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

Dissemblance et diversité

D. Chessel

Notes de cours cssb9

On aborde les bases d'un domaine plein d'avenir : celui de la mesure de la biodiversité. Dissimilarités métriques, euclidiennes, ultramétriques : définitions et mode de calcul. Leurs usages. Typologie et diversité sont deux concepts indissociables.

Table des matières

1 Introduction **2**

2 Calcul des dissimilarités **4**
2.1 Définitions 4
2.2 Distance de Mahalanobis 6
2.3 Dissimilarités écologiques 7
2.4 Distances génétiques, taxonomiques, phylogénétiques 8

3 Propriétés des dissimilarités **10**
3.1 Dissimilarités euclidiennes 10
3.2 Que faire des distances non euclidiennes 13
3.3 Dissimilarités ultramétriques 15

4 Des espèces aux communautés **18**
4.1 La question de la mesure de la diversité 18
4.2 Diversité et différences 23
4.3 Typologie sur différences 25
4.4 Maximiser la diversité 30

Références **35**

1 Introduction

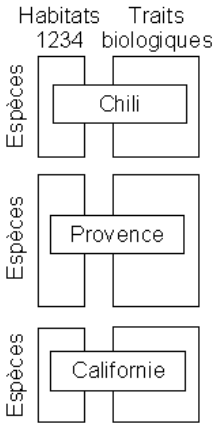
Nous partirons d'un problèmes posé par des ornithologues de renom [2] décrit dans :

`http://pbil.univ-lyon1.fr/R/pps/pps023.pdf`

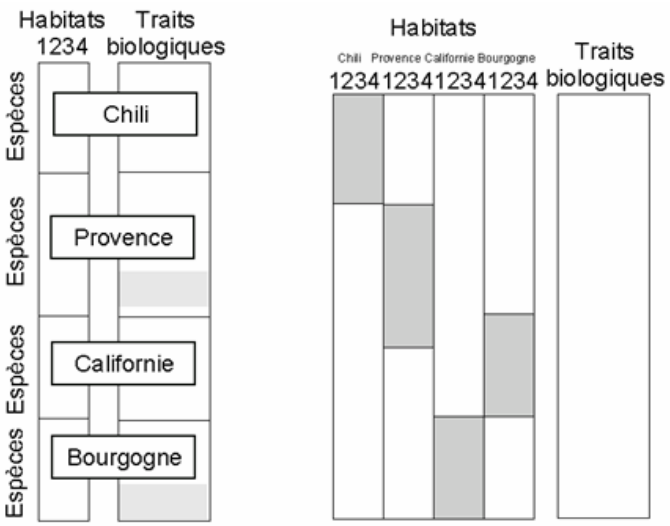
L'article contient toutes les données consignée dans l'objet `ecomor` :

```
library(ade4)
data(ecomor)
```

Les données originales, directement accessibles dans l'article et reproduites dans la liste `ecomor` présentent une petite erreur d'édition qui a une grande importance sémantique. Les auteurs confrontent trois cortèges faunistiques complètement disjoints. Trois régions soumis à un bioclimat méditerranéen (Chili, Californie et Provence) ont une avifaune nicheuse sans aucune espèce commune. Elles sont cependant structurées par le même gradient d'ouverture du milieu (prairies, maquis, forêts). La comparaison des trois univers conduit à la définition de 4 types d'habitats comparables. D'autre part un certain nombre de traits écologiques et de variables morphométriques sont enregistrées pour toutes les espèces nicheuses. Le schéma de la structure des données est :



Chaque tableau espèces-habitats est en présence-absence ce qui définit le profil de chaque espèce du type `1000` pour les spécialistes des milieux ouverts à `0001` pour les forestières strictes en passant par `1111` pour les espèces qu'on rencontre partout. Les auteurs ont rajouté une quatrième région de référence non méditerranéenne (Bourgogne), pour comparaison. Mais le nouveau cortège a la propriété d'avoir une partie de ses espèces en commun avec la Provence. Dans l'article une étoile indique qu'une espèce est commune au pool Provence et au pool Bourgogne mais seul est mentionné le profil en Provence ce qui laisse à penser qu'il est le même dans les deux régions. Or ce n'est pas vrai et J. Blondel m'a transmis les précisions nécessaires. La modification porte sur 13 des 129 espèces (et uniquement sur le profil d'habitat) mais oblige à choisir entre deux schémas :



A gauche, on a répété les espèces en commun (la partie grise signale la répétition). A droite on a utilisé leur profil global (la partie grise seule ne contient pas uniquement des 0). La perturbation est petite mais a grande signification car elle conduit à deux problèmes différents. La forme de gauche est un ensemble de quatre couples de tableaux écologiques qui demande la comparaison de quatre évolutions du contenu en traits biologiques entre les quatre types d'habitat. La forme de droite est un couple unique qui envisage la variation en traits biologiques soit entre régions (analyse inter), soit entre strates d'une même région (analyse intra), soit entre tous les éléments (analyse simple). La situation est encore compliquée par le schéma des hypothèses de l'article (figure 1). On y

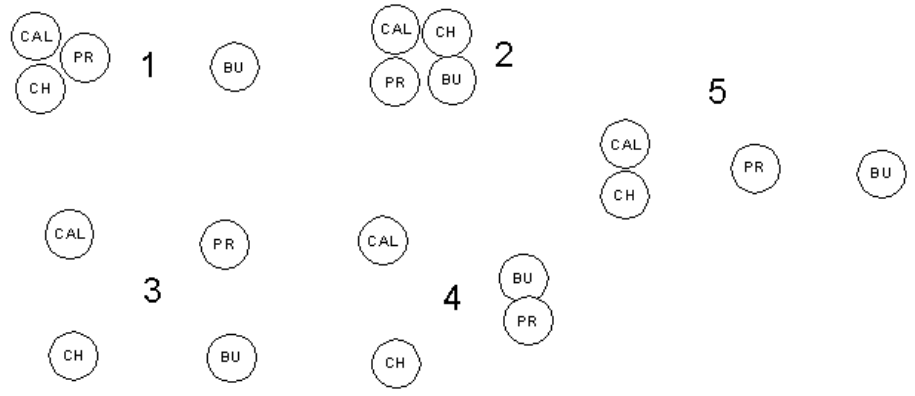
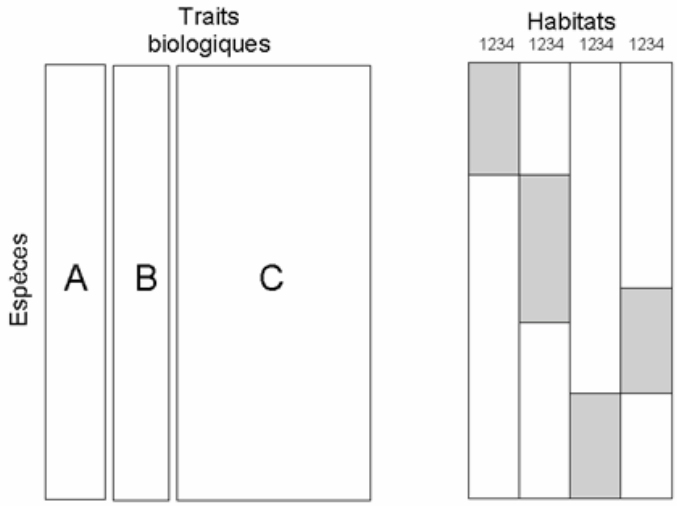


FIG. 1 – 1 - Modèle de la convergence bioclimatique, 2/3 - Absence de convergence, 4 - Modèle phylogénétique, 5 - Convergence bioclimatique ajustée (CAL Californie, CH Chili, PR Provence, BU Bourgogne)

voit une typologie de régions soit faible (2) soit forte (3) soit organisée (1, 4, 5) mais sans référence au gradient environnemental commun. Or comment s'intéresser à la variation entre régions sans tenir compte que chacune d'entre elles est

elle-même une structure (et bien connue pour être forte) ? La question est donc très ouverte. On a conservé les données sous la forme (A Lieu d'alimentation, B Régime alimentaire, C Morphométrie) :




```
names(ecomor)
```
```
[1] "forsub"  "diet"    "habitat" "morpho"  "taxo"    "labels"  "categ"
```

La question est comment mesurer, comparer, décrire des structures faunistiques avec des cortèges spécifiques distincts. Peut-on sortir du comptage pur et simple des espèces pour mesurer et comparer la biodiversité. Ces questions sont d'actualité [9], c'est le moins qu'on puisse dire ! L'essentiel des questions introduites ici sont développées dans la thèse de S. Pavoine [21].

2 Calcul des dissimilarités

Chacun sait qu'il y a plusieurs manières de mesurer une distance, à vol d'oiseau, en kilomètres d'autoroute ou en minutes de TGV. Il y en a encore bien plus de manières de mesurer la dissimilarité au sens large entre objets.

2.1 Définitions

En mathématiques, on appelle *distance* définie sur un ensemble $\mathcal{E}$ une fonction d de $\mathcal{E} \times \mathcal{E}$ dans $\mathbb{R}$ qui vérifie pour tout x, y et z éléments de $\mathcal{E}$:

$$\mathrm{d}(x,y) \geq 0$$

$$\mathrm{d}(x,y) = 0 \Rightarrow x = y$$

$$\mathrm{d}(x,y) = d(y,x)$$

$$\mathrm{d}(x,y) \geq \mathrm{d}(x,z) + \mathrm{d}(z,y)$$

En statistique, on appelle *dissimilarité* définie sur un ensemble fini $\mathcal{I}$ à n éléments (numérotés 1, 2,..., i, ..., n) une fonction d de $\mathcal{I} \times \mathcal{I}$ dans $\mathbb{R}$ qui vérifie pour tout i et j :

$$d_{ij} \geq 0$$

$$d_{ii} = 0$$

$$d_{ij} = d_{ji}$$

La dissimilarité est dite métrique (définition 1 dans [8]) si pour tout i, j et k éléments de $\mathcal{I}$ on a :

$$d_{ij} \leq d_{ik} + d_{kj}$$

Une dissimilarité est euclidienne (definition 2 dans [8]) si il existe n points dans un espace euclidien dont les distances deux a deux sont exactement les dissimilarités considérées. On parle aussi de distance euclidienne mais le terme est ambigu. Nous y reviendrons.

Une dissimilarité est ultramétrique si pour tout i, j et k éléments de $\mathcal{I}$ on a :

$$d_{ij} \leq \max(d_{ik}, d_{kj})$$

Nous verrons que cette notion est essentielle pour la mesure de la biodiversité.

En biologie, on utilise le terme de distance pour désigner la différence mesurée entre deux individus, deux populations, deux sites, deux espèces, ..., sans se préoccuper de définition. Pour suivre la coutume, on appellera *matrice de distances* une matrice contenant une *dissimilarité observée*. Les matrices de distances sont donc des matrices carrées (n lignes et n colonnes), contenant des nombres positifs ou nuls, symétriques et ayant des éléments nuls sur la diagonale. Ces objets sont stockés dans R dans la classe `dist` sous la forme d'un vecteur avec des attributs :

```
x <- matrix(rnorm(100), nrow = 5)
d <- dist(x)
d
as.matrix(d)
unclass(d)
```

Le nombre d'indices de dissimilarités proposés en écologie est proprement vertigineux. Calculer une dissimilarité relève de la partie expérimentale (acquisition), s'en servir relève de la partie statistique de l'exploitation des données. Quelques fonctions qui calcule des dissimilarités :

- les distances directement importées ou observées (`as.dist`)
- les distances issues des tableaux (`dist` et `dist.quant`)
- les distances issues des triplets statistiques (`dist.dudi`)
- Les distances issues des données floro-faunistiques en présence-absence (`dist.binary`)
- Les distances génétiques (`dist.genet`)
- Les distances de voisinages (`dist.neig`)
- les distances entre profils (`dist.prop`)
- Les distances phylogénétiques (`newick2phylog`)
- les distances taxonomiques (`dist.taxo`)

Pour utiliser d'autres options, examiner les librairies qui sont orientées écologie et qui sont citées dans :

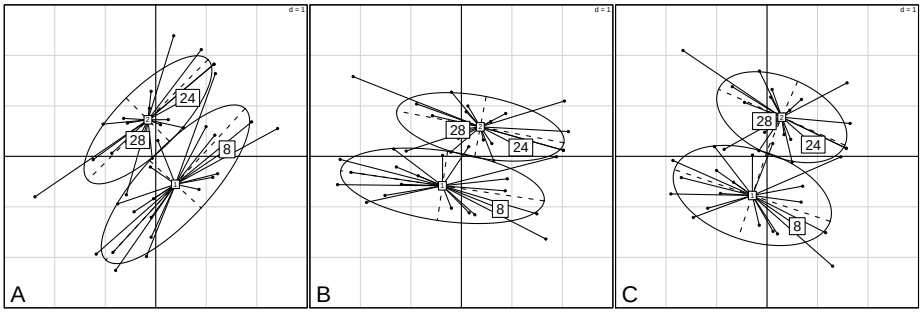
`http://cran.r-project.org/src/contrib/Views/Environmetrics.html`

Quelques cas importants sont à connaître.

