <- log(oribatid$fau + 1)
ori.pca <- dudi.pca(ori.log, scale = F, scannf = F, nf = 4)
scatter(ori.pca, posieig = "bottomleft")
```




$\mathbf{X}$ peut être un `data.frame` quelconque contenant des variables quantitatives (`numeric`) et des variables qualitatives (`factor`) voire même des qualitatives à modalités ordonnées (`ordered`).
Les quantitatives sont centrées et réduites, les qualitatives décomposées en indicatrices de classes puis centrées correctement et les pondérations font en sorte que chaque variable ait le même poids que les autres. La fonction `dudi.mix` assure cette opération :

```
ori.mix <- dudi.mix(oribatid$envir, scannf = F, nf = 3)
ori.mix
Duality diagramm
class: mix dudi
$call: dudi.mix(df = oribatid$envir, scannf = F, nf = 3)
$nf: 3 axis-components saved
$rank: 11
eigen values: 2.312 1.456 1.316 1.031 1 ...
  vector length mode    content
1 $cw    14     numeric column weights
2 $lw    70     numeric row weights
3 $eig   11     numeric eigen values

  data.frame nrow ncol content
1 $tab       70   14   modified array
2 $li        70   3    row coordinates
3 $l1        70   3    row normed scores
4 $co        14   3    column coordinates
5 $c1        14   3    column normed scores
other elements: assign index cr
barplot(ori.mix$eig)
```




Le poids d'une colonne vaut '1' si la variable associée à la colonne est quantitative (i.e. la variable `density`). Le poids d'une colonne vaut 'la fréquence de la modalité' si la colonne correspond à une modalité d'une variable qualitative. Ainsi la somme des poids des modalités d'une variable qualitative vaut 1 (i.e. la variable `substrate`).

```
ori.mix$cw
 subst.inter subst.litter   subst.peat   subst.sph1   subst.sph2   subst.sph3
  0.38571429   0.02857143   0.02857143   0.35714286   0.15714286   0.01428571
  subst.sph4    shrub.few   shrub.many   shrub.none topo.blanket topo.hummock
  0.02857143   0.37142857   0.35714286   0.27142857   0.62857143   0.37142857
     density        water
  1.00000000   1.00000000
ori.mix$cw[13]
density
      1
sum(ori.mix$cw[1:7])
[1] 1
sum(ori.mix$cw)
[1] 5
```

La pondération des lignes est uniforme :

```
unique(ori.mix$lw)
[1] 0.01428571
1/nrow(ori.mix$tab)
```

```
[1] 0.01428571
```

Les axes principaux ordinaires sont des vecteurs en colonnes dans une matrice $\mathbf{U}_s$ ($s$ est le nombre de facteurs conservés dans l'analyse simple) $\mathbf{Q}$-orthonormés :

$$\mathbf{U}_s{}^T\mathbf{Q}\mathbf{U}_s = \mathbf{I}_s$$

Les coordonnées de l'analyse simple $\mathbf{XQU}_s$ maximisent successivement l'inertie projetée sur un axe $\mathbf{u}$ soit $\|\mathbf{XQu}\|_{\mathbf{D}}{}^2$. Les maxima successifs sont les valeurs propres de l'analyse simple qu'on notera $\lambda_1, ..., \lambda_s$. Dans cet exemple, cela signifie que la première coordonnée est un score numérique $\mathbf{z}$ qui maximise la somme des quantités $corr^2(\mathbf{z}, \mathbf{X}[, j])$ quand la variable $\mathbf{X}[, j]$ est quantitative et $\eta^2(\mathbf{z}, \mathbf{X}[, j])$ quand elle est qualitative. Cette analyse est une ACP normée sur matrice de corrélation quand il n'y a que des quantitatives et une Analyse des Correspondances Multiples quand il n'y a que des qualitatives.

```
ori.mix$cr
```

```
                RS1         RS2         RS3
substrate 0.3654604 0.555784391 0.76607072
shrubs    0.4873543 0.354481827 0.25668105
topo      0.5242076 0.031934274 0.20871236
density   0.2370110 0.509573865 0.01215227
water     0.6982091 0.003817108 0.07214140
```

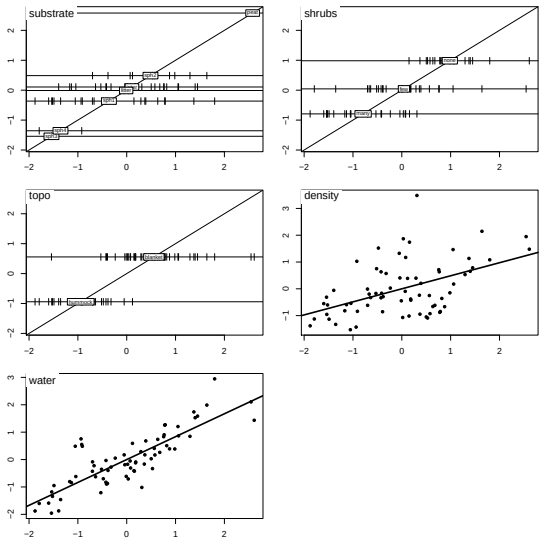
```
sum(ori.mix$cr[, 1])
```

