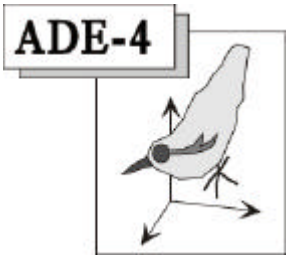


Fiche thématique 6.1

La corrélation entre deux matrices de distances euclidiennes

Résumé

La fiche décrit deux stratégies pour étudier la corrélation entre deux matrices de distances. La première est toujours possible par le test de Mantel. La seconde est réservée aux matrices euclidiennes. Elle utilise le test RV. Les deux tests sont voisins mais la signification du second est interprétable par une analyse de co-inertie entre deux représentations euclidiennes. On discute alors des possibilités qui existent entre couplages de tableaux et couplage de représentations euclidiennes de matrices de distances. L'usage de distances euclidiennes est alors vivement conseillé. L'approche géométrique unifie les deux points de vue.

Plan

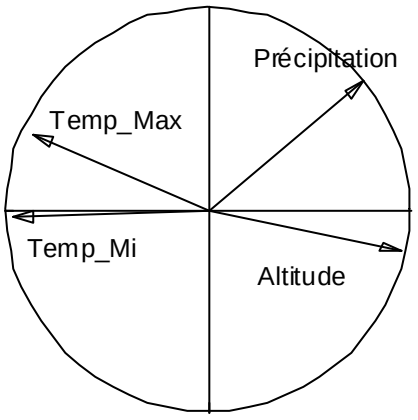
D. Chessel & J. Thioulouse

Introduction

Lorsqu'on dipose de deux matrices de distances sur les mêmes objets, on mesure habituellement la corrélation entre les deux mesures de distance par le test de Mantel [1]. Lorsque ces deux matrices sont euclidiennes, ADE-4 propose de mesurer cette corrélation par le test RV. La fiche explicite l'intérêt pratique de cette modification. On prendra les données de la carte Butterfl proposées dans [2] après [3].

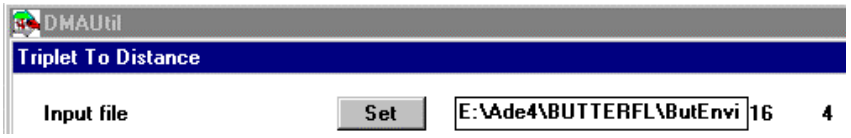
Le tableau de variables environnementales supporte une ACP normée :


```
Correlation matrix ------
[  1] 1000
[  2]  567 1000
[  3] -828 -479 1000
[  4] -936 -705  719 1000
-------------------------
```



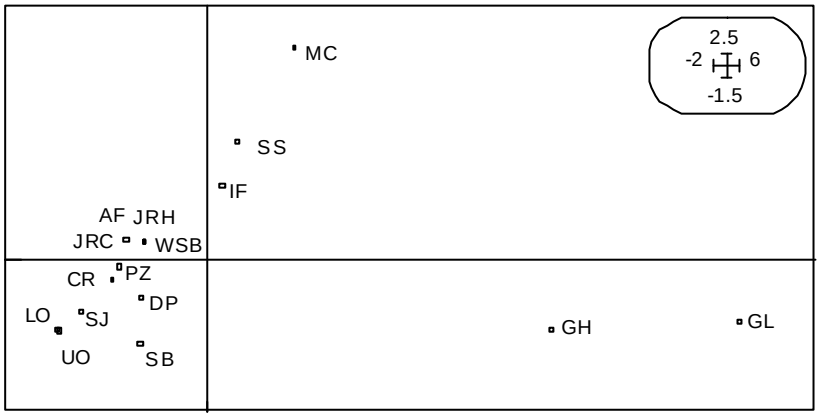
Cette analyse définit une distance environnementale entre deux stations, très simplement par :

$$d_e(i,j) = \sqrt{\sum_{k=1}^{p} \frac{\left(x_{ik} - x_{jk}\right)^2}{\mathrm{var}\left(\mathbf{x}^k\right)}}$$

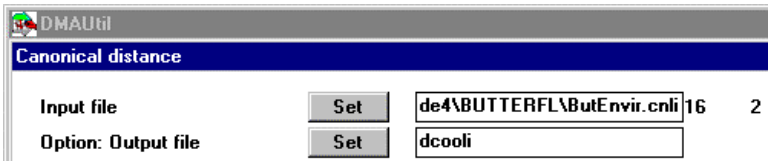



```
Distance matrix computation from a statistical triplet
----------------------------------------
Input file: E:\Ade4\BUTTERFL\ButEnvir.cnta
It has 16 rows and 4 columns
Distances are computed among rows
----------------------------------------
Computed distances use the diagonal metric and the centered table of the triplet
Output file: E:\Ade4\BUTTERFL\ButEnvir_MDcn
It has 120 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\BUTTERFL\ButEnvir_MDcn.dma
  1 -> 16
  2 -> 1
  3 -> Euclidean distance from triplet E:\Ade4\BUTTERFL\ButEnvir.cnta
  4 -> TRUE
----------------------------------------
```

Cette matrice de distances est euclidienne par définition. Les distances entre deux stations sont calculées comme distances entre points de $\mathbb{R}^4$. Le nuage des points dont les distances deux à deux sont dans cette matrice sont représentés au mieux par la carte factorielle :



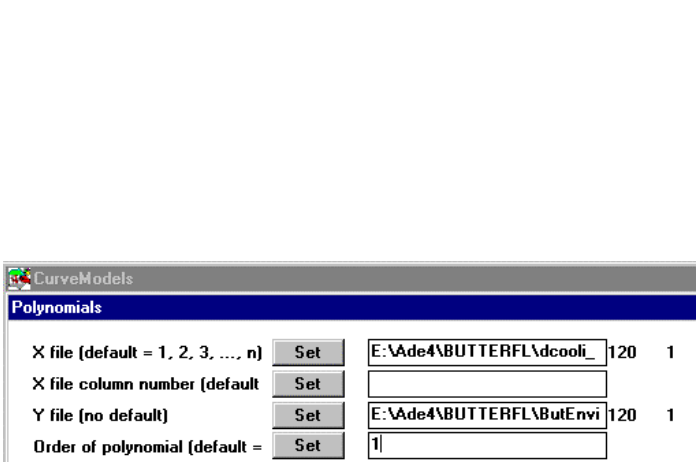
On peut calculer la distance effective entre points sur cette carte :

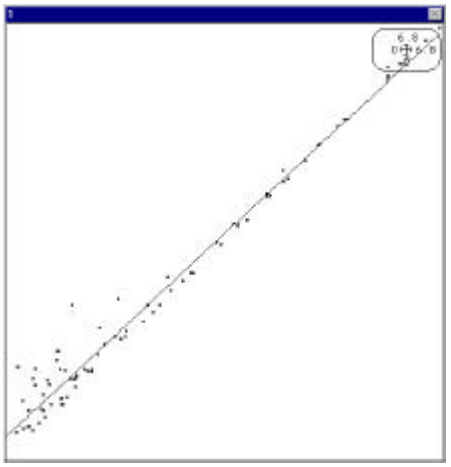



```
Distance matrix computation
----------------------------------------
Input file: E:\Ade4\BUTTERFL\ButEnvir.cnli
It has 16 rows and 2 columns
Distances are computed among rows
----------------------------------------
Canonical distances computed
Output file: dcooli_EU
It has 120 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: dcooli_EU.dma
```

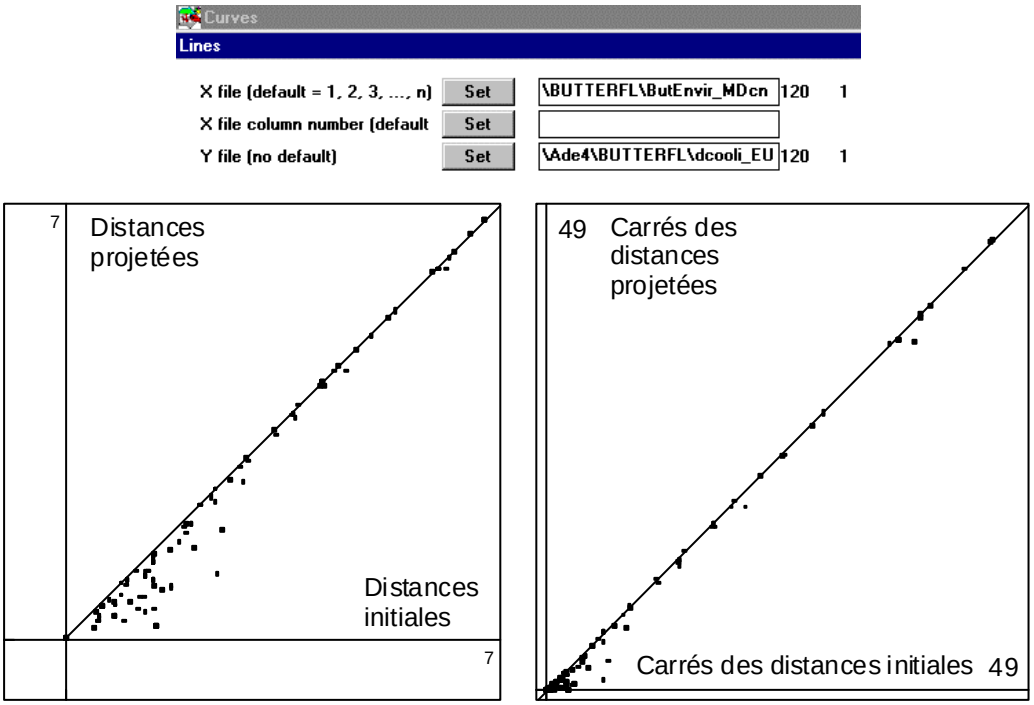
```
  1 -> 16
  2 -> 1
  3 -> Classical metric on E:\Ade4\BUTTERFL\ButEnvir.cnli
  4 -> TRUE
----------------------------------------
```

Le lien entre la distance représentée sur la carte et la distance effective dans l'espace est fort :