2.2 Distance de Mahalanobis

Pour les variables morphométriques, la distance de Mahalanobis s'impose. A l'origine [18] le nom désigne la distance entre populations au sens de l'inverse de la matrice de covariance intra-classe. Par extension on utilise le terme pour désigner la distance entre deux individus au sens de l'inverse de la matrice de covariances totales. Les liens entre les deux est affaire de modèle mais l'utilisateur a besoin de connaître la fonction de cette idée.

```
set.seed(29092006)
par(mfrow = c(1, 3))
library(MASS)
s <- matrix(c(1, 0.8, 0.8, 1), 2)
x1 <- mvrnorm(20, c(0.5, -0.5), s)
x2 <- mvrnorm(20, c(-0.5, 0.5), s)
x <- rbind.data.frame(x1, x2)
row.names(x) <- as.character(1:40)
fac <- factor(rep(1:2, rep(20, 2)))
s.class(x, fac, xlim = c(-3, 3), ylim = c(-3, 3), sub = "A", csub = 3)
choix <- c(8, 24, 28)
s.label(x[choix, ], clab = 2, add.p = T)
w <- dudi.pca(x, scan = F, scal = F)
s.class(w$li, fac, xlim = c(-3, 3), ylim = c(-3, 3), sub = "B",
    csub = 3)
s.label(w$li[choix, ], clab = 2, add.p = T)
s.class(w$l1, fac, xlim = c(-3, 3), ylim = c(-3, 3), sub = "C",
    csub = 3)
s.label(w$l1[choix, ], clab = 2, add.p = T)
```



Utiliser la distance de Mahalanobis, c'est utiliser la distance euclidienne ordinaire sur les coordonnées normalisées de l'ACP en lieu et place des données. On utilise des données décorrélées, en particulier en prenant en compte l'effet taille au niveau d'une seule variable (alors que cet effet compte dans la distance ordinaire comme 95 % des variables). Cette opération est essentielle en morphométrie et est malheureusement importée dans bien d'autres domaines où elle ne sert que de perturbation.

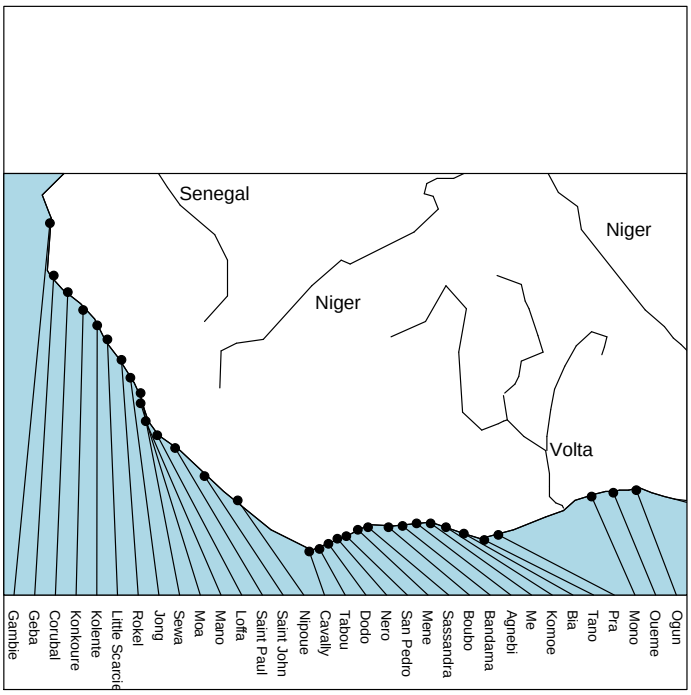
```
as.matrix(dist(x))[choix, choix]
          8       24       28
8  0.000000 1.282952 1.777313
24 1.282952 0.000000 1.294394
28 1.777313 1.294394 0.000000
as.matrix(dist(w$l1))[choix, choix]
          8       24       28
8  0.000000 1.646473 2.180075
24 1.646473 0.000000 1.073575
28 2.180075 1.073575 0.000000
as.matrix(dist.quant(x, meth = 3))[choix, choix]
          8       24       28
8  0.000000 1.646473 2.180075
24 1.646473 0.000000 1.073575
28 2.180075 1.073575 0.000000
```

2.3 Dissimilarités écologiques

Pour les données en présence-absence, les indices sont en grand nombre. Ce sont des indices de similarité S compris entre 0 et 1 qu'on transforme en distance avec $D = \sqrt{1 - S}$. Les données proposées par B. Hugueny et présentées dans :

`http://pbil.univ-lyon1.fr/R/pps/pps050.pdf`

posent clairement le problème (présence-absence de $n = 268$ espèces dans $p = 33$ bassins côtiers de l'Afrique de l'Ouest). Refaire la figure avec la fiche de documentation du *dataset* :



La distance entre deux bassins, du point de vue faunistique est une fonction du nombre d'espèces en commun $a = n_{11}$ et tient compte que du nombre d'espèces qui ne sont que dans le premier $b = n_{10}$ ou dans le second $c = n_{01}$. Doit-on tenir compte du nombre d'espèces absentes dans les deux $d = n_{00}$?

Pour tous les couples on a en tout cas $n = n_{11} + n_{10} + n_{01} + n_{00}$
La fonction `dist.binary` permet d'utiliser 10 indices indiqués par :

```
1 = JACCARD index (1901) S3 coefficient of GOWER & LEGENDRE
s1 = a/(a+b+c) --> d = sqrt(1 - s)
2 = SOCKAL & MICHENER index (1958) S4 coefficient of GOWER & LEGENDRE
s2 = (a+d)/(a+b+c+d) --> d = sqrt(1 - s)
3 = SOCKAL & SNEATH(1963) S5 coefficient of GOWER & LEGENDRE
s3 = a/(a+2(b+c)) --> d = sqrt(1 - s)
4 = ROGERS & TANIMOTO (1960) S6 coefficient of GOWER & LEGENDRE
s4 = (a+d)/(a+2(b+c)+d) --> d = sqrt(1 - s)
5 = CZEKANOWSKI (1913) or SORENSEN (1948) S7 coefficient of GOWER & LEGENDRE
s5 = 2*a/(2*a+b+c) --> d = sqrt(1 - s)
6 = S9 index of GOWER & LEGENDRE (1986)
s6 = (a-(b+c)+d)/(a+b+c+d) --> d = sqrt(1 - s)
7 = OCHIAI (1957) S12 coefficient of GOWER & LEGENDRE
s7 = a/sqrt((a+b)(a+c)) --> d = sqrt(1 - s)
8 = SOKAL & SNEATH (1963) S13 coefficient of GOWER & LEGENDRE
```

```
s8 = ad/sqrt((a+b)(a+c)(d+b)(d+c)) --> d = sqrt(1 - s)
9 = Phi of PEARSON = S14 coefficient of GOWER & LEGENDRE
s9 = ad-bc)/sqrt((a+b)(a+c)(b+d)(d+c)) --> d = sqrt(1 - s)
10 = S2 coefficient of GOWER & LEGENDRE
s10 =  a/(a+b+c+d) --> d = sqrt(1 - s) and unit self-similarity
```

2.4 Distances génétiques, taxonomiques, phylogénétiques

Pour les données génétiques, les définitions des distances sont indépendantes des modes d'utilisation. Elles sont associées aux mécanismes temporels de la dérive des fréquences alléliques et de la séparation des populations. Le pourcentage d'allèles partagés est une mesure de la proximité génétique entre deux populations et ramène au cas précédent.

Plus généralement, soit $\mathbf{A}$ un tableau de fréquences alléliques avec t populations (lignes) et m allèles (colonnes). Soit ν le nombre de loci. Le locus j donne $m(j)$ allèles.

$$m = \sum_{j=1}^{\nu} m(j)$$

Pour la ligne i $(1 \leq i \leq q)$ et l'allèle k $(1 \leq k \leq m(j))$ du locus j $(1 \leq j \leq \nu)$, on note a_{ij}^k la valeur initiale dans le tableau (effectifs).

$$a_{ij}^+ = \sum_{k=1}^{m(j)} a_{ij}^k \qquad p_{ij}^k = \frac{a_{ij}^k}{a_{ij}^+}$$

Soit $\mathbf{P}$ le tableau de terme général p_{ij}^k. On a :

$$p_{ij}^+ = \sum_{k=1}^{m(j)} p_{ij}^k = 1 \qquad p_{i+}^+ = \sum_{j=1}^{\nu} p_{ij}^+ = \nu \qquad p_{++}^+ = \sum_{j=1}^{\nu} p_{i+}^+ = tnu$$

La fonction `dist.genet` propose les distances entre populations à partir des fréquences alléliques p_{ij}^k les plus classiques :

Distance de Nei

$$D_1(a,b) = -ln(\frac{\sum_{k=1}^{\nu}\sum_{j=1}^{m(k)} p_{aj}^k p_{bj}^k}{\sqrt{\sum_{k=1}^{\nu}\sum_{j=1}^{m(k)} \left(p_{aj}^k\right)^2}\sqrt{\sum_{k=1}^{\nu}\sum_{j=1}^{m(k)} \left(p_{bj}^k\right)^2}})$$

Distance angulaire d'Edwards

$$D_2(a,b) = \sqrt{1 - \frac{1}{\nu}\sum_{k=1}^{\nu}\sum_{j=1}^{m(k)}\sqrt{p_{aj}^k p_{bj}^k}}$$

Coefficient de coancestralité ou distance de Reynolds

$$D_3(a,b) = \sqrt{\frac{\sum_{k=1}^{\nu}\sum_{j=1}^{m(k)}\left(p_{aj}^k - p_{bj}^k\right)^2}{2\sum_{k=1}^{\nu}(1 - \sum_{j=1}^{m(k)} p_{aj}^k p_{bj}^k)}}$$

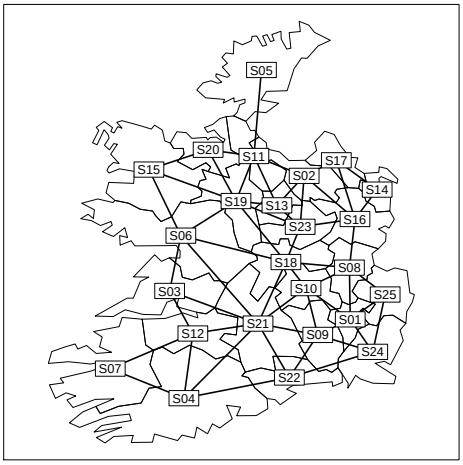
Distance euclidienne dite de Roger

$$D_4(a,b) = \frac{1}{\nu}\sum_{k=1}^{nuy}\sqrt{\frac{1}{2}\sum_{j=1}^{m(k)}\left(p_{aj}^k - p_{bj}^k\right)^2}$$

Distance absolue de Provesti

$$D_5(a,b) = \frac{1}{2\nu}\sum_{k=1}^{\nu}\sum_{j=1}^{m(k)}|p_{aj}^k - p_{bj}^k|$$