```
[1] 2.312242
```

```
ori.mix$eig[1]
```

```
[1] 2.312242
```

```
score(ori.mix)
```

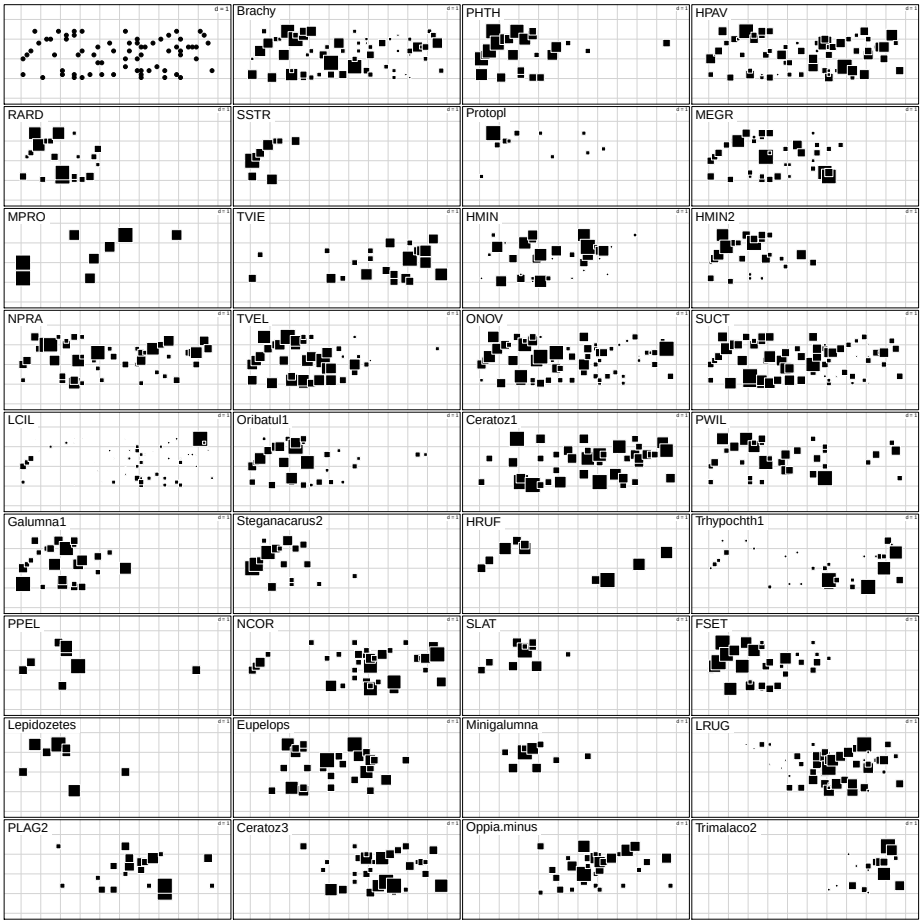



*Expression du lien entre les variables par le lien de chacune d'entre elles avec un score de synthèse.*

## 2.2 Structures spatiales

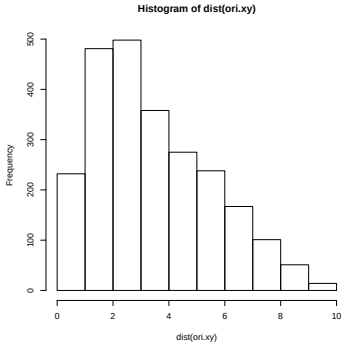
La distribution spatiale des espèces :

```
ori.xy <- oribatid$xy[, c(2, 1)]
names(ori.xy) <- c("x", "y")
par(mfrow = c(9, 4))
s.label(ori.xy, clab = 0, incl = F, addax = F, cpoi = 2)
for (j in 1:35) s.value(ori.xy, oribatid$fau[, j], cleg = 0, , incl = F,
    addax = F, sub = names(oribatid$fau)[j], csub = 2)
```




La question est de savoir si on peut faire une synthèse de ces distributions spatiales. Pour étudier une structure spatiale, il faut définir une relation de voisinage entre les points. Pour plus de détails, voir `http://pbil.univ-lyon1.fr/R/fichestd/ter4.pdf` :

```
par(mfrow = c(1, 1))
hist(dist(ori.xy))
```




On met en place un voisinage par la distance. Le nombre de voisins par point :

```
library(spdep)
ori.dn <- dnearneigh(as.matrix(ori.xy), 0, 1)
print(sort(card(ori.dn))[1:5])
```
```
[1] 0 1 1 3 3
```

Il y a un point sans voisin. On recommence :

```
ori.dn <- dnearneigh(as.matrix(ori.xy), 0, 1.5)
print(sort(card(ori.dn))[1:5])
```
```
[1] 3 4 5 6 7
```

Théoriquement, la relation de voisinage s'écrit dans une matrice à $n$ lignes et $n$ colonnes qui contient à la ligne $i$ et à la colonne $j$ la valeur 1 si les points $i$ et $j$ sont voisins, 0 sinon. De cette matrice de voisinage, on dérive une matrice de pondération spatiale $\mathbf{W} = [w_{ij}]$. Cette matrice de pondération peut subir une transformation. On considérera dans cette fiche, les transformations suivantes :

- $\mathbf{L} = [l_{ij}] = [w_{ij}/w_{i\bullet}]$
- $\mathbf{F} = [f_{ij}] = [w_{ij}/w_{\bullet\bullet}]$

Par défaut, la fonction `nb2listw` renvoie une pondération de type **L**. La matrice **L** n'a qu'une existence théorique, les calculs se faisant à partir de liste de voisins (classe `listw` de `spdep`) :

```
ori.listw <- nb2listw(ori.dn)
```

A partir de cette matrice de pondération spatiale, on peut calculer et tester une mesure d'autocorrélation. $n$ est le nombre de mesures (unités statistiques) et $\mathbf{W}$ est la matrice des poids de voisinages. $x_i$ est la valeur de l'unité statistique $i$ et $z_i = x_i - \bar{x}$. La notation classique est $\sum_{(2)} y_{ij} = \sum_{i,j=1;i\neq j}^{n} y_{ij}$.
Le $I$ de Moran est en général :

$$I = \frac{n \sum_{(2)} w_{ij} z_i z_j}{\sum_{(2)} w_{ij} \sum_{i=1}^{n} {z_i}^2}$$

ce qui désigne quelquefois (définition **F**) :

$$I = \frac{\mathbf{z}^T \mathbf{F} \mathbf{z}}{\sum_{i=1}^{n} {z_i}^2/n}$$

mais le plus souvent (somme par ligne de **L** vaut 1, la somme totale vaut $n$) :

$$I = \frac{\mathbf{z}^T \mathbf{L} \mathbf{z}/n}{\sum_{i=1}^{n} {z_i}^2/n}$$

Le vecteur **Lz** contient pour chaque point, la moyenne sur ses voisins. On l'apelle le *lag vector* (vecteur retard). On le calcule à l'aide de la fonction `lag.listw`. Anselin [1996] propose d'étudier la relation entre **z** et **Lz** par une régression linéaire. Les résultats sont représentés sur un graphique bivarié, le *Moran scatterplot*. En abscisse, on place les valeurs d'une variable, en ordonnée la moyenne des valeurs des voisins (*lag vector*).
Le $c$ de Geary vaut :

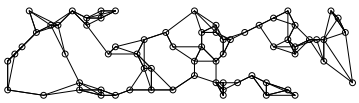
$$c = \frac{\sum_{(2)} w_{ij}(x_i - x_j)^2}{2\sum_{(2)} w_{ij} \sum_{i=1}^{n} {z_i}^2/(n-1)}$$

qui est utilisé souvent comme :

$$c = \frac{\sum_{(2)} f_{ij}(x_i - x_j)^2}{2\sum_{i=1}^{n} {z_i}^2/(n-1)}$$

Toutes ces approches considèrent une seule variable. Quand on en a plusieurs, on peut opérer variable par variable. On peut également travailler sur les scores d'une analyse multivariée.
La relation de voisinage aux 3 plus proches voisins est implantée. On s'intéresse aux variations liées au voisinage immédiat :