Cette manière de voir le lien entre deux matrices de distances est celui du test de Mantel. La distance calculée sur les deux premières coordonnées factorielles est proche de la distance calculée sur les 4 variables de départ. En fait, la représentation ci-dessus est inexacte. Il vaut mieux représenter la distance projetée en fonction de la distance initiale (ci-dessous à gauche) :



Le nuage de points est entièrement sous la droite $y = x$. Voir l'interprétation de cette figure dans [4] p. 118. En fait, il vaudrait mieux encore représenter les distances au carrés (ci-dessus, à droite). Cette opération fort simple montre que l'objectif de l'ACP est de représenter sur la carte factorielle la plus grande part possible des carrés des distances entre points du nuage initial. On a :

$$\sum_{i=1}^{n} d_e^2(i,j) = \sum_{i=1}^{n} d_p^2(i,j) + Erreur$$

La part représentée sur le plan factoriel est alors exactement :

$$\frac{\sum_{i=1}^{n} d_p^2(i,j)}{\sum_{i=1}^{n} d_e^2(i,j)} = \frac{l_1 + l_2}{l_1 + l_2 + \ldots + l_p}$$

Dans S-PLUS, on a :

```
> w_princomp(butenvir,cor=T)$sdev
> w
  Comp. 1   Comp. 2   Comp. 3   Comp. 4
 1.771848 0.7584027 0.5032749 0.1791484
> w*w
  Comp. 1   Comp. 2   Comp. 3    Comp. 4
 3.139446 0.5751746 0.2532856 0.03209416
> sum(w*w)
[1] 4
> w_w*w
> (w[1]+w[2])/sum(w)
   Comp. 1
 0.9286551
```

Dans ADE-4 :

```
Total inertia:        4
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +3.1394E+00 +0.7849 +0.7849   |02   +5.7517E-01 +0.1438 +0.9287 |
03   +2.5329E-01 +0.0633 +0.9920   |04   +3.2094E-02 +0.0080 +1.0000 |
```

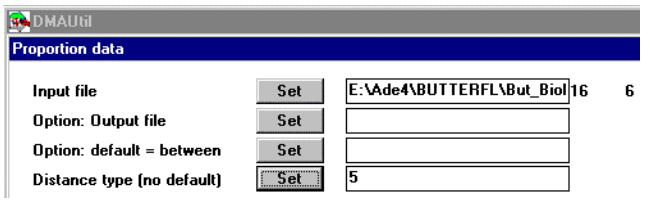
On vérifie alors que :

$$\frac{\sum_{i=1}^{n} d_p^2(i,j)}{\sum_{i=1}^{n} d_e^2(i,j)} = \frac{950.94}{1023.00} = 0.9287$$

On voit ici que le lien entre distances projetées (calculées sur les coordonnées) et distances de départ (calculées sur les variables) n'est pas du type corrélation simple. On dirait ici que 93% des carrés des distances de départ sont représentées sur la carte factorielle et qu'il est impossible de faire mieux par projection sur un plan. On part ici d'un nuage de points dont l'analyse représente au mieux les distances deux à deux. Partons maintenant d'une matrice de distances.

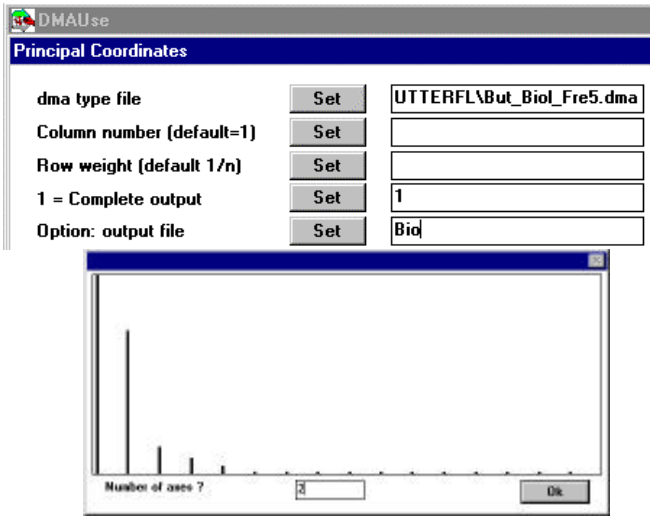
La représentation euclidienne d'une matrice de distances

Les données biologiques sont des profils de fréquences alléliques. On peut calculer les distances entre populations :



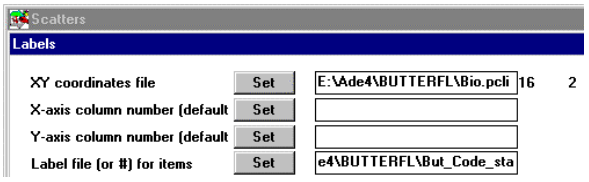

```
Distance amongst frequency distributions
Input file: E:\Ade4\BUTTERFL\But_Biol
It has 16 rows and 6 columns
Distances are computed among rows

d5 = sqrt (1 - (Sum(sqrt(p(i)q(i))))
Edwards 1971 in Hartl & Clark 1989
Output file: E:\Ade4\BUTTERFL\But_Biol_Fre5
It has 120 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\BUTTERFL\But_Biol_Fre5.dma
  1 -> 16
  2 -> 1
  3 -> EDWARDS on E:\Ade4\BUTTERFL\But_Biol
  4 -> TRUE
```

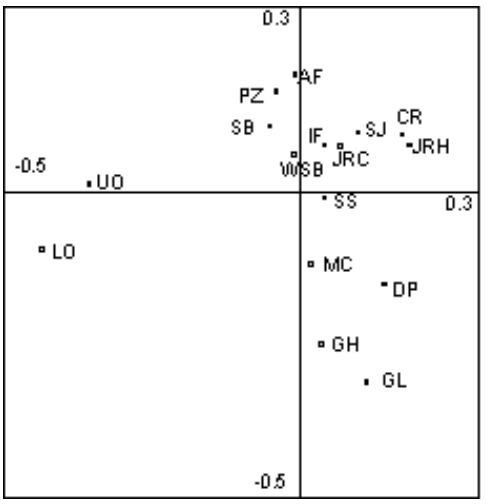
Cette distance est euclidienne. Il existe un nuage de points dans un sous-espace dont les distances sont celles de la matrice.




```
Input file: E:\Ade4\BUTTERFL\But_Biol_Fre5.dma
Distance file: E:\Ade4\BUTTERFL\But_Biol_Fre5
Row: 16 Col: 1 Col used: 1
Origin: EDWARDS on E:\Ade4\BUTTERFL\But_Biol
Euclidean distance / Uniform row weights

Rank : 6 Inertia  5.380e-02
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.7262E-02 +0.5068 +0.5068   |02   +1.9696E-02 +0.3661 +0.8729 |
03   +3.6052E-03 +0.0670 +0.9399   |04   +2.1534E-03 +0.0400 +0.9799 |
05   +9.6421E-04 +0.0179 +0.9979   |06   +1.1557E-04 +0.0021 +1.0000 |

File Bio.pcta contains the principal coordinates  (norm=sqrt(lambda))
--- It has 16 rows and 6 columns
File :Bio.pcta
```

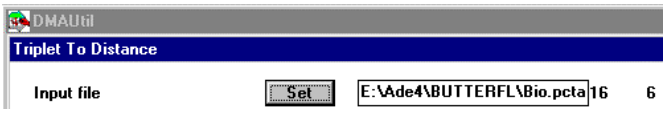
Le rang est 6, ce qui veut dire que le nuage de points reconstituant la matrice de données est dans $\mathbb{R}^6$. On peut, au vu du graphe des vecteurs propres, résumer ce nuage par une carte de dimension 2 :



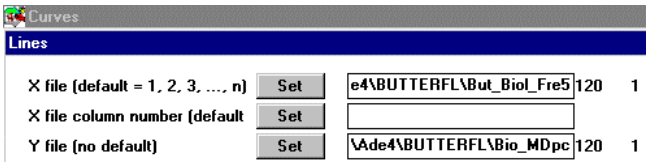


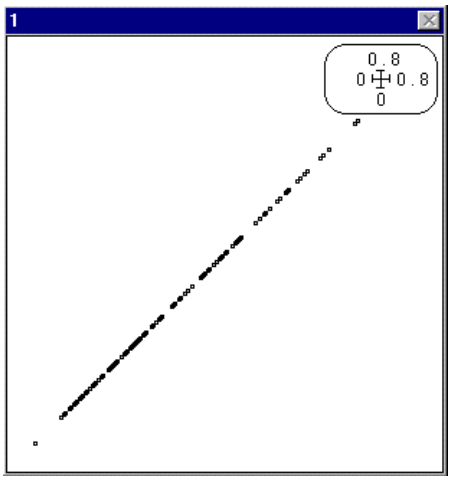

```
Inner product reconstitution : quality index
Drouet d'Aubigny 1989 p. 130.
HS norm squared  1.150e-03
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +7.4322E-04 +0.6464 +0.6464   |02   +3.8794E-04 +0.3374 +0.9838 |
03   +1.2997E-05 +0.0113 +0.9951   |04   +4.6370E-06 +0.0040 +0.9992 |
05   +9.2970E-07 +0.0008 +1.0000   |06   +1.3355E-08 +0.0000 +1.0000 |
```

On peut à nouveau comparer les distances qui sont dans la matrice de départ, celles qui dérivent du nuage de points et celles qui sont proviennent des points projetées sur le plan.




```
Distance matrix computation from a statistical triplet
----------------------------------------
Input file: E:\Ade4\BUTTERFL\Bio.pcta
It has 16 rows and 6 columns
Distances are computed among rows
----------------------------------------
Computed distances use the diagonal metric and the centered table of the triplet
Output file: E:\Ade4\BUTTERFL\Bio_MDpc
It has 120 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\BUTTERFL\Bio_MDpc.dma
  1 -> 16
  2 -> 1
  3 -> Euclidean distance from triplet E:\Ade4\BUTTERFL\Bio.pcta
  4 -> TRUE
----------------------------------------
```





Conformément à la théorie, la distance entre deux points de la représentation euclidienne est exactement la distance de la matrice de départ. La somme des carrés de ces distances est représentée à 87% sur le plan factoriel de l'analyse en coordonnées principales.