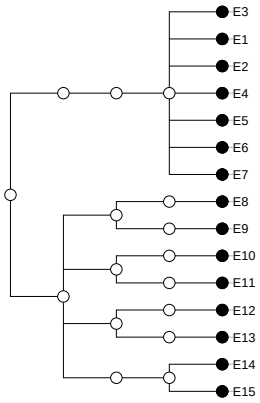
Pour les données spatiales, on utilise la distance ordinaire ou la distance de voisinage (longueur du plus court chemin d'un sommet à un autre) :




```
data(irishdata)
area.plot(irishdata$area.utm, clab = 0)
n0 <- neig(mat01 = irishdata$link > 0)
s.label(irishdata$xy.utm, lab = names(irishdata[[2]]), clab = 1,
    neig = n0, add.plot = T)
dist.neig(n0)
```

```
   1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
2  3
3  3 4
4  3 4 2
5  5 2 4 5
6  3 3 1 2 3
7  4 5 2 1 6 3
8  1 2 3 3 4 2 4
9  1 4 2 2 5 2 3 2
10 1 3 2 2 4 2 3 1 1
11 4 1 3 4 1 2 5 3 4  3
12 3 4 1 1 5 2 1 3 2  2  4
13 4 1 3 4 2 2 5 3 4  3  1  4
14 3 2 5 5 4 4 6 2 4  3  3  5  3
15 4 3 2 3 3 1 4 3 3  3  2  3  2  4
16 2 1 4 4 3 3 5 1 3  2  2  4  2  1  3
17 3 1 5 5 3 4 6 2 4  3  2  5  2  1  4  1
18 2 2 2 2 3 1 3 1 2  1  2  2  2  3  2  2  3
19 3 2 2 3 2 1 4 2 3  2  1  3  1  3  1  2  3  1
20 4 2 3 4 2 2 5 3 4  3  1  4  2  4  1  3  3  2  1
21 2 3 1 1 4 1 2 2 1  1  3  1  3  4  2  3  4  1  2  3
22 2 4 2 1 5 2 2 3 1  2  4  2  4  5  3  4  5  2  3  4  1
23 3 1 3 3 3 2 4 2 3  2  2  3  1  2  2  1  2  1  1  2  2  3
24 1 4 3 2 6 3 3 2 1  2  5  3  5  4  4  3  4  3  4  5  2  1  4
25 1 3 4 3 5 3 4 1 2  2  4  4  4  3  4  2  3  2  3  4  3  2  3  1
```

Pour les données taxonomiques, on utilise la racine de la longueur du plus court chemin d'une espèce à un autre.




```
data(taxo.eg)
w <- taxo.eg[[1]]
names(w)
```

```
[1] "genre"   "famille" "ordre"
```

```
row.names(w) <- sub("esp", "E", row.names(w))
w1 <- as.taxo(w)
wphy <- taxo2phylog(w1)
plot(wphy, cleaves = 2, cnod = 2)
dist.taxo(w1)^2
```

```
    E3 E1 E2 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14
E1   2
E2   2  2
E4   2  2  2
E5   2  2  2  2
E6   2  2  2  2  2
E7   2  2  2  2  2  2
E8   8  8  8  8  8  8  8
E9   8  8  8  8  8  8  8  4
E10  8  8  8  8  8  8  8  6  6
E11  8  8  8  8  8  8  8  6  6   4
E12  8  8  8  8  8  8  8  6  6   6   6
E13  8  8  8  8  8  8  8  6  6   6   6   4
E14  8  8  8  8  8  8  8  6  6   6   6   6   6
E15  8  8  8  8  8  8  8  6  6   6   6   6   6   2
```

Les distances phylogénétiques ont été abordées dans la fiche `cssb3` de la présente série. Chaque discipline utilise ou crée des dissimilarités, par exemple l'analyse des données textuelles :

`http://www.cavi.univ-paris3.fr/lexicometrica/jadt/`

On comprend que le passage par une matrice de dissimilarité, est un outil puissant pour associer des informations de type varié.

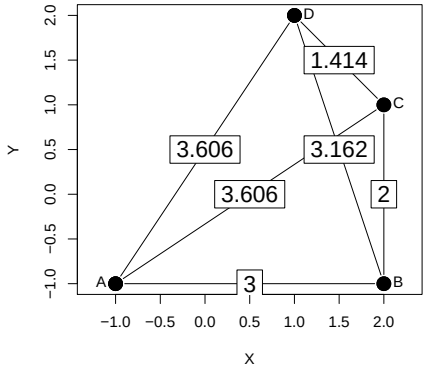
3 Propriétés des dissimilarités

3.1 Dissimilarités euclidiennes

La question est essentielle. Soient 4 points du plan :

A	-1	-1
B	2	-1
C	2	1
D	1	2

Représenter ces 4 points et les distances qui les séparent.



Oublier alors cette représentation et ne garder que les distances. Peut-on trouver 4 points sur un plan qui ont pour distances deux à deux les valeurs ci-dessous ?

```
      A     B     C
B 3.000
C 3.606 2.000
D 3.606 3.162 1.414
```

Sans doute ! Mais si on ne connaît que les distances, il peut ne pas y avoir de solutions. Le problème et la solution sont de Gower [7]. En général, la question est : étant donnée une matrice de distances $\mathbf{D} = [d_{ij}]$ entre n entités existe t-il n points M_i dans un espace euclidien de dimension p tels que :

$$d_{ij}^2 = \|M_i - M_j\|^2 \quad ?$$

Si c'est le cas, l'espace possède une base orthonormée dans laquelle le point M_i a des coordonnées qu'on peut mettre sur la ligne i d'un tableau $\mathbf{X}$. La matrice $\mathbf{X}\mathbf{X}^t$ contient alors les produits scalaires $\langle M_i \mid M_j \rangle$. Or :

$$d_{ij}^2 = \|M_i - M_j\|^2 = \|M_i\|^2 + \|M_j\|^2 - 2\langle M_i \mid M_j \rangle$$

Sans perte de généralités on peut supposer que le centre de gravité du nuage est à l'origine (sinon on le translate) et donc que $\sum_{i=1}^n M_i = 0$. On considère alors la matrice des valeurs $\mathbf{H} = \left[-\frac{1}{2}d_{ij}^2\right]$. On calcule la moyenne par lignes m_i, la moyenne par colonne m_j et la moyenne générale m :

$$m_i = -\frac{1}{2}\|M_i\|^2 - \frac{1}{2n}\sum_{j=1}^n \|M_j\|^2 \quad m_j = -\frac{1}{2}\|M_j\|^2 - \frac{1}{2n}\sum_{i=1}^n \|M_i\|^2$$

$$m = -\frac{1}{n}\sum_{i=1}^n \|M_i\|^2$$

La matrice $\mathbf{H}$ centrée par ligne et par colonne vaut alors :

$$\mathbf{H}_{\bullet\bullet} = \left[-\frac{1}{2}d_{ij}^2 - m_i - m_j + m\right] = \mathbf{X}\mathbf{X}^t$$

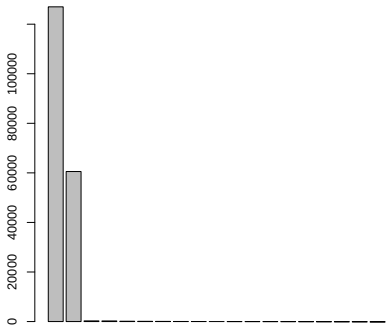
Toutes ses valeurs propres sont positives ou nulles. Réciproquement si la matrice des carrés des distances doublement centrées a toutes ses valeurs positives ou nulles :

$$\mathbf{H}_{\bullet\bullet} = \left[-\frac{1}{2}d_{ij}^2\right]_{\bullet\bullet} = \mathbf{U}\Lambda\mathbf{U}^t = \mathbf{U}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{U}^t = \mathbf{X}\mathbf{X}^t$$

Les carrés des distances entre les lignes de $\mathbf{X}$ sont les carrés des distances d'origine. Donc pour une matrice de distance $\mathbf{D}$ quelconque de deux choses l'une : - ou bien $\left[-\frac{1}{2}d_{ij}^2\right]_{\bullet\bullet}$ qui est symétrique a toutes ses valeurs propres positives ou nulles. On dit qu'elle est *euclidienne*. La matrice $\mathbf{X} = \mathbf{U}\Lambda^{\frac{1}{2}}$ donne les coordonnées d'un nuage de points dont les distances sont celles de départ. $\mathbf{X}$ est dite *représentation euclidienne* de $\mathbf{D}$. Ces coordonnées sont rangées par ordre décroissant de variance et si on utilise les premières pour voir une projection du nuage sur ses axes principaux on dit qu'on fait une *Analyse en coordonnées principales (PCOA)*. - ou bien $\left[-\frac{1}{2}d_{ij}^2\right]_{\bullet\bullet}$ qui est symétrique a des valeurs propres négatives. On dit qu'elle n'est pas *euclidienne*. La représentation euclidienne n'existe pas et la PCOA n'a pas lieu d'être. En pratique on fait souvent "comme si" en ignorant le problème des valeurs propres négatives. Gower et Legendre [8] ont étudié le caractère euclidien ou non de nombreuses distances et on sait souvent si on choisit une méthode de calcul qui donnera ou non une matrice de distances euclidiennes. Par exemple toutes les distances basées sur des indices de similarité sont euclidiennes alors que les distances les plus utilisées en génétique ne le sont pas.

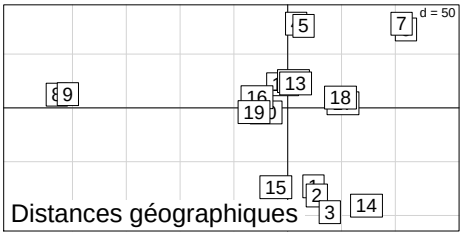
```
data(yanomama)
geo <- as.dist(yanomama$geo)
is.euclid(geo, T, T)
```

```
 [1]  1.270026e+05  6.054361e+04  2.596671e+02  2.420916e+02  1.291763e+02
 [6]  9.500950e+01  6.269167e+01  3.382556e+01  2.016325e+01 -2.469270e-13
[11] -1.318136e+01 -2.167582e+01 -2.942247e+01 -4.815153e+01 -7.937122e+01
[16] -1.178871e+02 -1.489993e+02 -1.755983e+02 -2.411742e+02
[1] FALSE
```



Cette matrice de dissimilarités n'est pas euclidienne pour des raisons simples d'arrondis impératives lors de l'édition. Il s'agit de distances à vol d'oiseau, donc euclidienne par définition. On corrige le problème par :

```
geo <- quasieuclid(geo)
s.label(dudi.pco(geo, scan = F)$li, sub = "Distances géographiques",
    csub = 2, clab = 1.5)
```



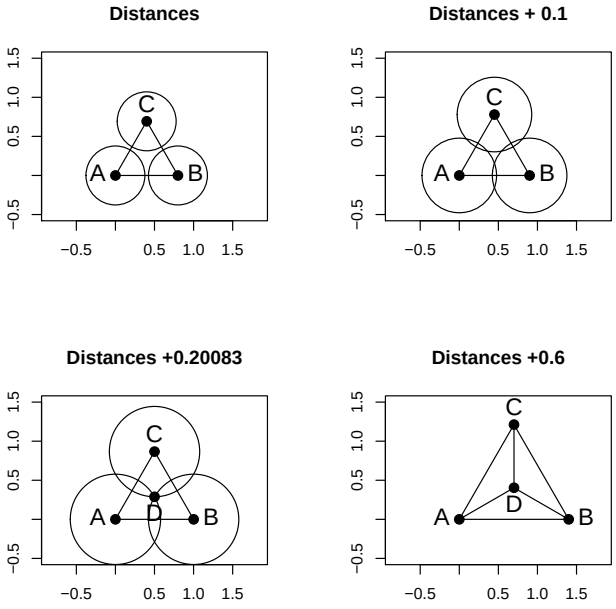
3.2 Que faire des distances non euclidiennes

Deux résultats fondamentaux confortent l'intérêt des distances euclidiennes.

Quand une matrice de distances n'est pas euclidienne, on sait calculer [3] la plus petite constante c qui assure que la distance $d_{ij} + c \quad \forall i \neq j$ est euclidienne.

L'exemple est décrit p. 435 dans l'ouvrage de référence de P. et L. Legendre [15]. On considère quatre points et les distances :

```
dmat
      A     B     C
B 0.800
C 0.800 0.800
D 0.377 0.377 0.377
```



Il est impossible de construire une figure qui respecte ces distances. Si on ajoute 0.1 la figure grossit et on pense que c'est la voie d'une solution car la petite distance augmente relativement plus vite que la grande.