```
nb.ori <- knn2nb(knearneigh(as.matrix(ori.xy), 3))
plot(nb.ori, ori.xy)
```

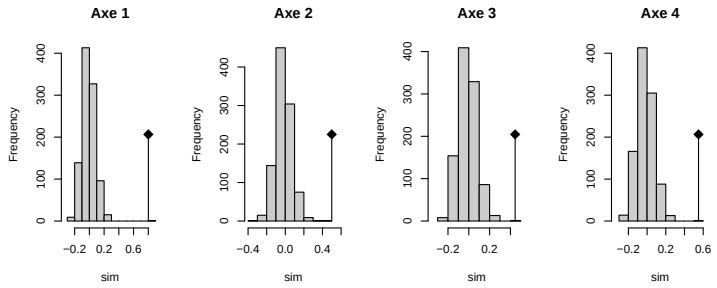


On obtient la pondération de voisinage :

```
ori.listw.3nn <- nb2listw(nb.ori)
```

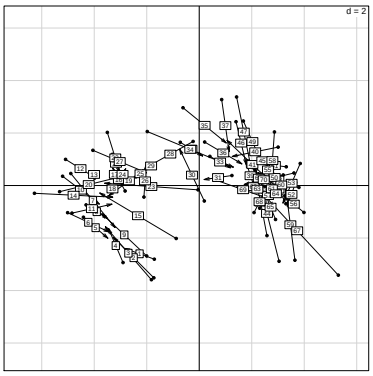
Les test d'autocorrélation sur les 4 premiers axes :

```
par(mfrow = c(1, 4))
for (i in 1:4) {
    w <- moran.mc(ori.pca$l1[, i], ori.listw.3nn, 999)
    plot(as.randtest(w$res, w$statistic), main = paste("Axe", i))
}
```



Pour avoir un *Moran scatterplot* bivarié :