Nous avons donc un tableau de variables environnementales qui a donné une matrice de distances environnementales et une matrice de distances biologiques qui a donné un tableau (abstrait) ou nuages de points représentant les distances biologiques. On peut donc comparer soit les distances soit les tableaux.

La corrélation entre les distances

Se pose alors la question de la corrélation entre matrices de distances. On peut comparer deux approches. La première, toujours possible, est traditionnelle et dérive du test de Mantel. Elle est décrite dans [5] p. 114 et consiste simplement à considérer que la demi-matrice de distances est un vecteur. La corrélation entre deux matrices de distances est la corrélation ordinaire entre deux variables qu'on notera $\mathrm{cor}(\mathbf{D},\mathbf{E})$.

La seconde est réservée aux matrices de distances euclidiennes et est mathématiquement très différente. Si $\mathbf{D}$ est une matrice de distances euclidienne, d'après le théorème de Gower [6], on considère la matrice $\Delta(\mathbf{D})$ et sa dérivée par double centrage :

$$\mathbf{D}=\left[d_{ij}\right]\Rightarrow\Delta(\mathbf{D})=-\frac{1}{2}\left[d_{ij}^2\right]\Rightarrow\Delta_0(\mathbf{D})=(\mathbf{I}-\mathbf{Q})\Delta(\mathbf{D})(\mathbf{I}-\mathbf{Q}^t)$$

$\mathbf{D}$ est euclidienne si et seulement si $\Delta_0(\mathbf{D})$ est semi-définie positive. La diagonalisation de $\Delta_0(\mathbf{D})$ est la base de l'analyse en coordonnées principales. Si $\mathbf{E}$ est également euclidienne, le produit scalaire d'Hilbert-Schmidt, introduit pour la première fois en analyse de données par Escoufier [7], entre les deux opérateurs défini par :

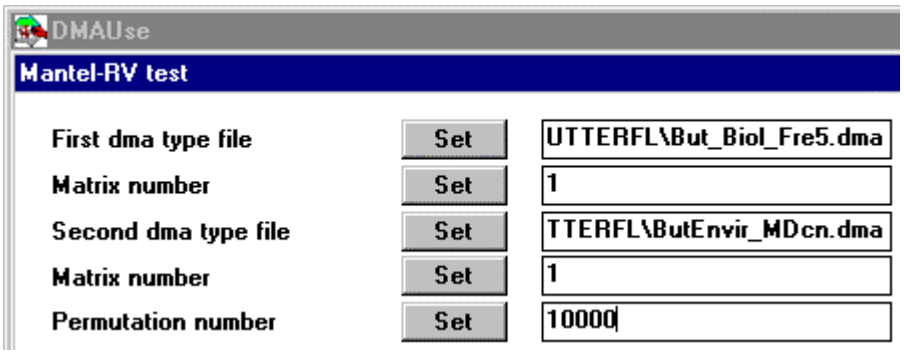
$$\left\langle \Delta_0(\mathbf{D}) \middle| \Delta_0(\mathbf{E}) \right\rangle_{HS} = \mathrm{trace}\left(\Delta_0(\mathbf{D})\Delta_0(\mathbf{E})\right)$$

donne une mesure de corrélation entre les deux matrices par :

$$\mathrm{RV}(\mathbf{D},\mathbf{E}) = \cos\left(\Delta_0(\mathbf{D}), \Delta_0(\mathbf{E})\right) = \frac{\left\langle \Delta_0(\mathbf{D}) \middle| \Delta_0(\mathbf{E}) \right\rangle_{HS}}{\left\|\Delta_0(\mathbf{D})\right\|_{HS} \left\|\Delta_0(\mathbf{E})\right\|_{HS}}$$

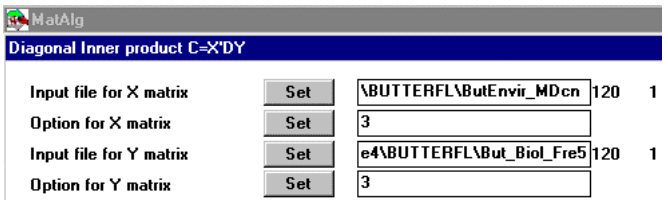
Le calcul est très simple puisqu'il reprend celui de Mantel sur les carrés des distances (en lieu des distances elles-mêmes) et *en utilisant le double centrage* de la matrice (qui implique le centrage simple sur le vecteur) pour calculer la corrélation. Le RV est toujours compris entre 0 et 1, tandis que la corrélation de Mantel est comprise entre -1 et 1.

On calcule les deux et le même test de permutations :




```
Correlation between two distance matrices
First input file: E:\Ade4\BUTTERFL\But_Biol_Fre5.dma
Text file: E:\Ade4\BUTTERFL\But_Biol_Fre5.dma
  1 -> 16
  2 -> 1
  3 -> EDWARDS on E:\Ade4\BUTTERFL\But_Biol
  4 -> TRUE
Matrix used : 1
Second input file: E:\Ade4\BUTTERFL\ButEnvir_MDcn.dma
Text file: E:\Ade4\BUTTERFL\ButEnvir_MDcn.dma
  1 -> 16
  2 -> 1
  3 -> Euclidean distance from triplet E:\Ade4\BUTTERFL\ButEnvir.cnta
  4 -> TRUE
Matrix used : 1
Permutation test on r value (Manly 1991 p. 114)
----------------------------------------
Matrix 1 of E:\Ade4\BUTTERFL\But_Biol_Fre5 versus matrix 1 of
E:\Ade4\BUTTERFL\ButEnvir_MDcn
r index :  3.643e-01
```

On retrouve cette valeur simplement comme une corrélation :




```
[  1]  3.6431e-01
```

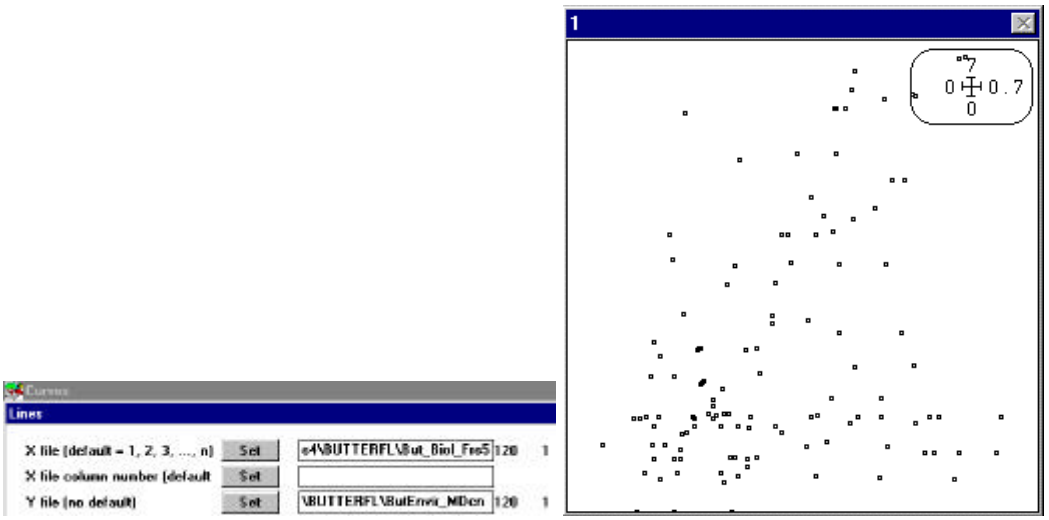
Le test de Mantel donne un niveau de signification de 5% :

```
number of random matching: 10000   Observed: 0.364315
Histogramm:  minimum = -0.339187, maximum = 0.799806
number of simulation X<Obs: 9509 (frequency: 0.950900)
number of simulation X>=Obs: 491 (frequency: 0.049100)

    |*
    |****************
    |********************************
    |**************************************************
    |***************************************************
    |****************************************
    |***************************
    |***********************
    |********************
    |**************
    |***********
    |**********
o->|*******
    |****
    |**
    |*
    |*
    |*
    |
    |
```

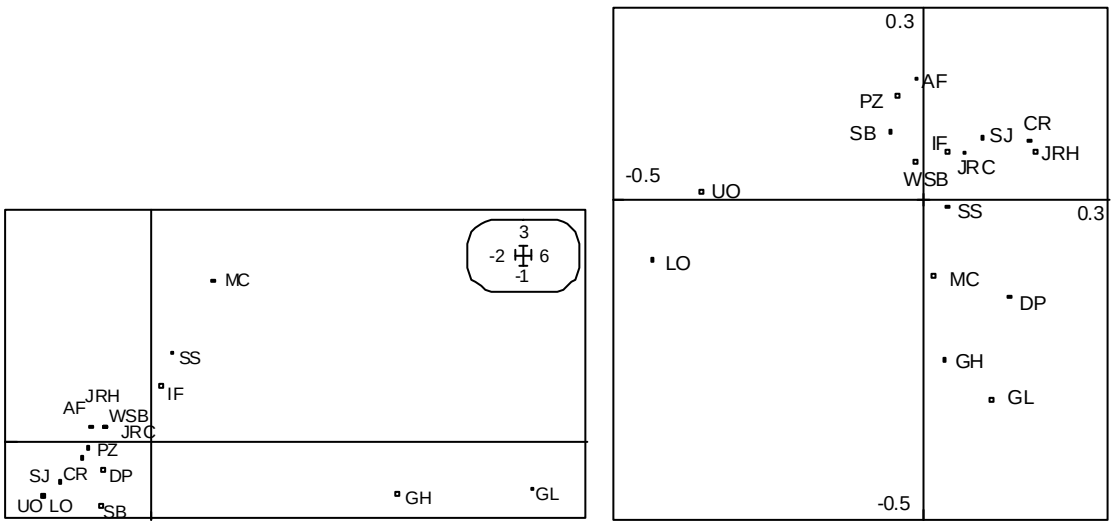
Le test RV est un peu différent :