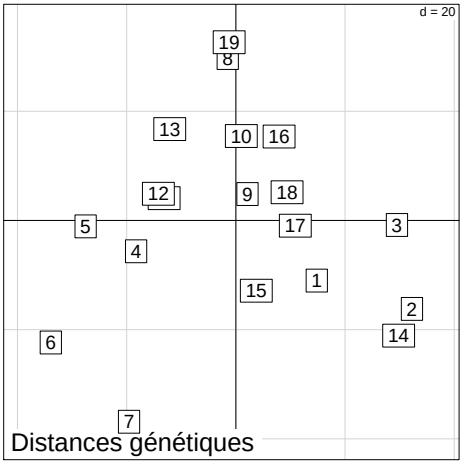
Il existe une valeur unique (0.20083) qui donne les quatre points dans un plan. Ensuite le quatrième point monte dans la troisième dimension et la figure n'est plus plane mais la configuration encore euclidienne.

La fonction `cailliez` donne la constante avec la solution décrite dans [3].

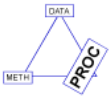
```
gen <- as.dist(yanomama$gen)
is.euclid(gen)
```

```
[1] FALSE
```

```
gen <- cailliez(gen)
s.label(dudi.pco(gen, scan = F)$li, sub = "Distances génétiques",
    csub = 2, clab = 1.5)
```



Quand une matrice de distances n'est pas euclidienne, on sait aussi calculer [16] la plus petite constante c qui assure que la distance $\sqrt{d_{ij}^2 + c} \quad \forall i \neq j$ est euclidienne.




```
ant <- as.dist(yanomama$ant)
is.euclid(ant)
```

```
[1] FALSE
```

```
ant <- lingoes(ant)
s.label(dudi.pco(ant, scan = F)$li, sub = "Distances anthropométriques",
    csub = 2, clab = 1.5)
```

On voit bien la question ainsi posée. Comment mesurer la ressemblance entre ces typologies faites par ces différentes distances ? La question était directement posée par R. Spielman, l'auteur des données [31]. On trouvera un exemple de même type dans un triplet spatial-environnemental et génétique (typologie de sites) avec :


```
data(butterfly)
```

3.3 Dissimilarités ultramétriques

Dans la classe des dissimilarités euclidiennes, un sous-ensemble de grand intérêt écologique (voir ci-après) est celui des distances ultramétriques. Les dissimilarités peuvent être métrique :

$$d_{ac} \leq d_{ab} + d_{bc}$$

Les distances euclidiennes sont toujours métriques. L'inégalité précédente est dite inégalité triangulaire. Une distance euclidienne peut en plus vérifier l'inégalité triangulaire ultramétrique :

$$d_{ac} \leq \max(d_{ab}, d_{bc})$$

La seconde est plus forte que la première. Les distances euclidiennes sont celles des nuages de points. Les distances ultramétriques sont celles des arbres de classification. On aura des détails dans :

`http://pbil.univ-lyon1.fr/R/stage/stage7.pdf`

Il suffit ici de savoir qu'une dissimilarités par le biais des méthodes de classification hiérarchique ascendante (CHA) induit une hiérarchie de partitions valuée qu'on voit par un dendrogramme. Les dendrogrammes sont des arbres (graphes sans cycles) dont toutes les feuilles sont à la même distance de la racine. Les phylogénies historiques et les taxonomies en sont (autant dire que ce sont des inventions de biologistes !).

Inversement si on a un dendrogramme, la distance entre feuilles définies par la hauteur du premier ancêtre commun est ultramétrique. En fait trouver un dendrogramme, c'est trouver une hiérarchie de partition valuée, c'est trouver une distance ultramétrique.

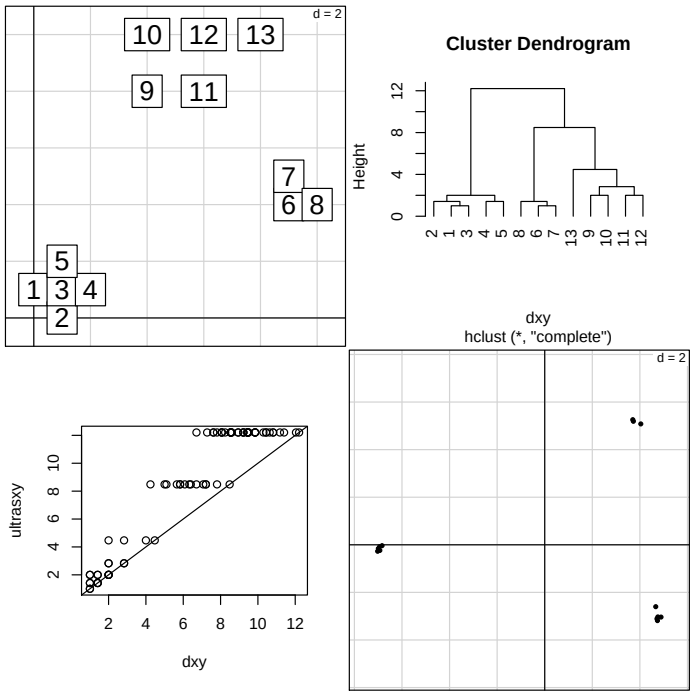
```
library(clue)
x = c(0, 1, 1, 2, 1, 9, 9, 10, 4, 4, 6, 6, 8)
y = c(1, 0, 1, 1, 2, 4, 5, 4, 8, 10, 8, 10, 10)
xy = cbind.data.frame(x, y)
par(mfrow = c(2, 2))
s.label(xy, clab = 2)
dxy <- dist(xy)
print(dxy, dig = 2)
```

```
      1    2    3    4    5    6    7    8    9   10   11   12
2   1.4
3   1.0  1.0
4   2.0  1.4  1.0
5   1.4  2.0  1.0  1.4
6   9.5  8.9  8.5  7.6  8.2
7   9.8  9.4  8.9  8.1  8.5  1.0
8  10.4  9.8  9.5  8.5  9.2  1.0  1.4
9   8.1  8.5  7.6  7.3  6.7  6.4  5.8  7.2
10  9.8 10.4  9.5  9.2  8.5  7.8  7.1  8.5  2.0
11  9.2  9.4  8.6  8.1  7.8  5.0  4.2  5.7  2.0  2.8
12 10.8 11.2 10.3  9.8  9.4  6.7  5.8  7.2  2.8  2.0  2.0
13 12.0 12.2 11.4 10.8 10.6  6.1  5.1  6.3  4.5  4.0  2.8  2.0
```

```
hsxy <- hclust(dxy, "complete")
plot(hsxy, hang = -1)
ultrasxy <- cl_ultrametric(hsxy)
print(ultrasxy, dig = 2)
```

```
Dissimilarities using Ultrametric distances:
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
 [1,]  1.414214
 [2,]  1.000000  1.414214
 [3,]  2.000000  2.000000  2.000000
 [4,]  2.000000  2.000000  2.000000  1.414214
 [5,] 12.206556 12.206556 12.206556 12.206556 12.206556
 [6,] 12.206556 12.206556 12.206556 12.206556 12.206556  1.000000
 [7,] 12.206556 12.206556 12.206556 12.206556 12.206556  1.414214  1.414214
 [8,] 12.206556 12.206556 12.206556 12.206556 12.206556  8.485281  8.485281
 [9,] 12.206556 12.206556 12.206556 12.206556 12.206556  8.485281  8.485281
[10,] 12.206556 12.206556 12.206556 12.206556 12.206556  8.485281  8.485281
[11,] 12.206556 12.206556 12.206556 12.206556 12.206556  8.485281  8.485281
[12,] 12.206556 12.206556 12.206556 12.206556 12.206556  8.485281  8.485281
      [,8]      [,9]      [,10]     [,11]     [,12]
 [1,]
 [2,]
 [3,]
 [4,]
 [5,]
 [6,]
 [7,]
 [8,]  8.485281
 [9,]  8.485281  2.000000
[10,]  8.485281  2.828427  2.828427
[11,]  8.485281  2.828427  2.828427  2.000000
[12,]  8.485281  4.472136  4.472136  4.472136  4.472136
```

```
plot(dxy, ultrasxy)
abline(c(0, 1))
z <- dudi.pco(ultrasxy, scan = F)$li
z$A1 <- jitter(z$A1, amount = 0.2)
z$A2 <- jitter(as.numeric(z$A2), amount = 0.2)
s.label(z, clab = 0)
```



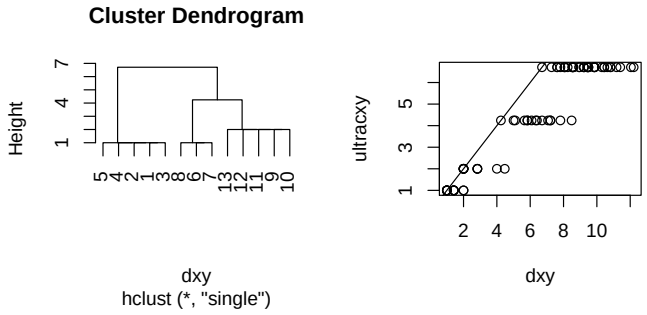
13 points sont dans un plan. On connait leurs coordonnées. La matrice de distances euclidiennes (au sens strict) :

$$d_{ab} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

est envoyé dans une classification hiérarchique (se méfier, il y en a beaucoup !) qui donne un dendrogramme (classe `hclust`). La classification donne une distance ultramérique, donc euclidienne et la représentation euclidienne de cette nouvelle distance est possible. Les CAH sont des pratiques qui concentrent l'aspect classification alors que la PCO concentre l'aspect ordination. On mélange souvent les deux.

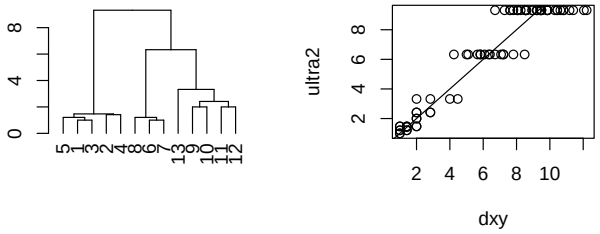
Noter que la nouvelle distance est supérieure ou égale à la première pour tous les couples (on dit qu'elle dominante) C'est la plus petite qui a cette propriété. On dit que c'est l'ultramétrique dominante minimale.

```
par(mfrow = c(1, 2))
hcxy <- hclust(dxy, "single")
plot(hcxy, hang = -1)
ultracxy <- cl_ultrametric(hcxy)
plot(dxy, ultracxy)
abline(c(0, 1))
```



Avec cette méthode, la nouvelle distance est toujours plus petite. On dit qu'elle est sous-dominante. Mais c'est la plus grande qui a cette propriété. On dit que c'est l'ultramétrique sous-dominante maximale. Dans la librairie de Kurt Hornik, on trouve une fonction qui donne l'ultramétrique la plus proche, au sens des moindres carrés (critère L2) ou des moindres différences en valeurs absolues (critère L1) :

```
par(mfrow = c(1, 2))
ultra2 <- ls_fit_ultrametric(dxy)
plot(ultra2)
plot(dxy, ultra2)
abline(c(0, 1))
```



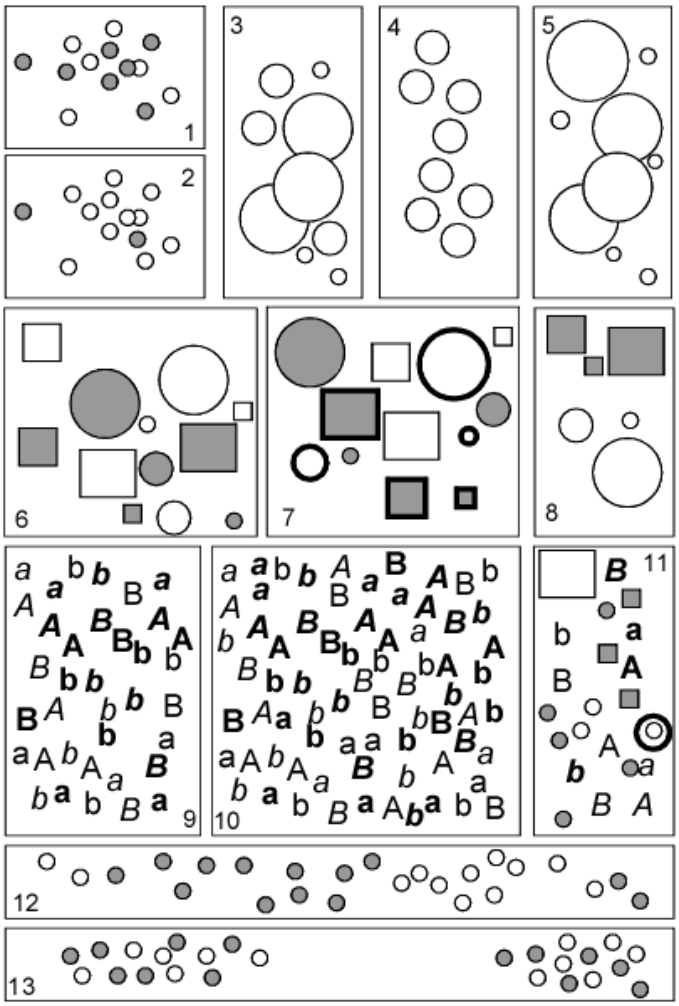
Pour en savoir plus, voir l'ouvrage de référence pour les écologues [14] ou les documents de cours [29] [1] [13] disponibles sur le réseau. [1].