```
w <- as.data.frame(apply(ori.pca$li, 2, function(var) lag.listw(x = ori.listw,
    var)))
row.names(w) <- row.names(ori.mix$li)
s.match(ori.pca$li, w, clab = 0.75)
```




La carte factorielle ordinaire est celle des points à l'origine des flèches. L'extrémité de la flèche est la position moyenne des voisins du point.
Ici, on réalise une synthèse du tableau puis on cherche une structure spatiale de cette synthèse. On a donc une approche indirecte qui n'est pas forcément optimale (une structure forte n'est pas forcément spatialisée). L'ordination sous contrainte spatiale a pour objectif de faire une synthèse (analyse multivariée) de structures spatiales (autocorrélation). Ces deux objectifs devant être satisfaits simultanément.

# 3 Hésitations méthodologiques

On arrive à la question de fond. $n$ unités statistiques portent une pondération de voisinage. Un schéma de dualité, objet de la classe `dudi` ou triplet statistique à $n$ lignes et $p$ colonnes s'écrit $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Le but est d'introduire le point de vue de voisinage $\mathbf{L}$ dans l'analyse de $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ en partant de l'indice de Moran. Ce choix est loin d'être naturel et Lebart a fait le contraire en partant de l'indice de Geary.

## 3.1 L'école de Lebart : variances et covariances locales

Une remarque s'impose sur le choix par Lebart de $c$ au détriment de $I$ dans l'approche multivariée. L'indice de Geary se réécrit :

$$c = \frac{\mathbf{z}^T (\mathbf{D} - \mathbf{F}) \mathbf{z}}{\sum_{i=1}^{n} z_i{}^2/(n-1)}$$

La matrice diagonale $\mathbf{D}$ contient des poids de voisinage pour les points $\mathbf{D} = diag(f_{1\bullet}, \cdots, f_{n\bullet})$ avec $f_{i\bullet} = \sum_{j=1}^{n} f_{ij}$.

$\mathbf{z}^T (\mathbf{D} - \mathbf{F}) \mathbf{z}$ mesure la variance locale. Elle s'écrit :

$$\mathbf{z}^T (\mathbf{D} - \mathbf{F}) \mathbf{z} = \frac{1}{2} \sum_{i,j} f_{ij} (x_i - x_j)^2 = \frac{1}{2} \sum_{i,j} f_{ij} (z_i - z_j)^2$$

La généralisation de Lebart [1969] introduit la matrice de covariance spatiale $\mathbf{X}^T(\mathbf{D}-\mathbf{F})\mathbf{X}$ à partir des graphes de voisinage.

L'idée a été reprise par Monestiez [1978] et généralisée aux pondérations quelconques dans le cadre de l'ACP par Le Foll [1982]. On citera également les travaux de Mom [1988] et Méot et al. [1993]. Le point de vue initial de la variance locale qui donne pour deux tableaux l'analyse de covariance locale [Chessel and Sabatier, 1993] est conservé. Toutes ces approches portent sur la variabilité de voisinage.

L'idée de Lebart ne s'est pas imposée. Le $c$ de Geary est indépendant du centrage puisqu'on ne prend en compte que des différences de valeurs. C'est une forme quadratique positive qui donne une métrique. La norme associée est la variance locale, le produit scalaire est la covariance locale et en introduisant en analyse de données cette métrique on obtient la famille des analyses locales. C'est simple et mathématiquement élégant, malheureusement ces analyses locales maximisent la variance locale et cet objectif est contraire à la majorité des intentions des expérimentateurs. En effet, que cherche-t-on en général ? Des combinaisons de variables les plus cartographiables, les plus lissées (des modèles spatiaux) donc des variables avec un *minimum* de variance locale (entre voisins). Que faire d'une analyse élégante qui est opposé au besoin le plus répandu ? La conséquence s'impose : les analyses locales sont peu utilisées.

## 3.2 L'école de l'auto-corrélation spatiale multivariée

Seul Wartenberg [1985b] a osé casser la contrainte qu'une analyse doit donner des valeurs propres positives. Il diagonalise $\mathbf{M} = \mathbf{X}^T\mathbf{W}\mathbf{X} = \mathbf{X}^T\mathbf{F}\mathbf{X}$ non sans précaution :

> *An important difference between this approach and PCA must be pointed out. Unlike* **R**, *the product-moment correlation matrix that is decomposed in PCA,* **M** *is not positive definite. That is,* **M** *can have negative eigenvalues, which* **R** *cannot. These negative eigenvalues are as important as positive eigenvalues but are of a qualitatively different type. They represent spatial interaction (covariance) that is more important than spatial pattern (variance). . . . To avoid this situation, data yielding negative eigenvalues are not used in this paper. All examples have large eigenvalues that are positive only.*

Il sait que son analyse pourrait donner de grandes valeurs propres négatives ayant du sens mais le cache provisoirement. Il y a cependant une contradiction en ce sens que l'indice de Moran prend tout son intérêt sur un lien **L** et que l'analyse utilise l'indice de Moran sur un lien **F**. Ces hésitations font qu'il y a peu d'utilisateurs de ces propositions auxquelles on préfère les classifications sous contraintes spatiales (*spatial clustering*) ou les méthodes géostatistiques multivariées [Wackernagel, 2003] comme dans Monestiez et al. [1994]. Les méthodes d'ordination sous contraintes spatiales sont les grandes absentes de la synthèse remarquable publiée dans *Ecography* (vol. 25, n. 5) par le groupe de travail *Integrating the Statistical Modeling of Spatial Data in Ecology* (`http://www.nceas.ucsb.edu/~liebhold/ecography/`).

On trouve d'autres approches similaires à celle de Wartenberg en géologie ou analyse d'image [Grunsky and Agterberg, 1991, Switzer and Green, 1984]. La seule différence concerne le type de pondération et une décorrélation des variables. Dans tous les cas, on utilise une matrice d'autocorrélation croisée entre variables.

La variance locale est une forme quadratique et a été intégrée naturellement en analyse des données. La notion d'autocorrélation spatiale ne l'est pas. L'autocorrélation n'est un coefficient de corrélation exactement que si le centrage est fait avec une moyenne calculée avec les poids de voisinage des points et si la normalisation est faite en divisant par l'écart-type utilisant la même pondération. En outre, cette forme quadratique n'est pas positive et l'analyse peut avoir des valeurs propres négatives. Cette insertion n'est pas optimum du point de vue mathématique, tout en étant très légitime du point de vue expérimental. C'est moins beau, mais c'est beaucoup plus utile.
On peut concilier les deux points de vue [Thioulouse et al., 1995] en n'utilisant que des données $\mathbf{D}$-normalisées, c'est-à-dire en prenant :

$$\bar{x} = \sum_{i=1}^{n} f_{i\bullet} x_i \quad var\left(\mathbf{x}\right) = \sum_{i=1}^{n} f_{i\bullet} \left(x_i - \bar{x}\right)^2 \quad z_i = \frac{x_i - \bar{x}}{\sqrt{var\left(\mathbf{x}\right)}}$$

Alors pour toute variable ayant subi ce traitement :

$$\mathbf{z}^T\mathbf{D}\mathbf{z} = \mathbf{1} = \mathbf{z}^T\left(\mathbf{D} - \mathbf{F}\right)\mathbf{z} + \mathbf{z}^T\mathbf{F}\mathbf{z}$$

Variance totale = 1 = variance locale + autocovariance.

Dans cette décomposition, deux termes seulement sur les trois sont toujours positifs. Pour un processus "lisse" donc fortement cartographiable, la variance locale est faible (mais positive) et la covariance locale est positive et forte. Pour un processus à forte variation entre voisins, autocorrélée négativement, la variance locale est plus forte que la variance et l'autocovariance est négative. Les deux statistiques disent la même chose tandis que leur somme est constante.

On pourrait croire la question résolue mais ce point de vue cache un gros inconvénient. Dans l'approche inférentielle, en effet, la pondération non uniforme qui intervient dans le calcul de la moyenne et la variance fait que cette moyenne et cette variance ne sont pas des invariants dans l'espace des $n!$ permutations des données.

Si une variable n'a pas de structure spatiale (hypothèse nulle), sa moyenne et sa variance sont estimées par des unités statistiques toutes égales. Si au contraire elle en a une, et c'est ce qu'on veut mettre en évidence au niveau multivarié, les points sont d'autant plus importants qu'ils ont un poids de voisinage plus grands. Un point relativement isolé dans une telle analyse est une perturbation qui n'a pas grand chose à dire, un point central joue un grand rôle dans l'analyse de la structure. La pondération uniforme des données impose la pondération uniforme, le bilan multivarié impose plutôt la pondération de voisinage. Nous nous en tiendrons ici au point de vue de Wartenberg dans la lignée de Moran.

# 4 La fonction multispati

## 4.1 Paramètres

La fonction utilise un quadruplet $\left(\underset{n\times p}{\mathbf{X}}, \underset{p\times p}{\mathbf{Q}}, \underset{n\times n}{\mathbf{D}}, \underset{n\times n}{\mathbf{L}}\right)$ dont les dimensions sont indiquées en associant une pondération de voisinage (classe `listw`) à un schéma de dualité (classe `dudi`).
Une analyse de base ou triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ ou schéma de dualité (Escoufier 1987) a de nombreuses variantes comme l'ACP, l'ACM, l'AFC notamment. Voir l'article de Chessel et al. [2004] pour un inventaire de la librairie `ade4`.
A ce triplet, on associe un opérateur de retard $\mathbf{L}$ qui permet de calculer $\mathbf{Y} = \mathbf{LX}$ où chaque valeur initiale au point $i$ de la variable $j$ est remplacée par la moyenne des valeurs des voisins de $i$ pour la même variable $j$. Pour une variable on a $\mathbf{y} = \mathbf{Lx}$ et le graphe du couple $(\mathbf{x}, \mathbf{y})$ est le *Moran scatterplot* d'Anselin. Ainsi étendu, l'opération génère un deuxième tableau totalement apparié au premier et donc un deuxième nuage de $n$ points de $\mathbb{R}^p$ qu'on peut projeter sur les axes principaux.
Un analyse ordinaire maximise la variance projetée. L'analyse sous contrainte spatiale garde une part de cette propriété mais intégre le lien de voisinage.

## 4.2 Principes

On comprend que chaque axe de $\mathbb{R}^p$ définit un système de coordonnées qui est plus ou moins autocorrélé. Les axes principaux de l'analyse simple maximise la variance projetée et n'ont aucune propriété d'autocorrélation particulière. On cherche alors ceux qui maximisent l'autocorrélation. La solution n'est pas ordinaire car le critère est :

$$I\left(\mathbf{XQv}\right) = \frac{\mathbf{v}^T\mathbf{Q}^T\mathbf{X}^T\mathbf{DLXQv}}{\mathbf{v}^T\mathbf{Q}^T\mathbf{X}^T\mathbf{DXQv}}$$

Considérons la matrice $\mathbf{H} = \frac{1}{2}\mathbf{X}^T\left(\mathbf{L}^T\mathbf{D} + \mathbf{DL}\right)\mathbf{XQ}$. Elle est $\mathbf{Q}$-symétrique et possède une base de vecteurs propres $\mathbf{Q}$-orthonormés. Le premier vecteur propre $\mathbf{v}_1$ associé à la plus grande valeur $\lambda_1$ réalise le maximum de :

$$\langle \mathbf{Hv} \mid \mathbf{v} \rangle_{\mathbf{Q}} = \mathbf{v}^T\mathbf{Q}^T\mathbf{X}^T\mathbf{DLXQv} = \|\mathbf{XQv}\|^2_{\mathbf{D}}\, I\left(\mathbf{v}\right) = var\left(\mathbf{XQv}\right) I\left(\mathbf{XQv}\right)$$

Le cas particulier pour une ACP normée est l'analyse de Wartenberg [1985a] quand on utilise un lien normalisé par lignes. La recherche de la base de vecteurs propres de $\mathbf{H}$ et les procédures associées sont implémentées dans la fonction `multispati`. Un test de permutations associé considère que les lignes du tableau et leur poids dans l'analyse sont attribués au hasard dans l'espace. La fonction `multispati.rtest` fait le calcul dans R et `multispati.randtest` le fait en C avec une fonction externe. La statistique utilisée est :

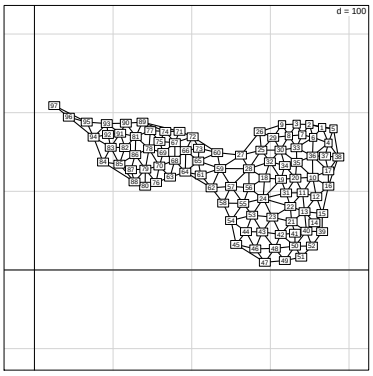
$$r = \frac{Trace\left(\mathbf{X}^T\mathbf{DLXQ}\right)}{Trace\left(\mathbf{X}^T\mathbf{DXQ}\right)}$$

# 5 Illustrations

## 5.1 Analyse des correspondances à composantes cartographiables

Les données :