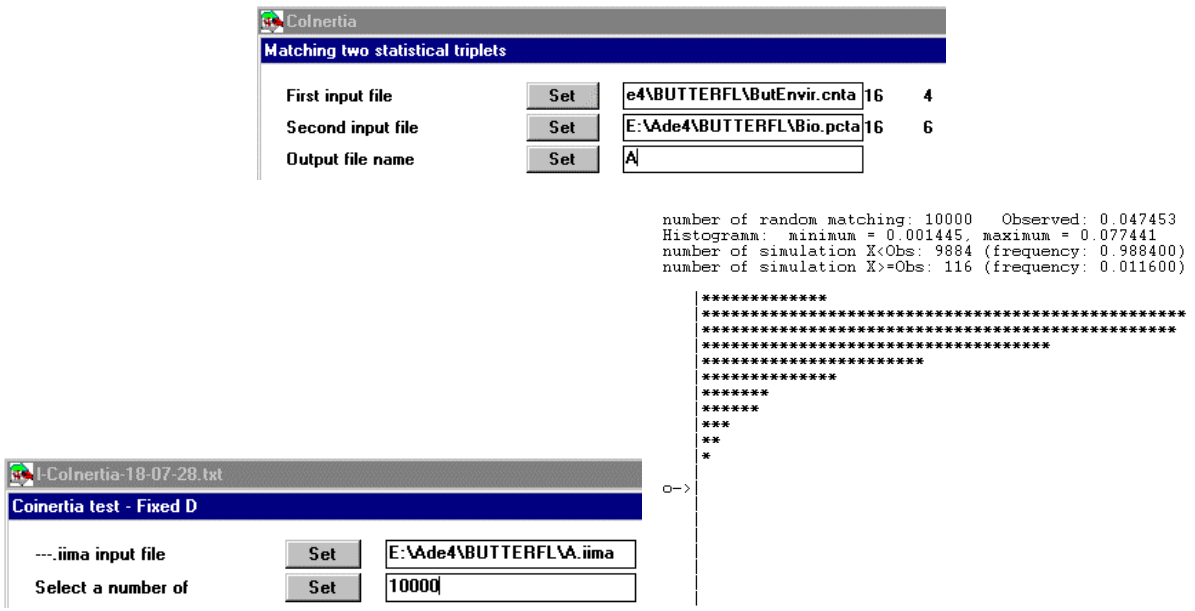
```
Matrix 1 of E:\Ade4\BUTTERFL\But_Biol_Fre5 versus matrix 1 of
E:\Ade4\BUTTERFL\ButEnvir_MDcn

RV index :  4.371e-01
number of random matching: 10000   Observed: 0.437075
Histogramm:  minimum = 0.015232, maximum = 0.718666
number of simulation X<Obs: 9890 (frequency: 0.989000)
number of simulation X>=Obs: 110 (frequency: 0.011000)

    |**************
    |***************************************************
    |********************************************
    |*******************************
    |********************
    |************
    |*******
    |*****
    |***
    |**
    |*
o->|
```

La distribution est plus dissymétrique mais le niveau de signification est de l'ordre de 1%. L'interprétation de ces deux notions sont fort différentes. Le premier est le simple examen de la corrélation entre distances. Le graphique associé est celui d'un nuage bivarié :

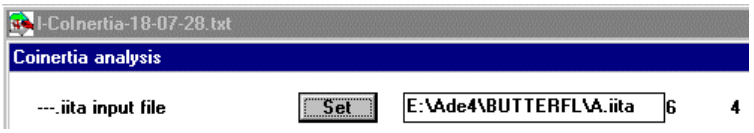


Il est difficile d'identifier pourquoi les matrices de distances sont corrélées. Avec le RV, c'est beaucoup plus clair. En effet, le test est exactement celui de la co-inertie entre les deux représentations euclidiennes. Nous avons deux analyses d'inertie et pouvons chercher ce qui les relie :





On obtient le même résultat. Ce qui est significatif s'exprime alors dans l'analyse de co-inertie :

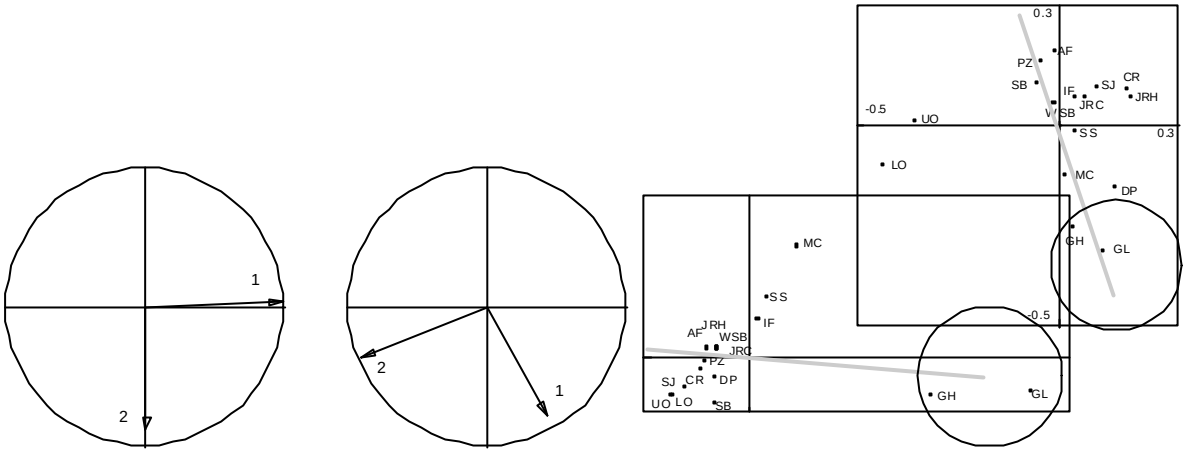



```
DiagoRC: General program for two diagonal inner product analysis
Input file: E:\Ade4\BUTTERFL\A.iita
--- Number of rows: 6, columns: 4
-----------------------
Total inertia: 0.0474528
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +4.5470E-02 +0.9582 +0.9582   |02   +1.8657E-03 +0.0393 +0.9975 |
03   +1.0905E-04 +0.0023 +0.9998   |04   +8.3988E-06 +0.0002 +1.0000 |
```

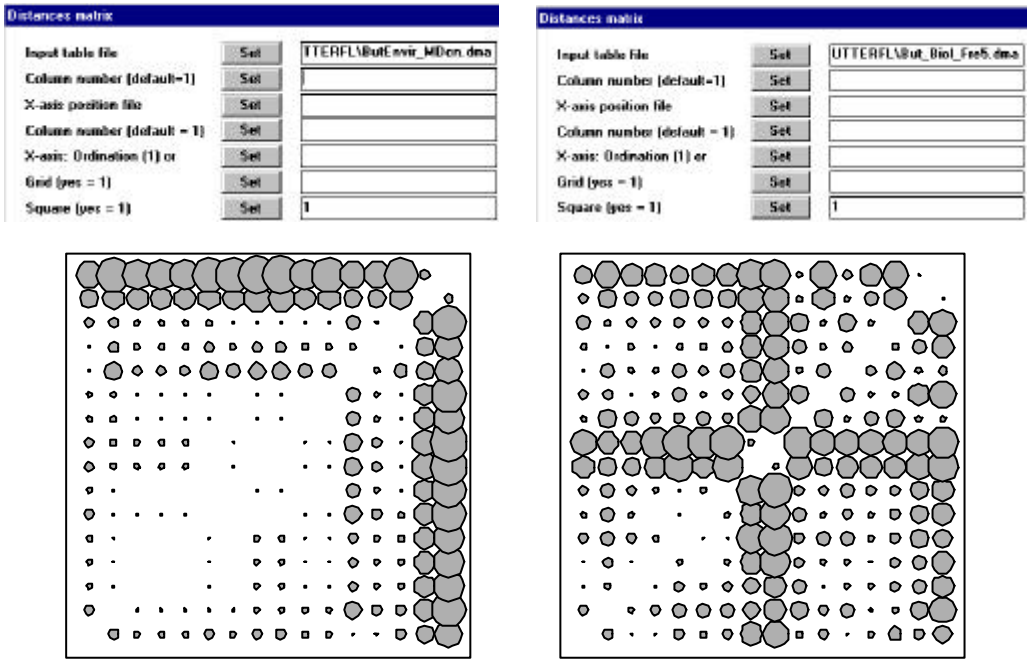
Un seul axe de co-inertie mérite d'être dépouillé. On retrouve le RV entre les deux matrices de distances, comme RV entre les deux représentations euclidiennes.

```
Co-inertia analysis between two statistical triplets

1 ---> E:\Ade4\BUTTERFL\ButEnvir.cnta (rows: 16, col: 4, axes: 2, inertia: 4.0)
2 ---> E:\Ade4\BUTTERFL\Bio.pcta (rows: 16, col: 6, axes: 2, inertia: 0.053797)
Co-inertia: 0.047453, RV coefficient: 0.43707
```

L'analyse repère qu'il faut associer l'axe 2 du nuage de la représentation euclidienne des données biologiques à l'axe 1 du nuage des données environnementales pour rendre compte de 96% de la co-inertie des deux nuages. Ceci se voit simplement sur les cartes des analyses simples :



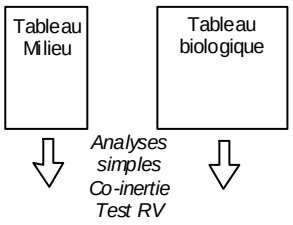
Ceci se voit sur les matrices de distances :



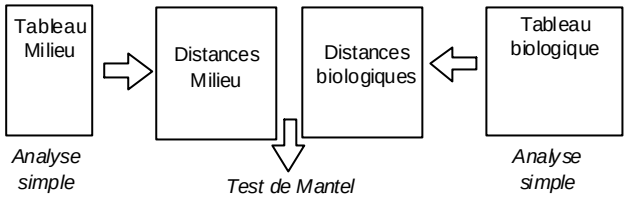
Sur la matrice de distances biologiques, on voit clairement l'élément qui la relie à la matrice de distances environnementales en même temps qu'il existe un autre élément de structure qui n'existe pas dans son vis-à-vis.