4 Des espèces aux communautés

4.1 La question de la mesure de la diversité

Elle se pose dès que l'objet d'étude est une collection, collection d'arbres dans un forêt, d'individus dans une communauté, collection de gènes, de séquences, collection d'entités non toutes identiques.

[1] Accès :
`www.imep-cnrs.com/mroux/algoclas.pdf`
`www.iup.univ-avignon.fr/Etudes/Master/TAIM//SupportsCours/Classification/coursClassification.pdf`
`ses.enst.fr/lebart/DEA/CLassif.pdf`



La figure est explicite. Que vaut la variabilité des éléments d'une collection pour une variable binaire (1 est-il plus divers que 2 ?) pour une variable quantitative (la diversité est nulle en 4 mais, entre 3 et 5, lequel est-il le plus divers ?) pour une variable quantitative et deux variables qualitatives (6) pour un ensemble de variables complexes (7) pour des variables redondantes (8) pour des proportions égales mais des effectifs différents (9-10) pour des choses non comparables (11) ou des objets spatialisés (12-13). La question n'est pas simple et la bibliographie associée est énorme.

Parmi les bases solides on retient que la diversité est une qualité définie pour une distribution d'objets rangés par catégories. On en a d'abord parlé pour les individus d'une communautés rangés par espèce (diversité taxonomique) ou pour les individus d'une société rangés par catégories de ressources (inégalité sociale). Le nombre de catégories (richesse) est l'indice le plus immédiat mais quand les classes sont artificielles c'est le mode de distribution entre classes qui doit être décrit :

> Plus particulièrement, le fait d'augmenter uniformément de cinq semaines le nombre minimal de semaines d'admissibilité exacerberait

> le problème de l'inégalité des revenus entre les hommes et entre les ménages au Canada. L'indice de Gini pour les hommes passerait de 0,448 à 0,459 pour le revenu avant impôt, ...[2]

Le coefficient de Gini [6] est celui de Simpson [30] en écologie et de Gini-Simpson pour les puristes. La diversité est une valeur pour les écologues, l'inégalité est une plaie pour les sociologues. Patil et Taillie ont synthétiser l'état de l'art dans les années 80 en parlant de fonction de la rareté.

Soit un relevé écologique contenant s espèces avec les proportions

$$\mathbf{p} = (p_1, \cdots, p_i, \cdots, p_s)$$

. Soit R une fonction caractéristique de la rareté de l'espèce du type :

$$\mathrm{R} : [0, 1] \rightarrow \mathbb{R}^+$$

La diversité est la moyenne de la rareté des espèces du relevé. La rareté diminue quand p augmente et vaut 0 quand p vaut 1. On peut comprendre ainsi :

La richesse

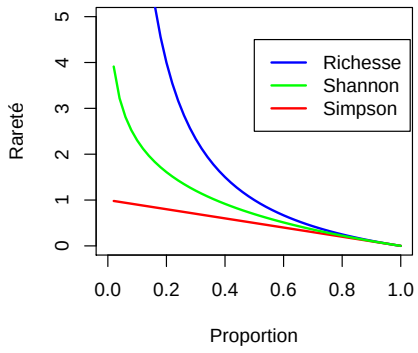
$$\mathrm{R}(p) = \frac{1}{p} - 1 \Rightarrow \sum_{i=1}^{s} p_i \left(\frac{1}{p_i} - 1 \right) = s - 1$$

L'indice de Simpson

$$\mathrm{R}(p) = (1 - p) \Rightarrow \sum_{i=1}^{s} p_i (1 - p_i) = 1 - \sum_{i=1}^{s} p_i^2$$

L'indice de Shannon

$$\mathrm{R}(p) = -\ln(p) \Rightarrow \sum_{i=1}^{s} p_i \ln(p_i)$$



[2] L'assurance-chômage et la redistribution du revenu - Août 1995 sur `http://www11.hrsdc.gc.ca/`

Définir la diversité par la richesse, c'est donner une préférence absolue aux espèces rares (valeur suprême de la coche naturaliste). Le formalisme de Hill [11] a enrichi la collection d'indices à partir de

$$N_a = \left(p_1^a + p_2^a + \cdots + p_s^a\right)^{\frac{1}{(1-a)}}$$

qui produit :

1 $a = -\infty$ $N_{-\infty}$ est l'inverse de la proportion de l'espèce la plus rare

2 $a = 0$ **Richesse** N_0 est le nombre d'espèces

3 $a = \frac{1}{2}$

$$N_{\frac{1}{2}} = \left(\sqrt{p_1} + \sqrt{p_2} + ...\sqrt{p_s}\right)^2$$

4 $a = 1$ **Shannon**

$$N_1 = \lim_{a \to 1} (N_a) = \exp\left(-\sum_{i=1}^{s} p_i \ln(p_i)\right)$$

5 $a = 2$ **Simpson inverse**

$$N_2 = \frac{1}{\sum_{i=1}^{s} p_i^2}$$

6 $a = \infty$ N_∞ est l'inverse de la proportion de l'espèce la plus abondante

La critique radicale d'Hurlbert [12] et une multitude d'approches voisines ont montré les difficultés d'installer des méthodes canoniques. Par exemple :

Rencontres *the proportion of potential interindividual encounters which is interspecific (as opposed to intraspecific), assuming avery individual in the collection can encounter all other individuals*

$$\Delta_1 = \left[\frac{N}{N-1}\right]\left[1 - \sum\nolimits_{i=1}^{s}\left(\frac{N_i}{N}\right)^2\right]$$

Espérance *the expected number of species in a sample of n individuals selected at random from a collection containing N individuals, S species, and N_i individuals in the ith species*

$$E(S_n) = \sum\nolimits_{i=1}^{s}\left(1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}}\right)$$

[12] $\Delta_1 = \left[\frac{N}{N-1}\right]\left[1 - \sum_{i=1}^{s} p_i^2\right]$ $\Delta_2 = \left[1 - \sum_{i=1}^{s} p_i^2\right]$ $\Delta_3 = \frac{1}{\sum_{i=1}^{s} p_i^2}$

$$\Delta_4 = \frac{1 - \sum_{i=1}^{s} p_i^2}{\sum_{i=1}^{s} p_i^2 - \frac{1}{N}} \quad \Delta_5 = \frac{1 - \sum_{i=1}^{s} p_i^2}{\sum_{i=1}^{s} p_i^2}$$

[17]

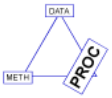
$$D = \sum\nolimits_{i=1}^{s}\left(\frac{n_i(n_i-1)}{N(N-1)}\right) \approx \sum\nolimits_{i=1}^{s} p_i^2$$

[19]

$$E = \frac{1 - \sqrt{\sum_{i=1}^{s} p_i^2}}{1 - \frac{1}{\sqrt{s}}}$$

[25]

$$D = \frac{1 - \sqrt{\sum_{i=1}^{s} p_i^2}}{1 - \frac{1}{\sqrt{N}}}$$



On sait depuis longtemps que ces quantités sont très corrélées et, globalement, explore le même point de vue.

Une petite fonction pour choisir des couleurs :

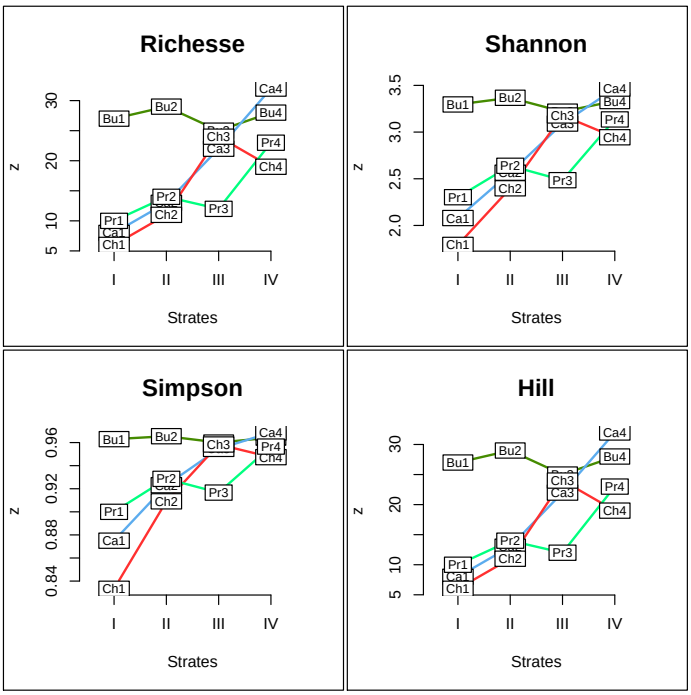
```
editcolor <- function() {
    w <- matrix(0, 21, 21)
    par(mar = c(0.1, 0.1, 0.1, 0.1))
    plot(c(1, 22), c(1, 22), type = "n")
    rect(col(w), row(w), col(w) + 1, row(w) + 1, col = couleurs)
    w <- as.numeric(locator(1))
    w <- floor(w)
    num <- (w[1] - 1) * 21 + w[2]
    return(couleurs[num])
}
```

Implanter une petite fonction pour représenter un vecteur dans le plan d'expérience qui nous concerne :

```
foo1 <- function(z = runif(16), sub = "") {
    sn1 <- c("I", "II", "III", "IV")
    col1 <- c("chartreuse4", "steelblue2", "firebrick1", "springgreen")
    reg.num <- rep(1:4, rep(4, 4))
    str.num <- rep(1:4, 4)
    plot(str.num, z, type = "n", xlim = c(0.5, 4.5), ax = F, xlab = "Strates")
    axis(1, 1:4, sn1)
    axis(2)
    for (k in 1:4) lines(1:4, z[reg.num == k], type = "b", col = col1[k],
        lwd = 2)
    s.label(cbind.data.frame(str.num, z), lab = names(ecomor$habitat),
        add.p = T)
    title(main = sub, cex.main = 1.5)
}
```

Représenter les mesures classiques de diversité :

```
library(vegan)
par(mfrow = c(2, 2))
foo1(apply(ecomor$habitat > 0, 2, sum), sub = "Richesse")
foo1(diversity(as.matrix(ecomor$habitat), index = "shannon", MARGIN = 2),
    "Shannon")
foo1(diversity(as.matrix(ecomor$habitat), index = "simpson", MARGIN = 2),
    "Simpson")
foo1(diversity(as.matrix(ecomor$habitat), index = "invsimpson",
    MARGIN = 2), "Hill")
```



4.2 Diversité et différences

Nous sommes à pied d'œuvre. Nous avons 16 habitats de 4 types dans 4 régions. Nous voulons comparer les habitats, comparer leur contenu, leurs propriétés. Nous voulons faire un bilan des ressemblances et des différences. Nous voulons comparer la variabilité et faire un bilan typologique. Nous voulons avoir plusieurs points de vue. Vaste programme.

Nous avons besoin d'une théorie générale pour repérer les possibles : C.R. Rao[26] [27] [28] nous la fournit. Les espèces sont différentes, plus ou moins. Les habitats sont diversifiés, plus ou moins en fonctions des différences des espèces qui y sont. Les habitats sont plus ou moins différents entre eux selon les différences qui existent entre les espèces de l'un et ceux de l'autre. S. Pavoine [, p. 46] montre la naissance de cette idée en écologie [10] où elle sera confidentielle jusqu'à sa mise en usage en biologie marine [32], puis en génétique (les espèces sont remplacées par des séquences nucléotidiques) [20] qui conduira aux travaux de Rao.