```
data(mafragh)
maf.xy <- mafragh$xy
maf.flo <- mafragh$flo
maf.listw <- nb2listw(neig2nb(mafragh$neig))
s.label(maf.xy, neig = mafragh$neig, clab = 0.75)
```

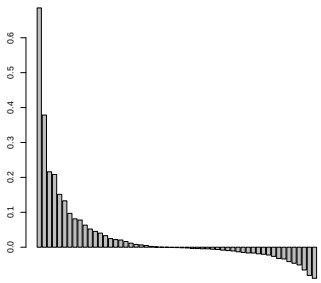



L'analyse simple :

```
maf.coa <- dudi.coa(maf.flo, scannf = F, nf = 3)
```

L'analyse spatialisée :

```
maf.coa.ms <- multispati(maf.coa, maf.listw, scannf = FALSE, nfposi = 2,
    nfnega = 0)
barplot(maf.coa.ms$eig)
```




Comparer l'analyse simple et l'analyse spatialisée par :

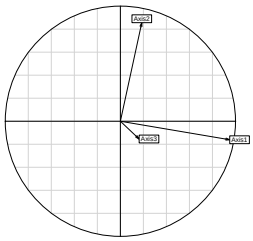
```
summary(maf.coa.ms)
```

```
Multivariate Spatial Analysis
Call: multispati(dudi = maf.coa, listw = maf.listw, scannf = FALSE,
    nfposi = 2, nfnega = 0)
Scores from the initial duality diagramm:
          var       cum     ratio     moran
RS1 0.8691476 0.8691476 0.1043473 0.7250457
RS2 0.6491089 1.5182565 0.1822775 0.4834366
RS3 0.5975380 2.1157944 0.2540161 0.2263971

Multispati eigenvalues decomposition:
          eig       var     moran
CS1 0.6854591 0.8332937 0.8225901
CS2 0.3785339 0.5865926 0.6453097
```

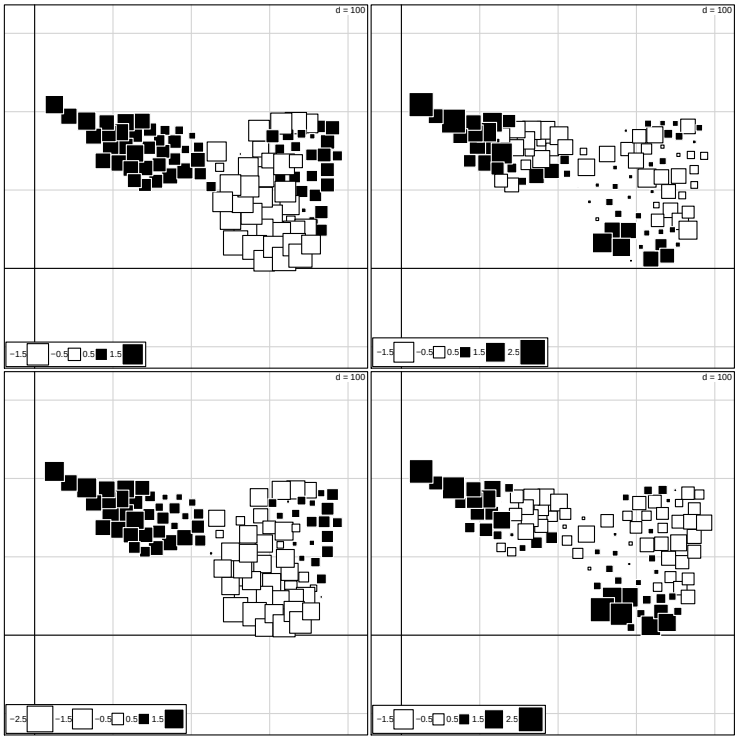
Représenter la projection des trois premiers axes de l'analyse simple sur le plan des deux premiers axes de l'analyse spatialisée. Globalement le plan 1-2 est largement conservé :