Diverses situations

Il convient alors d'énumérer plusieurs situations concrètes. La première est la plus simple :

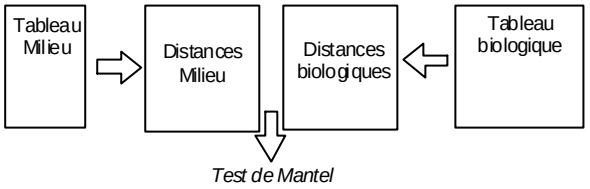


Quand on a deux tableaux qui supportent deux analyses, toute l'information est donnée par l'analyse de co-inertie du couple. Evidemment, des deux analyses simples on peut induire deux matrices de distances et étudier la corrélation entre matrices :



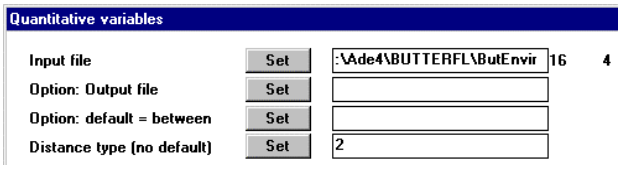
Cette stratégie est nettement plus faible que la première. On aura simplement un niveau de signification peu interprétable. Avec deux tableaux de données supportant naturellement deux analyses simples, le passage par les matrices de distances ne s'impose pas.

La situation se complique quand on désire utiliser une distance qui pour une raison ou une autre n'est pas simplement reliées à une analyse de base :



C'est une situation fréquente. Par exemple, dans le cas étudié on peut utiliser pour les variables environnementales la distance de Manhattan, ou de Cain & Harrison (référence p. 20 dans [8]), ou D_3 de Gower & Legendre [6] (distance non euclidienne) :

$$d_2(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{\left|x_{ik}-x_{jk}\right|}{r_k} \text{ avec } r_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ik}-m_k\right)^2} \text{ et } m_k = \frac{1}{n}\sum_{i=1}^{n}x_{ik}$$

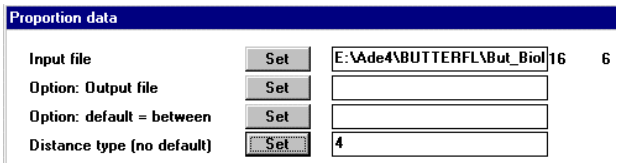



```
Distance matrix computation from dissimilarity coefficients
Dissimilarity coefficients amongst quantitative variables
Gower J.C. & Legendre P. (1986)
Metric and Euclidean properties of dissimilarity coefficients
Journal of Classication, 3, 5-48
Table 3 p. 27

Input file: E:\Ade4\BUTTERFL\ButEnvir
It has 16 rows and 4 columns
Distances are computed among rows

City-Block/Range = Manhattan/ Standard deviation
D3 coefficient of GOWER & LEGENDRE
Non Euclidean distance
Distances are computed by
```

```
dij = (1/p)Sum|xik-xjk|/sk 1<=k<=p
sk = standard deviation  1<=k<=p
Output file: E:\Ade4\BUTTERFL\ButEnvir_Dqv2
It has 120 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\BUTTERFL\ButEnvir_Dqv2.dma
  1 -> 16
  2 -> 1
  3 -> S2 coefficient of GOWER & LEGENDRE on E:\Ade4\BUTTERFL\ButEnvir
  4 -> FALSE
```

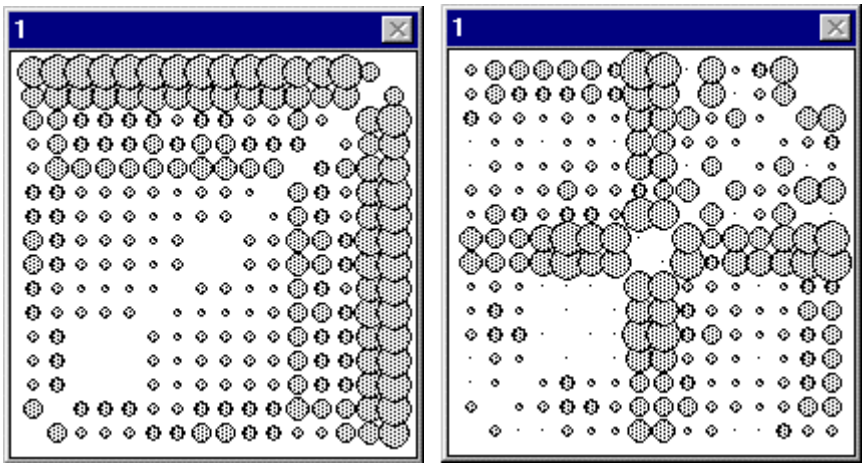
Pour les variables biologiques, on peut de même préférer la distance de Nei [9] :

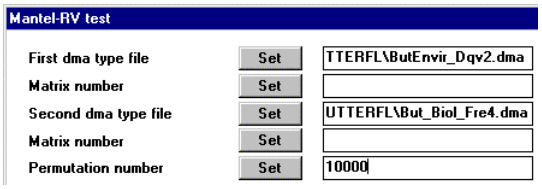
$$d_4 = -\ln\left(\frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2}\sqrt{\sum_i q_i^2}}\right)$$




```
Distance amongst frequency distributions
Input file: E:\Ade4\BUTTERFL\But_Biol
It has 16 rows and 6 columns
Distances are computed among rows

d4=-ln( (Sum(p(i)q(i)) / sqrt( Sum(p(i)*p(i))*Sum(q(i)*q(i)) ) )
Nei 1972 in Avise 1994 p. 95
Test of the euclidean property by diagonalization (theorem of GOWER)
Output file: E:\Ade4\BUTTERFL\But_Biol_Fre4
It has 120 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\BUTTERFL\But_Biol_Fre4.dma
  1 -> 16
  2 -> 1
  3 -> NEI on E:\Ade4\BUTTERFL\But_Biol
  4 -> FALSE
```

Dans ce cas, on doit se contenter d'éditer les matrices (on représente alors les distances plutôt que les carrés des distances, par souci de cohérence) et de tester la statistique de Mantel :

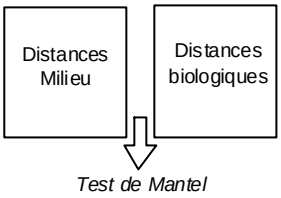





```
Matrix 1 of E:\Ade4\BUTTERFL\ButEnvir_Dqv2 versus matrix 1 of
E:\Ade4\BUTTERFL\But_Biol_Fre4
r index :  2.399e-01
number of random matching: 10000   Observed: 0.239945
Histogramm:  minimum = -0.320741, maximum = 0.769745
number of simulation X<Obs: 8563 (frequency: 0.856300)
number of simulation X>=Obs: 1437 (frequency: 0.143700)

    |**
    |****************
    |**********************************
    |******************************************
    |************************************************
    |***************************************
    |*************************
    |*******************
    |************
    |*******
o->|************
    |*************
    |*********
    |****
    |*
    |
    |
    |
    |*
    |*
Non euclidean distances. HS cosinus non available
```

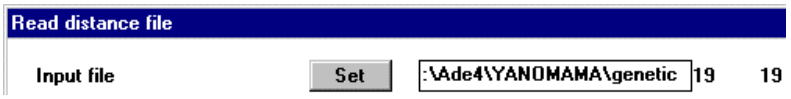
Le résultat n'est pas fameux et on peut passer à côté de l'essentiel.

Il se peut également qu'on ne dispose, pour une raison quelconque que des matrices de distances. C'est souvent le résultat disponible dans la littérature. Si ces distances ne sont pas euclidiennes, on se contentera de la situation :

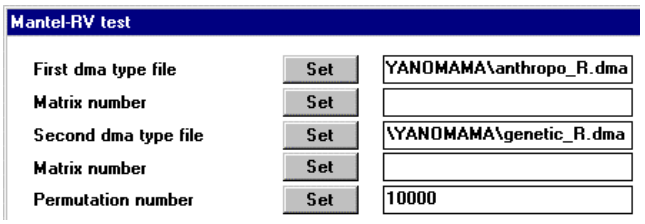


Par exemple, en partant de la carte Yanomama (données de **10** dans **5**) :


```
Input file: E:\Ade4\YANOMAMA\anthropo
E:\Ade4\YANOMAMA\anthropo is a binary file with 19 rows and 19 columns
Squared matrix: Ok
Non negative value: Ok
Dii = 0 for all i: Ok
Symetric matrix: Ok
Test of the euclidean property by diagonalization (theorem of GOWER)
Output file: E:\Ade4\YANOMAMA\anthropo_R
It has 171 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\YANOMAMA\anthropo_R.dma
  1 -> 19
  2 -> 1
  3 -> Input distance file E:\Ade4\YANOMAMA\anthropo
  4 -> FALSE
```

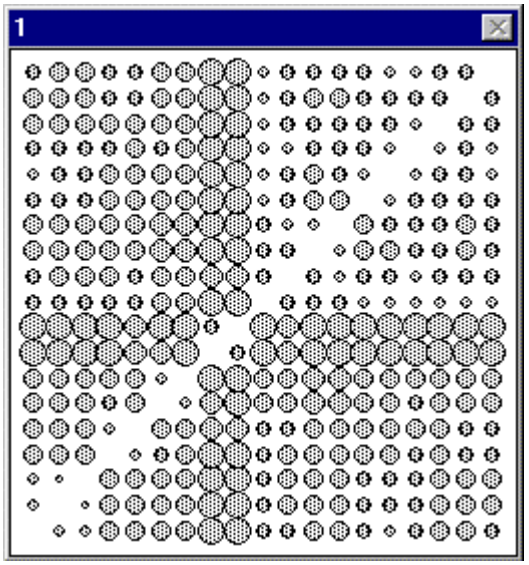


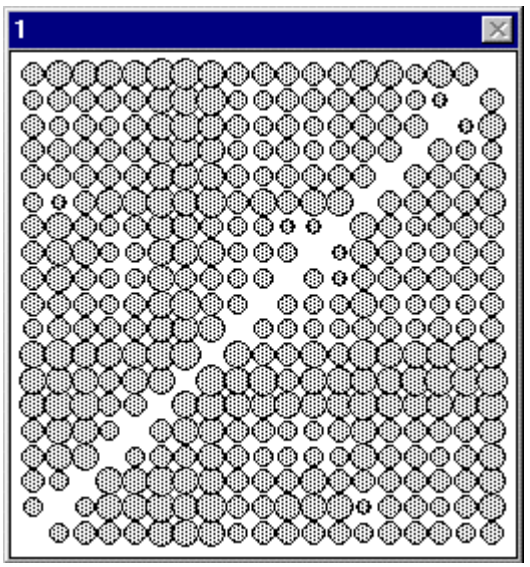

```
Input file: E:\Ade4\YANOMAMA\genetic
E:\Ade4\YANOMAMA\genetic is a binary file with 19 rows and 19 columns
Squared matrix: Ok
Non negative value: Ok
Dii = 0 for all i: Ok
Symetric matrix: Ok
Test of the euclidean property by diagonalization (theorem of GOWER)
Output file: E:\Ade4\YANOMAMA\genetic_R
It has 171 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\YANOMAMA\genetic_R.dma
  1 -> 19
  2 -> 1
  3 -> Input distance file E:\Ade4\YANOMAMA\genetic
  4 -> FALSE
```