On peut utiliser, comme on voudra :

- le langage écologique : un communauté est une collection pondérée d'espèces ;
- le langage génétique : une population est une collection pondérée de gènes ;
- le langage statistique : une collection est une distribution pondérée de catégories.

Les catégories de la collection sont au nombre de s. Une collection est un élément de l'ensemble :

$$\mathfrak{P} = \{\mathbf{p} = (p_1, \cdots, p_i, \cdots, p_s) \diagup \sum_{i=1}^{s} p_i = 1\}$$

Les catégories montrent des différences consignées dans une matrice de distances à s lignes et s colonnes :

$$\mathbf{D} = [d_{ij}] \quad 1 \leq i \leq s \quad 1 \leq j \leq s$$

La diversité de $\mathbf{p}$ du point de vue de $\mathbf{D}$ est définie par la diversité quadratique ou encore entropie quadratique :

$$div_{\mathbf{D}}(\mathbf{p}) = \frac{1}{2}\sum_{i,j} p_i p_j d_{ij}^2$$

La seule contrainte imposée pour que ce calcul ait un sens est que le mélange de deux collections soit plus diverses que chacune des deux, soit, pour deux nombres positifs quelconques :

$$\alpha + \beta = 1 \Rightarrow div_{\mathbf{D}}(\alpha\mathbf{p} + \beta\mathbf{q}) \geq \alpha div_{\mathbf{D}}(\mathbf{p}) + \beta div_{\mathbf{D}}(\mathbf{q})$$

Si l en est ainsi, non seulement la diversité a un sens mais, l'augmentation de la diversité par mélange est une mesure de distances entre collections :

$$\delta_{\mathbf{D}}(\mathbf{p}, \mathbf{q}) = \sqrt{2}\sqrt{2div_{\mathbf{D}}(\frac{\mathbf{p}+\mathbf{q}}{2}) - div_{\mathbf{D}}(\mathbf{p}) - div_{\mathbf{D}}(\mathbf{q})}$$

Ces quelques éléments d'introduction suffisent pour comprendre qu'il ne s'agit pas d'un bricolage mais d'une structure de raisonnement très ouverte. En effet, supposons que toutes les espèces soient également différentes entre elles :

$$i \neq j \Rightarrow d_{ij} = \sqrt{2}$$

Alors la diversité d'une collection est :

$$div(\mathbf{p}) = \sum_{i=1}^{i=s}\sum_{i=1}^{i=s} p_i p_j \delta_{ij} = \sum_{i=1}^{i=s} p_i(1-p_i) = 1 - \sum_{i=1}^{i=s} p_i^2$$

Et la distance entre deux collections est :

$$\delta^2(\mathbf{p}, \mathbf{q}) = 4(1 - \sum_{i=1}^{i=s}\frac{(p_i+q_i)^2}{4}) - 2(1 - \sum_{i=1}^{i=s} p_i^2) - 2(1 - \sum_{i=1}^{i=s} q_i^2) = \sum_{i=1}^{i=s}(p_i - q_i)^2$$

Ce qui signifie que penser chaque espèce différente de chacune des autres exactement de la même manière, c'est mesurer la diversité par l'indice de Simpson et assure la typologie des sites par une ACP centrée.

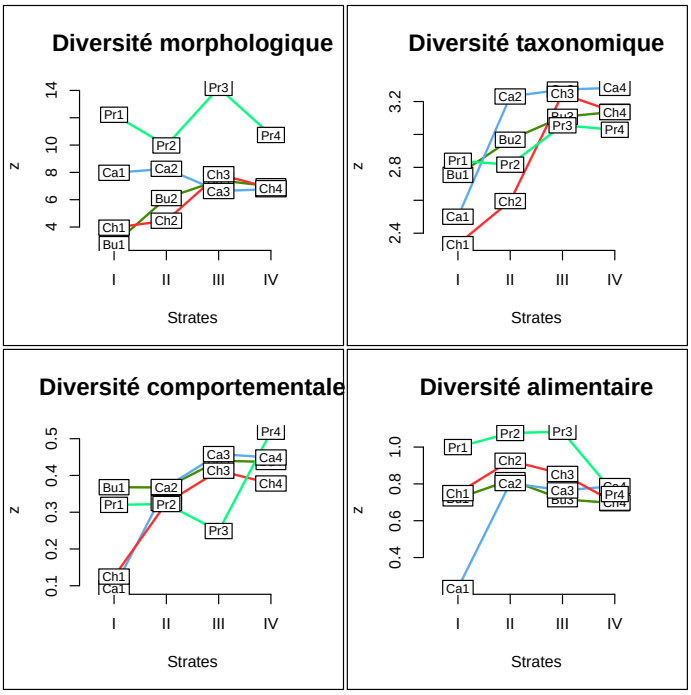
```
d0 <- matrix(sqrt(2), nrow(ecomor$habitat), nrow(ecomor$habitat))
diag(d0) <- 0
d0 <- as.dist(d0)
quad1 <- divc(ecomor$habitat, d0)[, 1]
sim1 <- diversity(as.matrix(ecomor$habitat), index = "simpson",
    MARGIN = 2)
all(abs(quad1 - sim1) < 1e-14)
```

```
[1] TRUE
```

La différence est vite compréhensible. Dans l'indice classique, on ne fait aucune différence dans les différences entre espèces : un lion et deux impalas, c'est la même chose qu'un gnou et deux zèbres ou qu'un éland et deux koudous.

On peut maintenant introduire des points de vue très différents :

```
par(mfrow = c(2, 2))
d0 <- dist.quant(ecomor$morpho, 3)
q0 <- divc(ecomor$habitat, d0)[, 1]
foo1(q0, sub = "Diversité morphologique")
d0 <- dist.taxo(ecomor$taxo)
q0 <- divc(ecomor$habitat, d0)[, 1]
foo1(q0, sub = "Diversité taxonomique")
forsub <- data.frame(t(apply(ecomor$forsub, 1, function(x) x/sum(x))))
d0 <- dist(forsub)
q0 <- divc(ecomor$habitat, d0)[, 1]
foo1(q0, sub = "Diversité comportementale")
d0 <- dist(ecomor$diet)
q0 <- divc(ecomor$habitat, d0)[, 1]
foo1(q0, sub = "Diversité alimentaire")
```



Un outil assainir le débat sur la biodiversité ? C'est possible.

4.3 Typologie sur différences

La manipulation simultanée des variabilités intra communautés (diversité) et inter communautés (typologie) est un élément fondamental de cette approche. On trouvera dans la thèse de S. Pavoine[3] une approche complète de la décomposition de la diversité dans des plans d'observations complexes, croisés ou hiérarchiques. Cette introduction s'achèvera sur deux questions importantes traitées dans ce travail.

La première touche à l'usage de la distance induite sur les communautés par une distance observée sur les espèces. L'exemple a été choisie parce que l'analyse du tableau espèces-sites n'a aucun sens sauf à enfoncer une porte ouverte : les cortèges faunistiques n'ont aucun points communs d'un continent à l'autre. Mais

[3] `http://pbil.univ-lyon1.fr/R/liens/these_sp.pdf`

en passant par des traits biologiques, on retrouve la possibilité de comparer ce qui ne l'était pas par la seule taxonomie.

Considérons le lieu d'alimentation des espèces d'oiseaux. `forsub` signifie *foraging substrate*et contient des variables binaires (1 = oui, 0 = non) passées en pourcentage par lignes :

foliage alimentation dans le feuillage

ground alimentation au sol

twig alimentation sur les rameaux

bush alimentation dans les buissons

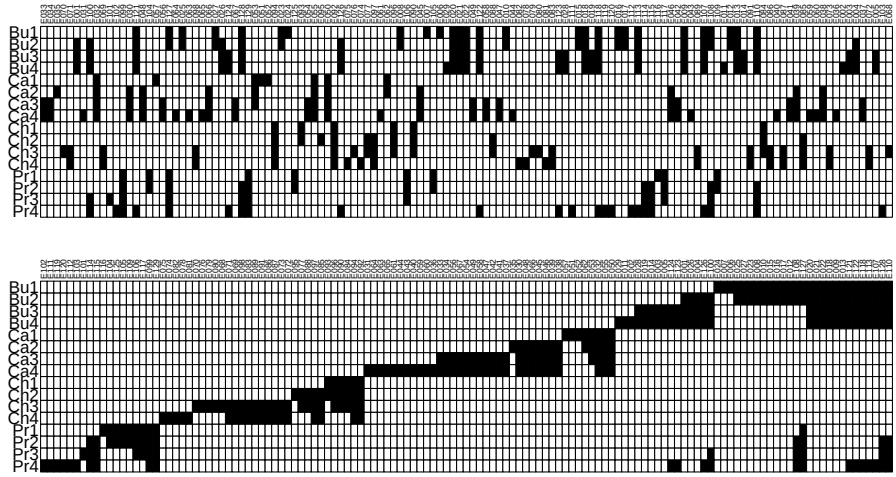
trunk alimentation sur les troncs

aerial alimentation aérienne

Les lieux d'alimentation font une typologie d'espèces, les espèces font une typologie d'habitats : la question est comment faire une typologie d'habitats sur la base des comportements des espèces qu'on y trouve. Cette question est proprement écologique : expertiser un milieu par les espèces qui l'habitent est une activité quotidienne. Pourtant la solution statistique est récente[22].

Les espèces font une typologie d'habitats, triviale, dès qu'on la connaît :

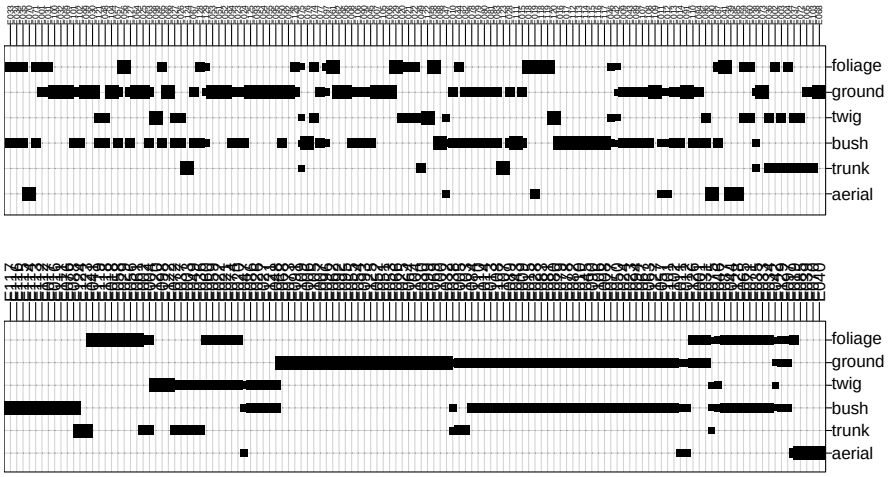
```
par(mfrow = c(2, 1))
wt <- as.data.frame(t(ecomor$habitat))
table.paint(wt, clabel.r = 1, clabel.c = 0.5, cleg = 0)
a <- -ecomor$habitat
a <- t(apply(a, 1, cumsum))
w <- paste(a[, 1], a[, 2], a[, 3], a[, 4], a[, 5], a[, 6], a[, 7],
    a[, 8], a[, 9], a[, 10], a[, 11], a[, 12], a[, 13], a[, 14],
    a[, 15], a[, 16], sep = "")
table.paint(wt[, order(w)], clabel.r = 1, clabel.c = 0.5, cleg = 0)
```



Les colonnes du tableau habitats-espèces sont ordonnées, en haut, par la taxonomie. En bas, c'est exactement le même tableau à une permutation des colonnes près.