```
s.corcircle(maf.coa.ms$as)
```



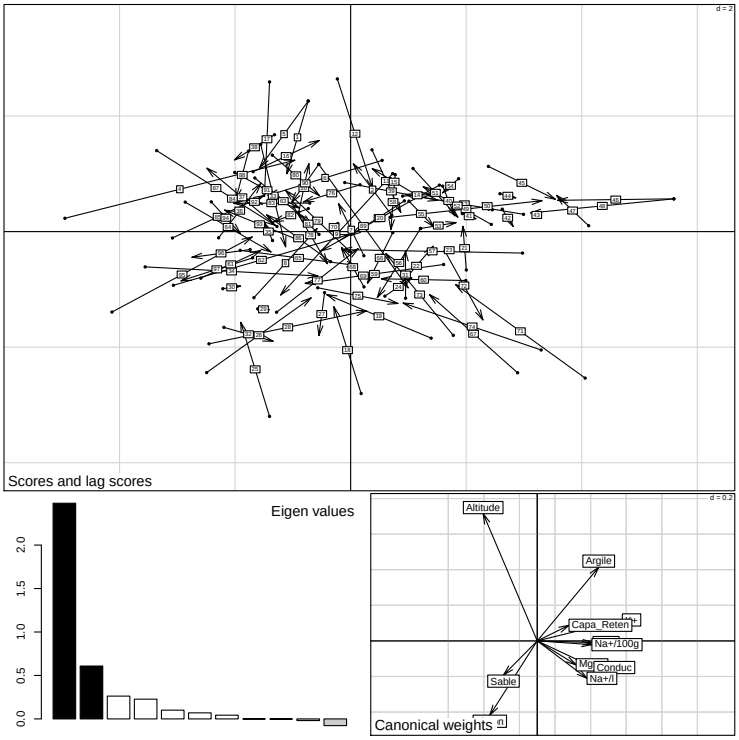

La cartographie des scores :

```
par(mfrow = c(2, 2))
s.value(mafragh$xy, maf.coa$li[, 1])
s.value(mafragh$xy, maf.coa$li[, 2])
s.value(mafragh$xy, maf.coa.ms$li[, 1])
s.value(mafragh$xy, maf.coa.ms$li[, 2])
```




La lecture des résultats est simplifiée.
L'analyse nous apprend que la partie spatialisée de l'interprétation se réduit à deux dimensions. `maf.coa$c1` contient des scores des espèces de moyenne nulle et de variance unité pour la distribution de `maf.coa$cw`. `maf.coa$li` place les sites à la moyenne des espèces qu'ils contiennent. On a les mêmes propriétés pour `maf.coa.ms$c1` et `maf.coa.ms$li`.

```
plot(multispati(dudi.pca(mafragh$mil, scannf = F), maf.listw, scannf = F,
    nfposi = 2, nfnega = 2))
```

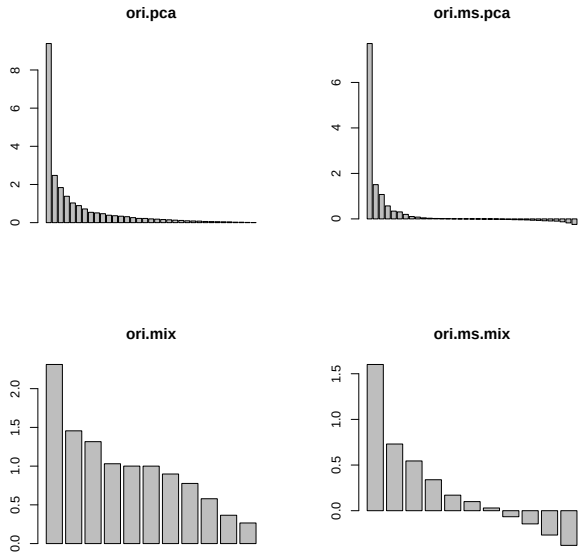



La végétation est beaucoup plus spatialisée que l'analyse de sol.

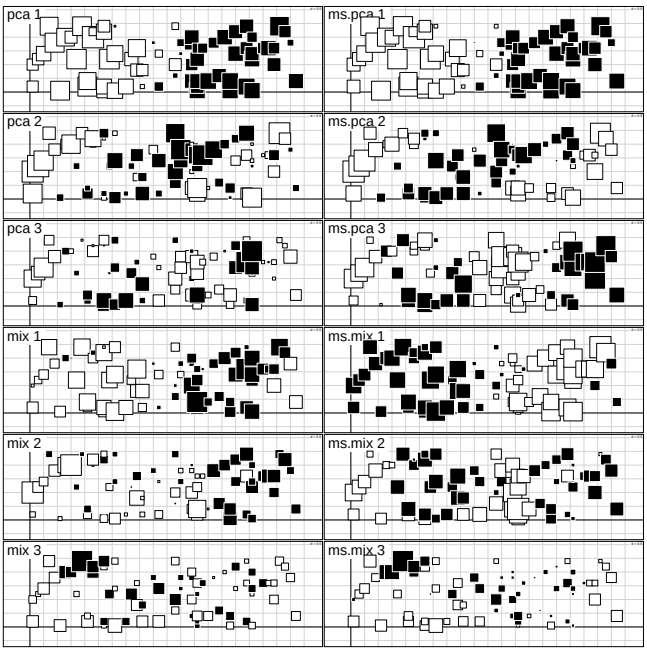
## 5.2 Gradients

On reprend l'exemple des oribates :
La structure faunistique est-elle spatialisée ? Et celle du tableau milieu ?

```
ori.ms.pca <- multispati(ori.pca, ori.listw.3nn, scannf = F, nfposi = 4,
    nfnega = 0)
ori.ms.mix <- multispati(ori.mix, ori.listw.3nn, scannf = F, nfposi = 4,
    nfnega = 0)
par(mfrow = c(2, 2))
barplot(ori.pca$eig, main = "ori.pca")
barplot(ori.ms.pca$eig, main = "ori.ms.pca")
barplot(ori.mix$eig, main = "ori.mix")
barplot(ori.ms.mix$eig, main = "ori.ms.mix")
```



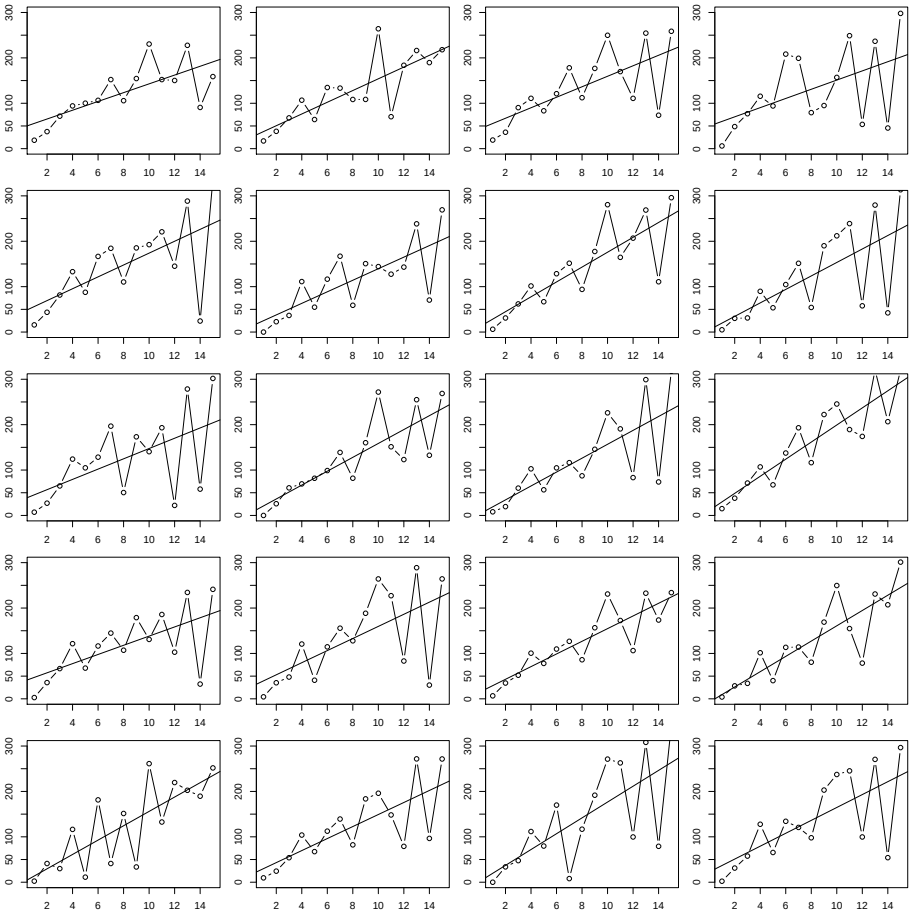