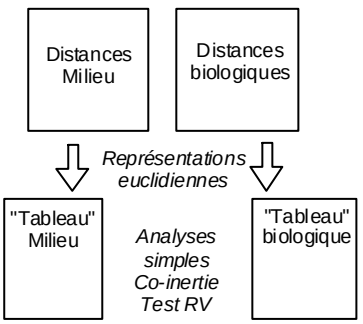

```
Matrix 1 of E:\Ade4\YANOMAMA\anthropo_R versus matrix 1 of
E:\Ade4\YANOMAMA\genetic_R
r index :  2.996e-01
number of random matching: 10000   Observed: 0.299551
```

```
Histogramm:  minimum = -0.362376, maximum = 0.591415
number of simulation X<Obs: 9534 (frequency: 0.953400)
number of simulation X>=Obs: 466 (frequency: 0.046600)

    |
    |*****
    |****************
    |********************************
    |*************************************************
    |***************************************************
    |****************************************
    |******************************
    |**************************
    |**************************
    |***************************
    |************************
    |****************
o->|**********
    |******
    |***
    |**
    |**
    |*
    |
Non euclidean distances. HS cosinus non available
```

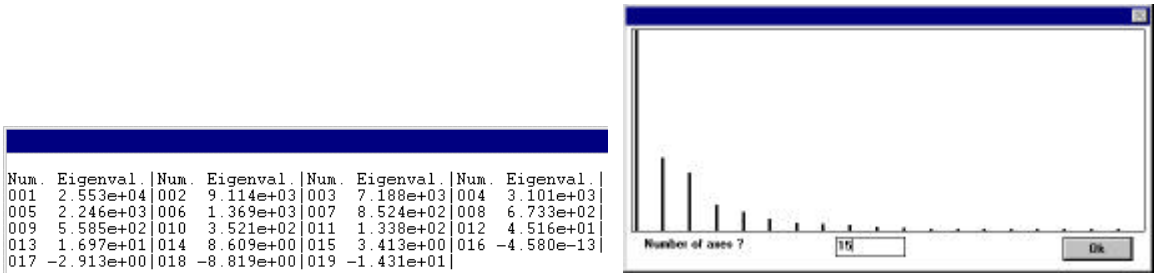




Lorsque ces distances sont euclidiennes, on a alors une possibilité nouvelle par :

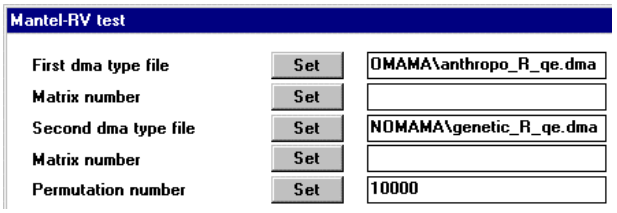


Dans l'exemple qui précède, si les matrices ne sont pas euclidienne à la lecture, c'est uniquement à cause du format d'édition dans l'ouvrage cité. Elles sont quasi-euclidiennes et supportent une excellente approximation :




```
File E:\Ade4\YANOMAMA\anthropo_R_qe contains distance half-matrices
Rows = 171  Cols = 1
One half-matrix per column
d21, d31, d32, d41, d42, d43, ..
Text file: E:\Ade4\YANOMAMA\anthropo_R_qe.dma
  1 -> 19
  2 -> 1
  3 -> Euclidean matrix by positive eigenvalues from E:\Ade4\YANOMAMA\anthropo_R
  4 -> TRUE
----------------------------------------
File E:\Ade4\YANOMAMA\genetic_R_qe contains distance half-matrices
Rows = 171  Cols = 1
One half-matrix per column
d21, d31, d32, d41, d42, d43, ..
Text file: E:\Ade4\YANOMAMA\genetic_R_qe.dma
  1 -> 19
  2 -> 1
  3 -> Euclidean matrix by positive eigenvalues from E:\Ade4\YANOMAMA\genetic_R
  4 -> TRUE
```

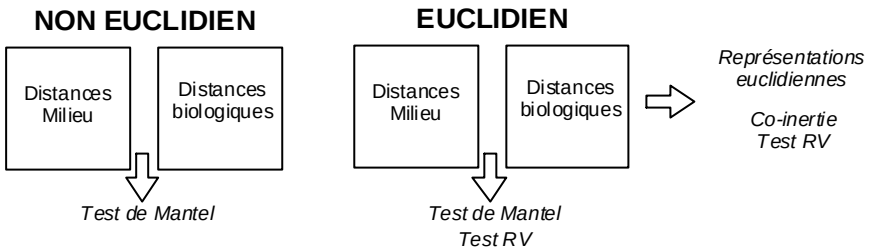
Alors :




```
Matrix 1 of E:\Ade4\YANOMAMA\anthropo_R_qe versus matrix 1 of
E:\Ade4\YANOMAMA\genetic_R_qe
RV index :  4.273e-01
number of random matching: 10000   Observed: 0.427270
Histogramm:  minimum = 0.111225, maximum = 0.505320
number of simulation X<Obs: 9921 (frequency: 0.992100)
number of simulation X>=Obs: 79 (frequency: 0.007900)
   |
   |******
   |*****************
   |*******************************
   |*******************************************
   |*******************************************
```

```
    |************************************************
    |*****************************************
    |*********************************
    |*************************
    |******************
    |**************
    |*********
    |*****
    |***
    |**
o->|*
    |
    |
    |
```

On est passé d'un seuil de 5% à un seuil de 1 pour mille. On a de plus la possibilité d'interpréter la corrélation par la co-inertie des deux représentations euclidiennes. En bref , le schéma est :



Les deux tests RV mentionnés sont mathématiquement identiques. Ceci conduit à préférer, sauf raisons très particulières, les distances euclidiennes à celle qui ne le sont pas.

Distances et analyses simples

On peut alors faire une remarque simple. Toute analyse simple définit une distance et toute distance euclidienne définit une analyse simple. Il arrive que les deux points de vue soient identiques. C'est le cas de la distance d'Edwards (**11**, **12** p. 167) :

$$d(i,j)=\sqrt{1-\sum_{k=1}^{p}\sqrt{p_{ik}p_{jk}}}$$

On a :

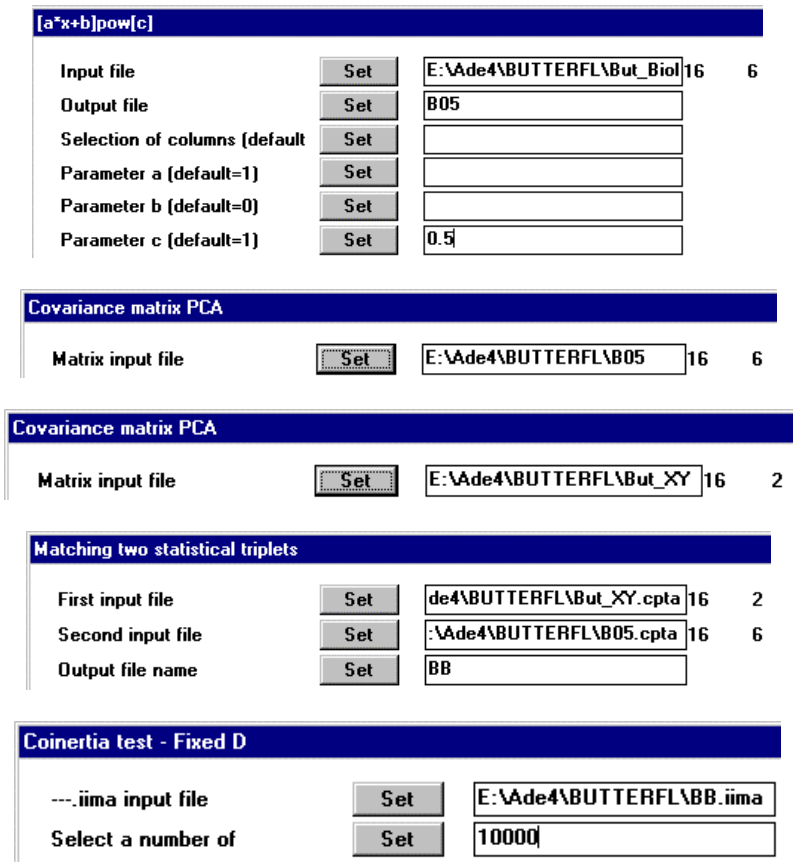
$$d^2(i,j)=1-\sum_{k=1}^{p}\sqrt{p_{jk}}\sqrt{p_{ik}}=\frac{1}{2}\left(2-\sum_{k=1}^{p}\sqrt{p_{jk}}\sqrt{p_{ik}}\right)$$

soit :