Les catégories de site d'alimentation font une typologie des espèces, simple avec trois grands types et des nuances. Voir le tableau suffit à comprendre son contenu :

```
par(mfrow = c(2, 1))
wt <- as.data.frame(t(forsub))
table.value(wt, clabel.r = 1, clabel.c = 0.5, cleg = 0, csi = 0.25)
hc1 <- hclust(dist(forsub))
table.value(wt[, hc1$order], clabel.c = 1, cleg = 0, csi = 0.25)
```



Mais deux tableaux simples ne rendent pas simple la question de la typologie des colonnes du premier par la distance induite sur les lignes par le second ! Un tableau et une matrice de distance s'associe dans une double PCO :

```
d0 <- dist(forsub)
dpco1 <- dpcoa(ecomor$habitat, d0, scan = F)
plot(dpco1)
```

L'essentiel est dans le fait que la distance entre espèces, quand elle est euclidienne, introduit une distance entre sites qui est aussi euclidienne. On peut alors faire une représentation euclidienne des espèces et des sites d'un seul coup. Ils sont en effet dans un même espace euclidien, parceque les espèces définissent cet espace par leur distance et parce que les sites sont des centres de gravité (distributions de fréquences sur les espèces). On représente ainsi la typologie des sites induite par la distances entre espèces. Un exemple d'utilisation en microbiologie est dans [5]. En fait, cette stratégie ici effleurée recouvre une très large gamme d'analyses connues (voir [22]).

```
par(mfrow = c(2, 2))
d1 <- dist.quant(ecomor$morpho, 3)
q0 <- dpcoa(ecomor$habitat, d1, scan = F)$l2[, 1]
foo1(q0, sub = "Typologie morphologique")
d2 <- dist.taxo(ecomor$taxo)
q0 <- dpcoa(ecomor$habitat, d2, scan = F)$l2[, 1]
foo1(q0, sub = "Typologie taxonomique")
d3 <- dist(forsub)
q0 <- dpcoa(ecomor$habitat, d3, scan = F)$l2[, 1]
foo1(q0, sub = "Typologie comportementale")
d4 <- dist.prop(ecomor$diet, 5)
q0 <- dpcoa(ecomor$habitat, d4, scan = F)$l2[, 1]
foo1(q0, sub = "Typologie alimentaire")
```

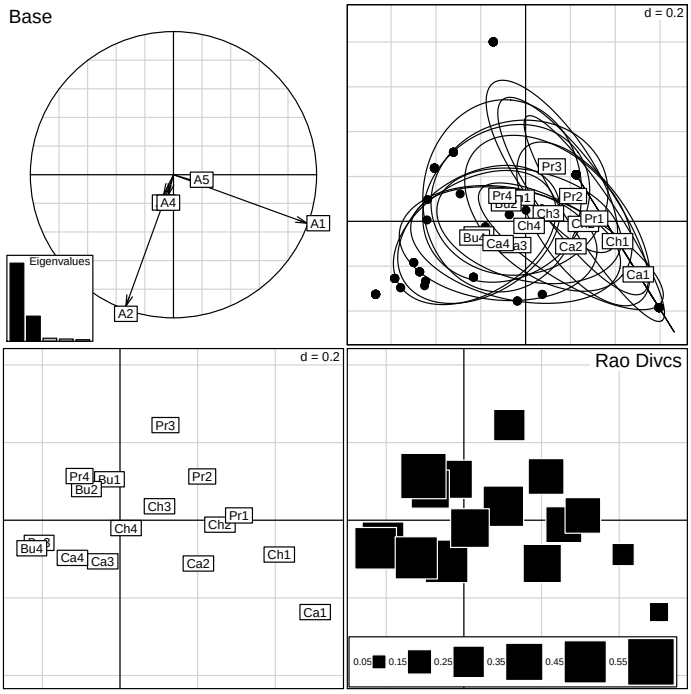
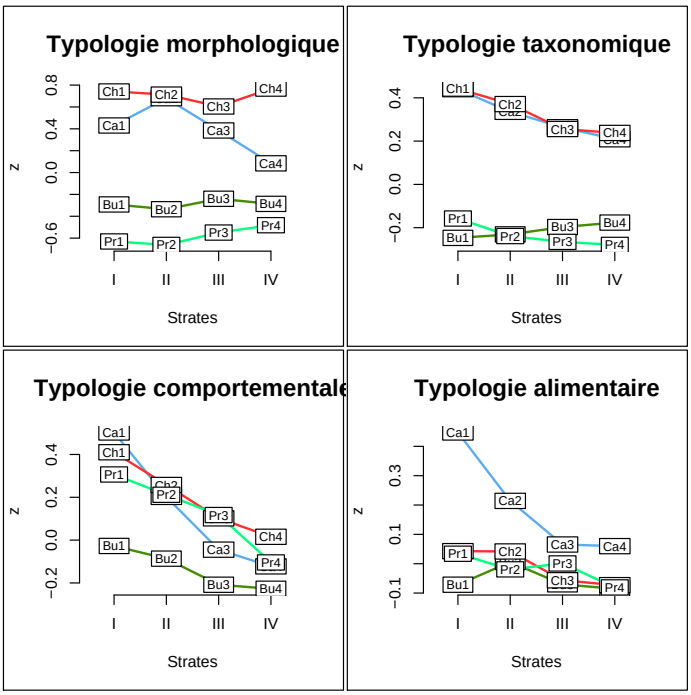
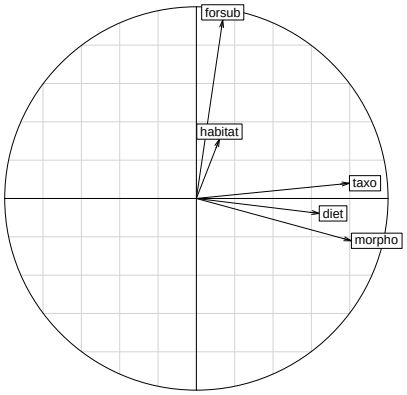


FIG. 2 – Analyse en coordonnées principales d'un tableau faunistique par rapport à une distance euclidienne entre espèces. En haut à gauche, graphe des valeurs propres (les barres noires indiquent le numéro des axes servant à la représentation) et projections des axes de la représentation euclidienne de la distance entre espèces sur les axes principaux de l'analyse. En haut à droite, projection sur les axes principaux des espèces (points noirs) et des relevés (étiquettes et ellipses de dispersion). En bas à gauche, nuage des relevés (recadrage du précédent). En bas à droite, représentation sur la carte de la diversité au sens de Rao des relevés. Le plan de la représentation euclidienne des distances induites entre relevés est sensiblement celui de la représentation euclidienne des distances entre espèces. La typologie induite est essentiellement intra-continent et se caractérise par l'augmentation de la diversité des lieux d'alimentation par simple augmentation de la complexité des architectures végétales. L'analyse de ce trait simple prouve surtout la pertinence de l'entreprise de comparaison de cortèges faunistiques complètement disjoints.



Cette figure est assez extraordinaire. On n'utilise de chaque analyse que le premier axe, qui est toujours clairement pertinent (vérifier sur les valeurs propres) et on représente la coordonnée factorielle des habitats. Ce qui frappe est la diversité des résultats. Chacun peut être discuté. La morphologie, profondément liée à la taxonomie, donne une séparation continentale. La convergence méditerranéenne est explicite pour les lieux d'alimentation alors que l'originalité de la Californie apparaît dans les régimes alimentaires. L'architecture végétale intervient surtout dans le troisième. Ce qui conduit à la classe des `kdist` :

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



On pourra alors aborder les tableaux de traits biologiques qui comportent un grand nombre de dimensions.

4.4 Maximiser la diversité

La seconde question abordée par S. Pavoine est ouverte dans [4]. Elle est posée par des biologistes qui ont souhaité avoir une normalisation des mesures de diversité. Si en multiplie par 2 une distance, la diversité quadratique est multipliée par 4 et les problèmes d'unité de mesure sont directement posés. On peut penser à normer par :

$$div_{\mathbf{D}}(\mathbf{p}) = \frac{1}{2}\sum_{i,j} p_i p_j d_{ij}^2 \Rightarrow div_{\mathbf{D}}^*(\mathbf{p}) = \frac{div_{\mathbf{D}}(\mathbf{p})}{max_{\mathbf{q}\in\mathfrak{P}}(div_{\mathbf{D}}(\mathbf{q}))}$$

Il faut alors savoir si ce maximum existe, le calculer éventuellement et observer pour quelles distributions il est obtenu. On tombe sur une difficulté sévère. Sur une seule variable quantitative, la diversité quadratique est exactement la variance :

$$div_{\mathbf{D}}(\mathbf{p}) = \frac{1}{2}\sum_{i,j} p_i p_j (x_i - x_j)^2 = \sum_i p_i (x_i - \overline{\mathbf{x}})^2$$

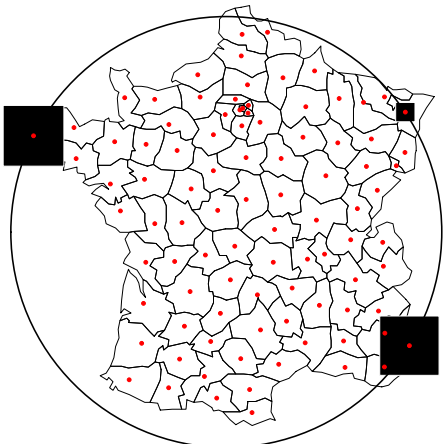
Pour maximiser cette quantité, il suffit de mettre 50% sur le plus petit et 50% sur le plus grand. Pour augmenter la diversité du point de vue de la taille, il suffirait d'enlever toutes les espèces sauf la plus petite et la plus grande. C'est mathématiquement indiscutable et écologiquement absurde. Par contre, si chaque espèce est également distante de toutes les autres, la diversité quadratique est l'indice de Simpson qui est maximum pour la distribution uniforme.

Pour une variable quantitative, on apprécie la variabilité en divisant l'écart-type par la moyenne. Mais cette pratique n'a plus de sens en dimension quelconque. Par contre l'effet persiste :

```
data(elec88)
par(mar = c(0.1, 0.1, 0.1, 0.1))
area.plot(elec88$area)
d0 <- dist(elec88$xy)
France.m <- divcmax(d0)
w0 <- France.m$vectors$num
v0 <- France.m$value
(1:94)[w0 > 0]
```

```
[1]  6 28 66
```

```
w1 = elec88$xy[c(6, 28, 66), ]
w.c = apply(w1 * w0[c(6, 28, 66)], 2, sum)
s.value(elec88$xy, w0, add.plot = TRUE, cleg = 0, csize = 2)
symbols(w.c[1], w.c[2], circles = sqrt(v0), inc = FALSE, add = TRUE,
    lwd = 2, col = "blue")
points(elec88$xy, pch = 20, col = "red")
```



Pour une distance euclidienne, les points sont dans un espace euclidien et il existe toujours une sphère plus petite qui contient tous les points. Les points de cette sphère ont seuls un poids non nuls pour maximiser la diversité quadratique. La fonction `divcmax` donne une solution numérique dans tous les cas. La figure montre comment répartir les français pour maximiser le carré moyen de la distance entre deux individus, ce qui expérimentalement est sans intérêt. La théorie dit que c'est un fait général et que l'entropie quadratique ne se maximise pas avec une signification écologique.

On montre alors [24] que ce défaut majeur disparaît avec les distances ultramétriques. En effet, dans ce cas, il existe un point centre d'une sphère sur laquelle on trouve toutes les observations. La distribution qui maximise la diversité ne montre alors aucun poids nul. Ce poids est une mesure de l'originalité de chaque élément[23]. On comprend facilement son fonctionnement. Reprenons les données morphométriques avec plus d'attention :

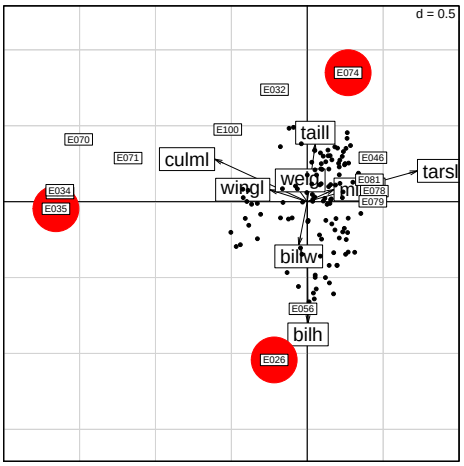
```
 morlog <- log(ecomor$morpho)
 round(apply(morlog, 2, var), dig = 3)
wingl taill culml  bilh billw tarsl midtl  weig
0.150 0.159 0.157 0.224 0.156 0.208 0.160 1.213
```

Toutes les variables ont des variances comparables : le poids fait exception. Or le poids est en g, équivalent d'un mm^3 en dimension, il aurait fallu travailler avec la racine cubique du poids, donc avec le tiers du logarithme, donc avec une variance 9 fois plus petite.

```
 morlog$weig = morlog$weig/3
 round(apply(morlog, 2, var), dig = 3)
wingl taill culml  bilh billw tarsl midtl  weig
0.150 0.159 0.157 0.224 0.156 0.208 0.160 0.135
 morlog <- morlog - apply(morlog, 1, mean)
```

On a débarrassé les données de l'effet taille et on ne discute maintenant que de la forme.