```
par(mfcol = c(6, 2))
for (i in 1:3) s.value(ori.xy, ori.pca$l1[, i], sub = paste("pca",
    i), csub = 3, cleg = 0)
for (i in 1:3) s.value(ori.xy, ori.mix$l1[, i], sub = paste("mix",
    i), csub = 3, cleg = 0)
for (i in 1:3) s.value(ori.xy, ori.ms.pca$li[, i], sub = paste("ms.pca",
    i), csub = 3, cleg = 0)
for (i in 1:3) s.value(ori.xy, ori.ms.mix$li[, i], sub = paste("ms.mix",
    i), csub = 3, cleg = 0)
```



Le lien très fort des deux tableaux avec l'espace est une contrainte énorme qui cache peut-être les relations espèces-environnement ayant un intérêt écologique. D'où les travaux qui visent à débarrasser les données des composantes spatiales [Borcard and Legendre, 1994, Borcard et al., 1992, Méot et al., 1998].
D'une manière générale, la contrainte simplifie l'interprétation quand on la cherche dans l'espace.

## 5.3 Croissance et alternance, global et local

```
data(clementines)
par(mfrow = c(5, 4), mar = c(2, 2, 1, 1))
w0 <- 1:15
for (i in 1:20) {
    plot(w0, clementines[, i], type = "b", ylim = c(0, 300))
    abline(lm(clementines[, i] ~ w0))
}
```

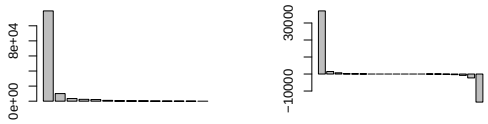


*Production de 20 arbres fruitiers pendant 15 ans*

Les analyses multivariées :

```
clem.nb <- cell2nb(15, 1)
clem.listw <- nb2listw(clem.nb)
clem.pca <- dudi.pca(clementines, scale = F, scannf = F)
```

```
clem.ms <- multispati(clem.pca, clem.listw, scannf = F, nfposi = 2,
    nfnega = 2)
par(mfrow = c(1, 2))
barplot(clem.pca$eig)
barplot(clem.ms$eig)
```



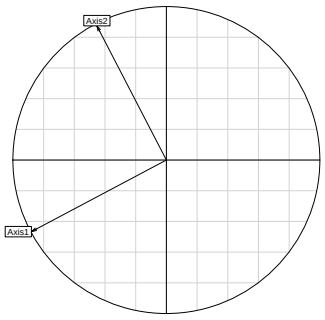

```
summary(clem.ms)
```

```
Multivariate Spatial Analysis
Call: multispati(dudi = clem.pca, listw = clem.listw, scannf = F, nfposi = 2,
    nfnega = 2)
Scores from the initial duality diagramm:
          var      cum     ratio      moran
RS1 119761.38 119761.4 0.8510445  0.2095731
RS2  10059.13 129820.5 0.9225264 -0.5008158

Multispati eigenvalues decomposition:
               eig          var         moran
CS1   3.710362e+04 94737.595443  3.916462e-01
CS2   1.474828e+03  2156.711342  6.838320e-01
CS13 -1.842444e-11     1.733618  1.717053e-15
CS14 -5.295258e+00    88.552101 -5.979822e-02
```

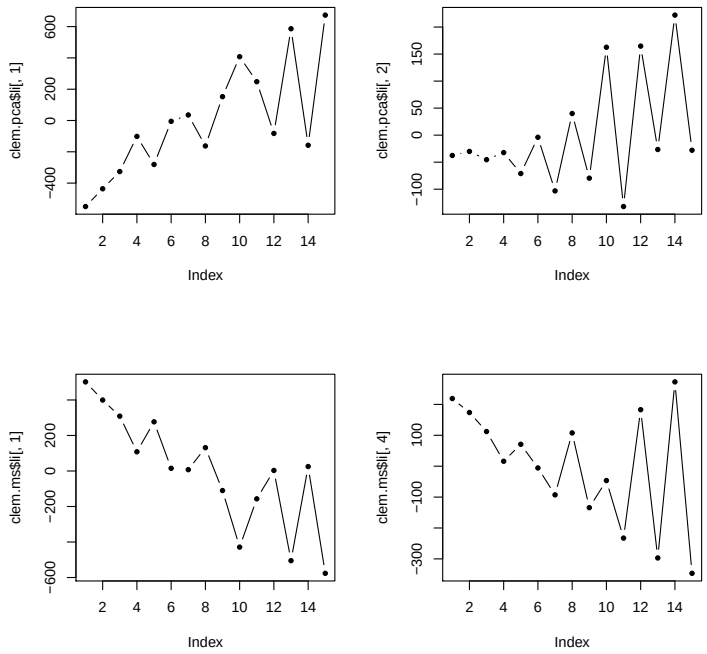
Le plan des deux premiers axes de l'ACP simple maximise l'inertie projetée. Le total est 129821 soit 92.25 %. Il ne peut être dépassé par le plan 1-20 de l'ACP spatiale qui donne cependant un résultat très proche.