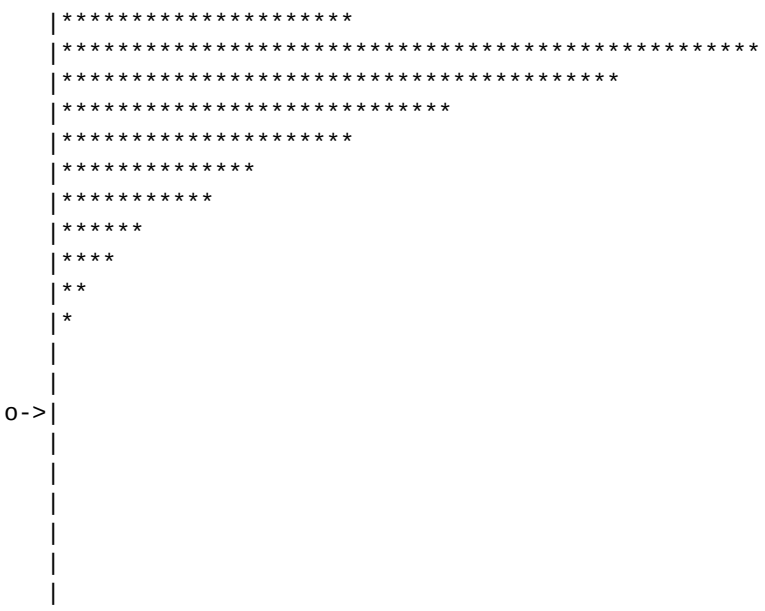
$$\sum_{k=1}^{p}\left(\sqrt{p_{ik}}\right)^2=1\Rightarrow d^2(i,j)=\frac{1}{2}\sum_{k=1}^{p}\left(\sqrt{p_{jk}}-\sqrt{p_{ik}}\right)^2$$

A une constante près qui n'intervient pas dans les tests de corrélation, la distance utilisée est celle de l'ACP des fréquences transformée par $y=\sqrt{x}$.

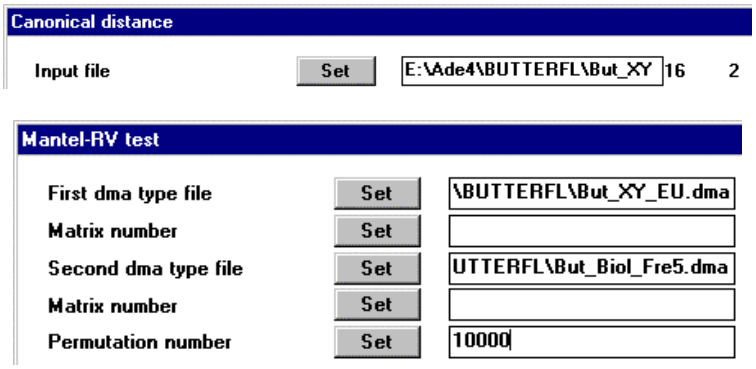
Tester le co-inertie entre l'ACP du tableau de variables biologiques et l'ACP du tableau de coordonnées spatiales sera exactement tester la corrélation entre distances biologiques d'Edwards et distances spatiales traditionnelles. On peut vérifier. Dans la stratégie tableau-tableau :




```
number of random matching: 10000   Observed: 11377.297852
Histogramm:  minimum = 45.699284, maximum = 16615.679688
number of simulation X<Obs: 9959 (frequency: 0.995900)
number of simulation X>=Obs: 41 (frequency: 0.004100)
```



Dans la stratégie distances-distances :

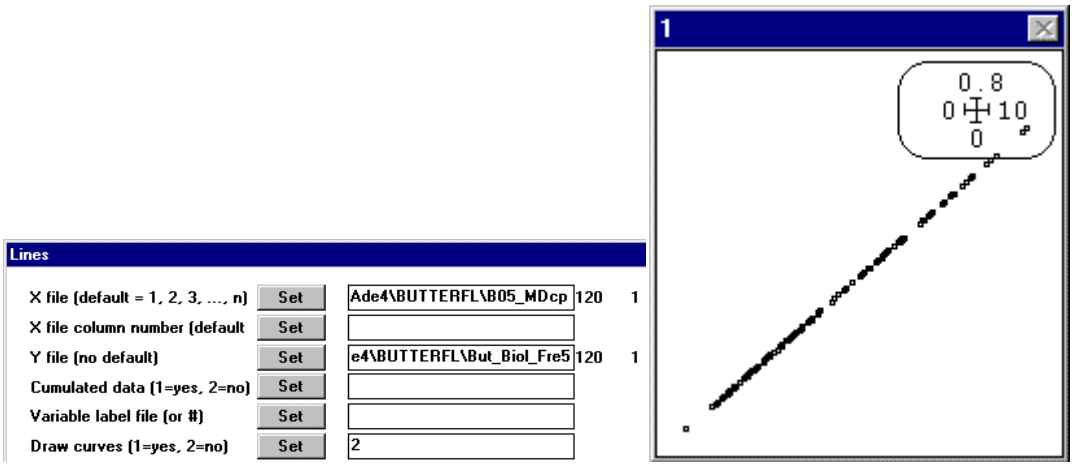



```
Correlation between two distance matrices
First input file: E:\Ade4\BUTTERFL\But_XY_EU.dma
Text file: E:\Ade4\BUTTERFL\But_XY_EU.dma
  1 -> 16
  2 -> 1
  3 -> Classical metric on E:\Ade4\BUTTERFL\But_XY
  4 -> TRUE
Matrix used : 1
Second input file: E:\Ade4\BUTTERFL\But_Biol_Fre5.dma
Text file: E:\Ade4\BUTTERFL\But_Biol_Fre5.dma
  1 -> 16
  2 -> 1
  3 -> EDWARDS on E:\Ade4\BUTTERFL\But_Biol
  4 -> TRUE
Matrix used : 1
Permutation test on r value (Manly 1991 p. 114)
----------------------------------------
Matrix 1 of E:\Ade4\BUTTERFL\But_XY_EU versus matrix 1 of
E:\Ade4\BUTTERFL\But_Biol_Fre5
r index :  5.250e-01
number of random matching: 10000   Observed: 0.524990
Histogramm:  minimum = -0.354353, maximum = 0.696797
number of simulation X<Obs: 9946 (frequency: 0.994600)
number of simulation X>=Obs: 54 (frequency: 0.005400)

   |***
   |*************
   |******************************
   |*********************************************
   |************************************************
   |*****************************************
   |*************************************
   |*********************************
   |******************************
   |**********************
   |****************
   |***********
   |*******
   |***
   |***
o->|*
   |
   |
   |

Matrix     1     of     E:\Ade4\BUTTERFL\But_XY_EU     versus     matrix     1     of
E:\Ade4\BUTTERFL\But_Biol_Fre5
RV index :  4.695e-01
```

```
number of random matching: 10000   Observed: 0.469482
Histogramm:  minimum = 0.005164, maximum = 0.628806
number of simulation X<Obs: 9962 (frequency: 0.996200)
number of simulation X>=Obs: 38 (frequency: 0.003800)

   |*********************
   |************************************************
   |************************************
   |*******************************
   |**********************
   |*****************
   |*************
   |********
   |*****
   |***
   |*
   |*
   |
   |
o->|
   |
   |
   |
   |
   |
```

Les deux tests sont équivalents et les deux stratégies donnent le même test. On vérifie également que la distance issue de l'analyse est la distance calculée directement :


```
Distance matrix computation from a statistical triplet
------------------------------------------
Input file: E:\Ade4\BUTTERFL\B05.cpta
It has 16 rows and 6 columns
Distances are computed among rows
------------------------------------------
Computed distances use the diagonal metric and the centered table of the triplet
Output file: E:\Ade4\BUTTERFL\B05_MDcp
It has 120 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: E:\Ade4\BUTTERFL\B05_MDcp.dma
  1 -> 16
  2 -> 1
  3 -> Euclidean distance from triplet E:\Ade4\BUTTERFL\B05.cpta
  4 -> TRUE
------------------------------------------
```



Le lien entre la distance spatiale et la distance de Edwards est donc le lien implicite entre deux ACP sur les mêmes individus. Il est identique de faire directement l'ACP sur les racines des fréquences ou de calculer la matrice de distances et de faire sa représentation euclidienne. Pour des similarités sur données en 0-1, on verrait de même que le passage par la distance issue de l'indice de Sokal et Michener (voir **13**) :

$$S_2 = \frac{a+d}{n}$$

renvoie à l'ACP centrée du tableau en 0-1 car :

$$d^2(i,j) = 1 - \frac{a+d}{n} = \frac{c+d}{n} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ik} - y_{jk}\right)^2$$

En effet $c+d$ est le nombre de couples de valeurs 0-1 ou 1-0 soit :

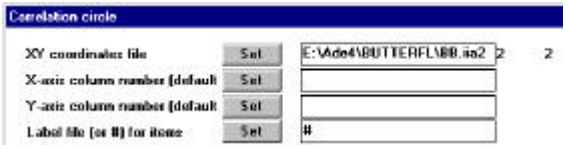
$$c+d = \sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2$$

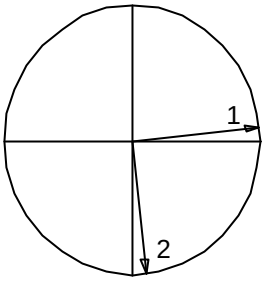
puisque les couples 0-0 et 1-1 donne dans la somme une valeur nulle. La représentation euclidienne d'une matrice de distances d'Edwards ou de Sokal et Michener est une ACP simple implicite. Une classification sur la distance issu de S_2, souvent pratiquée est une classification sur la métrique canonique sur les données binaires.