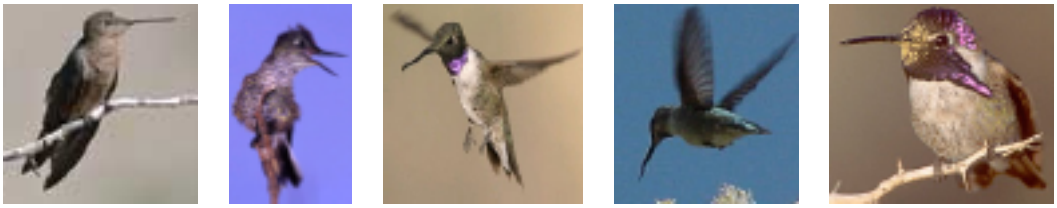
```
 w <- dudi.pca(morlog, scal = F, scan = F)
 par(mar = rep(0.1, 4))
 s.arrow(w$c1, xlim = c(-2, 1), clab = 1.5)
 points(w$li[c(3, 28, 49), 1], w$li[c(3, 28, 49), 2], pch = 19, cex = 8,
     col = "red")
 s.label(w$li, clab = 0, add.p = T)
 sel <- c("E033", "E034", "E035", "E070", "E071", "E100", "E032",
     "E074", "E046", "E081", "E078", "E079", "E056", "E026")
 s.label(w$li[sel, ], clab = 0.75, add.p = T)
```



Pour les points rouges, voir ci-dessous. Pour l'illustration :

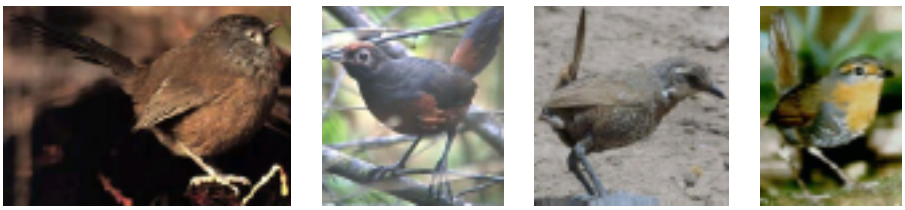
A gauche : tout dans le bec [4]

E070 E071 E033 E034 E035



A droite : tout dans les pattes [5]

E046 E078 E079 E081



[4]Sources :
E070 *Patagona gigas* http://www.arthurgrosset.com/sabirds/gianthummingbird.html
E071 *Sephaniodes sephaniodes* http://www.greglasley.net/gbfire.html
E033 *Archilochus alexandri* http://weaselhead.org/learn/birds_black-chinned_hummingbird.asp
E034 *Calypte anna* http://www.oceanoasis.org/fieldguide/caly-ann.html
E035 *Calypte costae* http://www.birdphotography.com/species/cohu.html

[5]Sources :
E046 *Chamaea fasciata* http://stockpix.com/stock/animals/birds/songbirds/wrens/4484.htm
E078 *Pteroptochos tarnii* http://www.avesdechile.cl/184.htm
E079 *Pteroptochos megapodius* http://www.camacdonald.com/birding/MoustachedTurca(NL).jpg
E081 *Scelorchilus rubecula* http://www.avesdechile.cl/174.htm

En haut : finesse du bec et de la queue [6]

E074

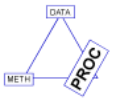
E032

E100

En bas : puissance du bec [7]

E056

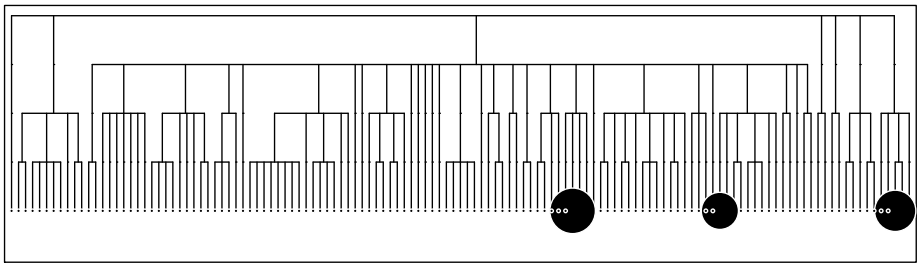
E026



Pour ordonner, la distance euclidienne s'impose. Pour mesurer la diversité d'une collection, c'est moins sûr :

```
deuc <- dist(morlog)
wtax <- divcmax(deuc)$vectors$num
phyt <- taxo2phylog(ecomor$taxo, add = F)
symbols.phylog(phyt, wtax, cleg = 0)
ecomor$labels[which(wtax > 0), ]
```

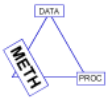
```
     latin                          abbr
E035 "Calypte costae"               "Cal|cos"
E026 "Pyrrhula pyrrhula"            "Pyr|pyr"
E074 "Sylviorthorhynchus desmursii" "Syl|des"
```



[6] Sources :
E074 *Sylviorthorhynchus desmursii* http://www.avesdechile.cl/327.htm
E032 *Zenaida macroura* http://animaldiversity.ummz.umich.edu/site/accounts/information/Zenaida_macroura.html
E100 *Streptopelia turtur* http://www.oiseaux.net/oiseaux/columbiformes/tourterelle.des.bois.html

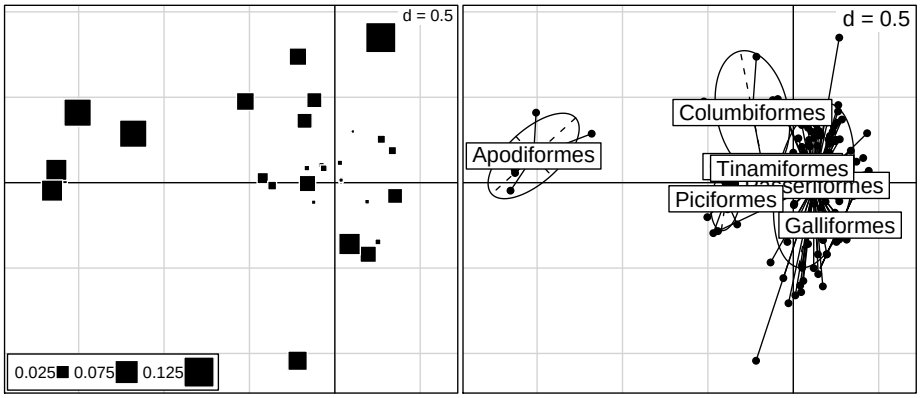
[7] Sources :
E056 *Pheucticus melanocephalus* http://www.birdphotography.com/species/bhgr.html
E026 *Pyrrhula pyrrhula* http://www.oiseaux.net/oiseaux/passeriformes/bouvreuil.pivoine.html

La variance n'est pas une bonne mesure biologique. Elle s'intéresse bien trop aux extrêmes. Elle considère qu'on augmente la diversité en enlevant les valeurs moyennes. Éliminer les modérés dans une assemblée pour ne conserver que des extrémistes (des deux bords !) augmente la diversité. Mais la variance est une forme quadratique et le théorème de Pythagore en fait un pilier. Avec une exponentielle, bonne fonction si il en est, on aura la loi normale et toute la statistique inférentielle. Les mathématiques ont des raisons que l'écologie ne connaît pas.



On aura donc intérêt à introduire des nuances en prenant des distances ultramétriques. La classification hiérarchique permet d'obtenir des solutions qui restent à explorer. Au lieu de mesurer une diversité qui augmente avec la disparition de la quasi totalité des taxons, on aura la possibilité de mesurer une diversité qui augmente avec la préservation en bonnes proportions de tout ce qui fait des différences. C'est clair quand on compare avec la taxonomie :

```
hcmor <- hclust(deuc, met = "single")
ultramor <- cl_ultrametric(hcmor)
wmor <- divcmax(ultramor, epsilon = 0.001)$vectors$num
par(mfrow = c(1, 2))
s.value(w$li, wmor)
s.class(w$li, ecomor$taxo[, 3])
```



Le débat méthodes-données-procédures est alors une source inépuisable de recherches. Il faut bien s'arrêter à un moment donné !

Références

[1] P. Bellot. Méthodes de classification et de catégorisation de textes, 2003.

[2] J. Blondel, F. Vuilleumier, L.F. Marcus, and E. Terouanne. Is there ecomorphological convergence among mediterranean bird communities of chile, california, and france. In M.K. Hecht, B. Wallace, and R.J. MacIntyre, editors, *Evolutionary Biology*, pages 141–213. Vol. 18. Plenum Press, New York, 1984.

[3] F. Cailliez. The analytical solution of the additive constant problem. *Psychometrika*, 48 :305–310, 1983.

[4] S. Champely and D. Chessel. Measuring biological diversity using euclidean metrics. *Environmental and Ecological Statistics*, 9 :167–177, 2002.

[5] P.B. Eckburg, E.M. Bik, C.N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S.R. Gill, K.E. Nelson, and D.A. Relman. Diversity of the human intestinal microbial flora. *Science*, 308 :1635–1638, 2005.

[6] C. Gini. Variabilitá e mutabilitá. Technical report, Universite di Cagliari III, Parte II, 1912.

[7] J.C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53 :325–338, 1966.

[8] J.C. Gower and P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3 :5–48, 1986.

[9] D.L. Hawksworth. *Biodiversity. Measurement and Estimation.* Chapman & Hall with the Royal Society, 1995.

[10] J.A. Jr. Hendrickson and P.R. Ehrlich. An expanded concept of "species diversity". *Notulae Naturae*, 439 :1–6, 1971.

[11] M.O. Hill. Diversity and evenness : a unifying notation and its consequences. *Ecology*, 54 : 427-432 :54 : 427–432, 1973.

[12] S.H. Hurlbert. The non-concept of species diversity : a critique and alternative parameters. *Ecology*, 52 : 577-586 :52 : 577–586, 1971.

[13] L. Lebart. Quelques méthodes de classification, 2003.

[14] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle.* Dunod, Paris, 1995.

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

[16] J.C. Lingoes. Somme boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, 36 :195–203, 1971.

[17] A.E. Magurran. *Ecological diversity and its measurement.* Croom Helm Limited, London, 1988.

[18] P.C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 12 :49–55, 1936.

[19] R.P. McIntosh. An index of diversity and the relation of certains concepts of diversity. *Ecology*, 48 :392–404, 1967.

[20] M. Nei and Hong-Bin Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76 :5269–5273, 1979.

[21] S. Pavoine. *Méthodes statistiques pour la mesure de la biodiversité*. Thèse de doctorat, Université Lyon 1, 2005.

[22] S. Pavoine, A-B. Dufour, and D. Chessel. From dissimilarities among species to dissimilarities among communities : A double principal coordinate analysis. *Journal of Theoretical Biology*, 228 :523–537, 2004.

[23] S. Pavoine, S. Ollier, and A.-B. Dufour. Is the originality of a species measurable ? *Ecology Letters*, 8 :579–586, 2005.

[24] S. Pavoine, S. Ollier, and D. Pontier. Measuring diversity from dissimilarities with rao's quadratic entropy : are any dissimilarity indices suitable ? *Theoretical Population Biology*, 67 :231–239, 2005.

[25] E.C. Pielou. *An introduction to mathematical ecology*. Wiley, New York, 1969.

[26] C.R. Rao. Diversity and dissimilarity coefficients : a unified approach. *Theoretical Population Biology*, 21 :24–43, 1982.

[27] C.R. Rao. Diversity : its measurement, decomposition, apportionment and analysis. *Sankhya, A*, 44 :1–22, 1982.

[28] C.R. Rao. Rao's axiomatization of diversity measures. In S. Kotz and N.L. Johnson, editors, *Encyclopedia of Statistical Sciences, Vol. 7*, pages 614–617. Wiley & Sons, New York, 1986.

[29] M. Roux. Algorithmes de classification, 2006.

[30] E.H. Simpson. Measurement of diversity. *Nature*, 163 : 688 :163 : 688, 1949.

[31] R.S. Spielman. Differences among yanomama indian villages : do the patterns of allele frequencies, anthropometrics and map locations correspond ? *American Journal of Physical Anthropology*, 39 :461–480, 1973.

[32] R.M. Warwick and K.R. Clarke. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology - Progress Series*, 129 :301–305, 1995.