```
s.corcircle(clem.ms$as, 1, 4)
```

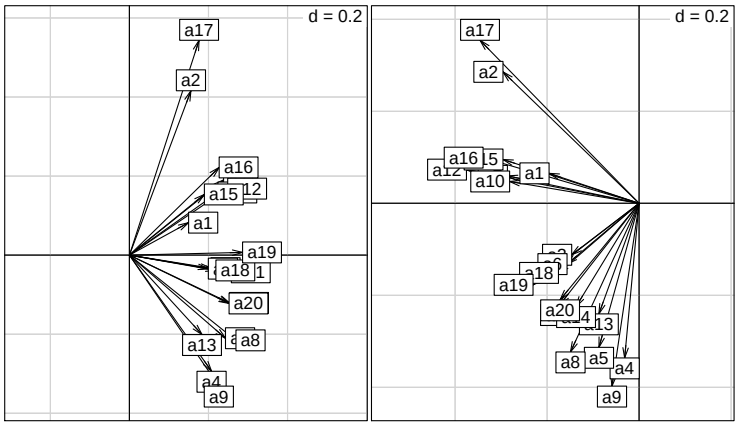



Les deux plans sont les mêmes.

```
par(mfrow = c(2, 2))
plot(clem.pca$li[, 1], type = "b", pch = 20)
plot(clem.pca$li[, 2], type = "b", pch = 20)
plot(clem.ms$li[, 1], type = "b", pch = 20)
plot(clem.ms$li[, 4], type = "b", pch = 20)
```

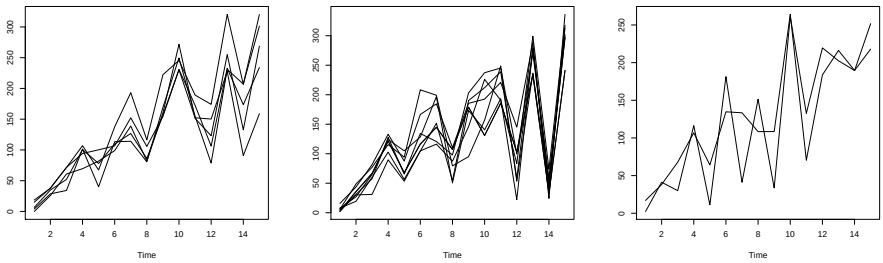



```
par(mfrow = c(1, 2))
s.arrow(clem.pca$c1)
s.arrow(clem.ms$c1, 1, 4)
```




La croissance et l'alternance sont deux composantes de la variabilité. L'analyse spatiale les sépare clairement et identifie le groupe des croissances les plus régulières (1,10,12,15 et 16) et des alternances les plus marquées (4,9,5,13,8,11,20,13).

```
par(mfrow = c(1, 3))
ts.plot(clementines[, c(1, 10, 12, 15, 16)])
ts.plot(clementines[, c(4, 9, 5, 13, 8, 11, 20, 13)])
ts.plot(clementines[, c(2, 17)])
```




L'analyse sous contrainte spatiale reste une analyse d'inertie fortement orientée dans l'interprétation vers la lecture de *l'autocorrélation*.

# Références


L. Anselin. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M.M. Fischer, H.J. Scholten, and D. Unwin, editors, *Spatial analytical perspectives on GIS*, pages 111–125. Taylor and Francis, London, 1996. Absent.

D. Borcard and P. Legendre. Environmental control and spatial structure in ecological communities : an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics*, 1 :37–61, 1994.

D. Borcard and P. Legendre. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, 153 :51–68, 2002.

D. Borcard, P. Legendre, and P. Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73 :1045–1055, 1992.

D. Borcard, P. Legendre, C. Avois-Jacquet, and H. Tuomisto. Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, 85(7) :1826–1832, 2004.

D. Chessel and R. Sabatier. Couplage de triplets statistiques et graphes de voisinage. In B. & Coll. Asselain, editor, *Biométrie et Données spatio-temporelles*, pages 28–37. Société Française de Biométrie, ENSA, Rennes, 1993.

D. Chessel, A.-.B. Dufour, and J. Thioulouse. The ade4 package - I : One-table methods. *R News*, 4 :5–10, 2004.

S. Dray, P. Legendre, and P.R. Peres-Neto. Spatial modeling : a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). *Ecological Modelling*, 196 :483–493, 2006.

E.C. Grunsky and F.P. Agterberg. Spfa : a fortran-77 program for spatial factor analysis of multivariate data. *Computers & Geosciences*, 17 :133–160, 1991.

Y. Le Foll. Pondération des distances en analyse factorielle. *Statistique et Analyse des données*, 7 :13–31, 1982.

L. Lebart. Analyse statistique de la contiguïté. *Publication de l'Institut de Statistiques de l'Université de Paris*, 28 :81–112, 1969.

A. Mom. *Méthodologie statistique de la classification des réseaux de transports*. PhD thesis, USTL - Montpellier, 1988.

P. Monestiez. *Présentation de deux méthodes utilisant la notion de contiguïté pour l'analyse des données géographiques*. Thèse de docteur-ingénieur, Paris VI, 1978.

P. Monestiez, M. Goulard, and G. Charmet. Geostatistics for spatial genetic structures : study of wild populations of perennial ryegrass. *Theoretical and Applied Genetics*, 88 :33–41, 1994.

A. Méot, D. Chessel, and R. Sabatier. Opérateurs de voisinage et analyse des données spatio-temporelles. In J.D. Lebreton and B. Asselain, editors, *Biométrie et Environnement*, pages 45–72. Masson, Paris, 1993.

A. Méot, P. Legendre, and D. Borcard. Partialling out the spatial component of ecological variation : questions and propositions in the linear modelling framework. *Environmental and Ecological Statistics*, 5 :1–27, 1998.

P.R. Peres-Neto, P. Legendre, S. Dray, and D. Borcard. Variation partitioning of species data matrices : estimation and comparison of fractions. *Ecology*, 87 :2614–2625, 2006.

P. Switzer and A. A. Green. Min/max autocorrelation factors for multivariate spatial imagery. Technical report, Tech. rep. 6, Stanford University, 1984.

C.J.F. ter Braak. Partial canonical correspondence analysis. In H.H. Bock, editor, *Classification and related methods of data analysis*, pages 551–558. Elsevier Science Publishers B. V., Amsterdam, 1988.

J. Thioulouse, D. Chessel, and S. Champely. Multivariate analysis of spatial patterns : a unified approach to local and global structures. *Environmental and Ecological Statistics*, 2 :1–14, 1995.

H. Wackernagel. *Multivariate Geostatistics : An introduction with applications*. Springer, Berlin, 2003.

D. Wartenberg. Multivariate spatial correlation : a method for exploratory geographical analysis. *Geographical Analysis*, 17(4) :263–283, 1985a.

D.E. Wartenberg. Multivariate spatial correlations : a method for exploratory geographical analysis. *Geographical Analysis*, 17 :263–283, 1985b.