Sur l'exemple traité, on peut ainsi interpréter le RV entre matrice de distances par la co-inertie des deux ACP :

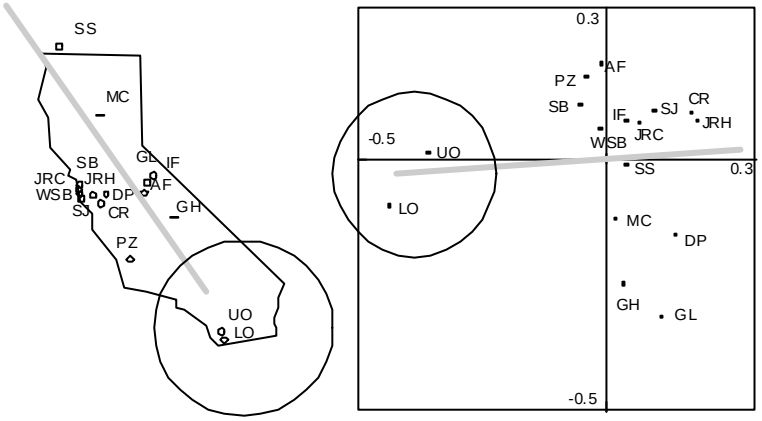

```
Co-inertia analysis between two statistical triplets
1 ---> E:\Ade4\BUTTERFL\But_XY.cpta (rows: 16, col: 2, axes: 2, inertia: 3963.5)
2 ---> E:\Ade4\BUTTERFL\B05.cpta (rows: 16, col: 6, axes: 2, inertia: 10.759)
Co-inertia: 11377, RV coefficient: 0.46948
```

```
E:\Ade4\BUTTERFL\BB.iia2 is a binary file with 2 rows and 2 columns
It contains the coordinates of the projections of inertia
axes onto the co-inertia axes (table 2)
In earlier version of ADE this file uses the extension cp2
File :E:\Ade4\BUTTERFL\BB.iia2
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1| 1.019e-01| 9.816e-01|
|   2|-9.832e-01| 9.706e-02|
|----|----------|----------|
```

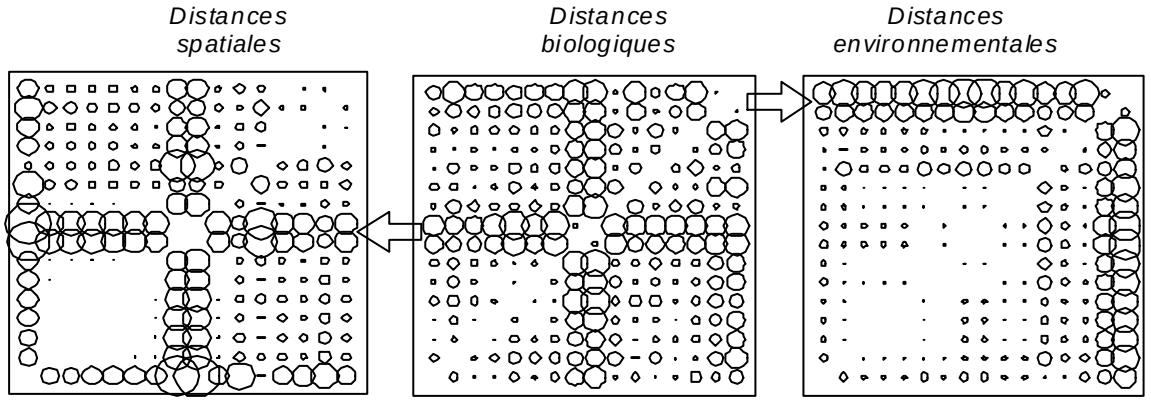




C'est l'axe 1 de l'inertie du nuage des données biologiques qui est maintenant liée aux structures spatiales :



On voit alors clairement les composantes spatiales et environnementales dans la matrice de distances biologiques :



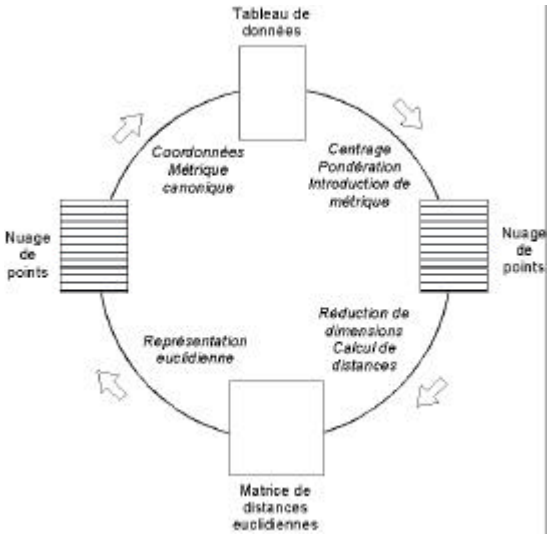
Cet exemple montre que les co-inerties sous-jacentes aux liens entre matrices de distances permettent **d'interpréter** les corrélations significatives du test de Mantel ou du test RV. Le coefficient RV qui permet de faire de la géométrie dans les espaces d'opérateurs est le premier pas

vers la régression multiple ou la régression partielle ([14]) entre matrices de distances par le biais des matrices de produits scalaires.

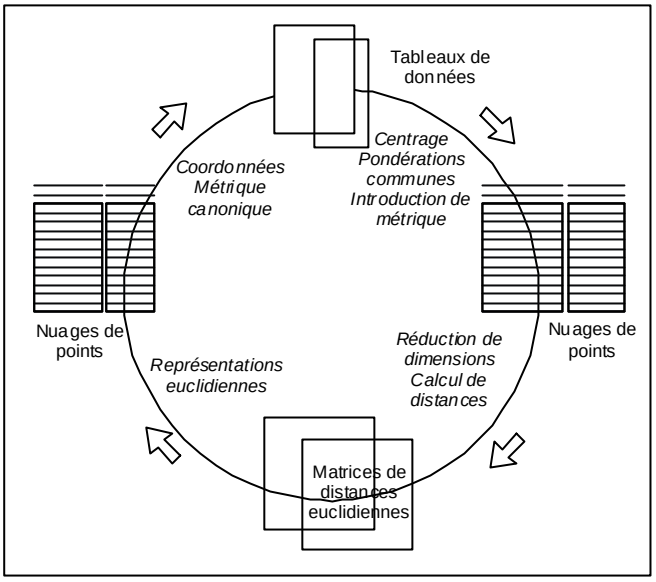
Notons encore que le test RV est basé sur les carrés des distances, valeurs associées dans les espaces euclidiens aux produits scalaires. Manly [15] note :

> *The trouble is that distance measures are often arbitrary to the extent that if thety are changed by a monotonic transformation then the transformed distances may be equally valid.*

C'est particulièrement juste pour le passage de la distance à son carré et c'est pourquoi Dietz [16] propose de passer aux rangs. Nous avons ici une raison de fond pour s'en tenir aux carrés des distances euclidiennes. Le test est unique et s'il est significatif, il est interprétable. Dans ce contexte, on peut résumer la situation par une boucle dans laquelle on peut entrer par le haut (tableau de données) ou le bas (matrice de distances) :



Dans ce schéma, toutes les situations dérivent de la précédente et sont toutes équivalentes. Quand on a deux schémas de ce type le RV rend l'ensemble cohérent :



L'analyse de co-inertie est alors la méthode de couplage naturel de deux matrices de distances et étend l'analyse en coordonnées principales à un couple de tableaux. Il convient de la *Correction for negative values* [17] n'a pas sa place ici. La PCoa (*Principal coordinates analysis*) ne se justifie que sur la base d'un théorème : s'il n'y a pas de valeurs propres négatives, alors il existe un nuage de points dans un espace euclidien, alors les coordonnées principales sont les coordonnées de ces points dans une des représentations euclidiennes possibles. Nous avons vu que, pour des raisons de précision numérique, une matrice peut sembler non euclidienne et que l'approximation par les valeurs propres positives corrigent ce détail, mais l'usage de distances qui ne sont pas euclidiennes par définition n'ont pas à subir cette transformation pragmatique.

Le test RV est alors le même sur les tableaux (test de co-inertie) ou sur les distances (test de Mantel utilisant le RV). Le passage à *K* matrices de données ou *K* matrices de distances est alors possible.

Références

1 Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* : 27, 209-220.

2 Manly, B.F. (1994) *Multivariate Statistical Methods*. A primer. Second edition. Chapman & Hall, London. 1-215.

3 McKechnie, S.W., Ehrlich, P.R. & White, R.R. (1975) Population genetics of Euphydryas butterflies. I. Genetic variation and the neutrality hypothesis. *Genetics* : 81, 571-594.

4 Legendre, L. & Legendre, P. (1984) *Ecologie numérique*. Tome 2 - La structure des données écologiques. Masson, Paris. 2ème édition revue et augmentée : 1-344.

[5] Manly, B.F.J. (1991) *Randomization and Monte Carlo methods in biology*. Chapman and Hall, London. 1-281.

[6] Gower, J.C. & Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* : 3, 5-48.

[7] Escoufier, Y. (1973) Le traitement des variables vectorielles. *Biometrics* : 29, 750-760.

[8] Digby, P. G. N. & Kempton, R. A. . (1987*) Multivariate Analysis of Ecological Communities*. Chapman and Hall, Population and Community Biology Series, London. 1-205. (p.96).

[9] Avise, J.C. (1994) *Molecular markers, natural history and evolution*. Chapman & Hall, London. 1-511 (p. 95).

[10] Spielman, R.S. (1973) Differences among Yanomama Indian villages: do the patterns of allele frequencies, anthropometrics and map locations correspond*?*. *American Journal of Physical Anthropology* : 39, 461-480.

[11] Edwards, A.W.F. (1971) Distance between populations on the basis of gene frequencies. *Biometrics* : 27, 873-881.

[12] Hartl, D.L. & Clark, A.G. (1989) *Principles of population genetics*. Sinauer Associates, Sunderland, Massachussetts. 1-682.

[13] Legendre, L. & Legendre, P. (1984b) *Ecologie numérique*. Tome 2 - La structure des données écologiques. Masson, Paris. 2ème édition revue et augmentée : 1-344. Voir p. 6 et suivantes.

[14] Smouse, P.E., Long, J.C. & Sokal, R.R. (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* : 35, 627-632.

[15] Manly, B.F.J. (1986) Randomization and regression methods for testing for associations with geographical environmental and biological distances between populations. *Researches on Population Ecology* : 28, 201-218.

[16] Dietz, E.J. (1983) Permutation tests for association between two distance matrices. *Systematic Zoology* : 32, 21-26.

[17] Legendre, P. & Anderson, M.J. (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* : 69, 1-24.