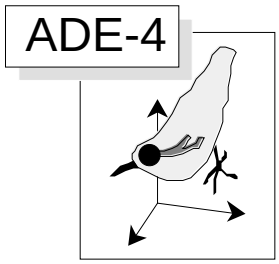


Distances

Distances : Additive constante

Méthode heuristique rendant euclidienne une matrice de distances.

En mathématiques, on appelle **distance** définie sur un ensemble E une fonction *d* de ExE dans R qui vérifie pour tout x, y et z éléments de E :

$$(1) \quad d(x,y) \geq 0$$
$$(2) \quad d(x,y) = 0 \Leftrightarrow x = y$$
$$(3) \quad d(x,y) = d(y,x)$$
$$(4) \quad d(x,y) \leq d(x,z) + d(z,y)$$

En statistiques, on appelle **dissimilarité** définie sur un ensemble fini I à *n* éléments (numérotés 1, 2, ..., *i*, …, *n*) une fonction de IxI dans R qui vérifie pour tout *i* et *j* :

$$(1) \quad \delta_{ij} \geq 0$$
$$(2) \quad \delta_{ii} = 0$$
$$(3) \quad \delta_{ij} = \delta_{ji}$$

En biologie, on utilise le terme de distance pour désigner la différence mesurée entre deux individus, deux populations, deux sites, ... sans se préoccuper de définition. Par exemple la distance génétique de Autem & Bonhomme (1980)[1] utilisée dans Agnese (1989)[2] n'est pas nulle pour deux populations ayant les mêmes profils de fréquences alléliques. 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 (1), symétriques (3), ayant des éléments nuls sur la diagonale (2).

On dit qu'une dissimilarité est euclidienne si elle admet une image euclidienne, c'est-à-dire si on peut associer au point *i* ($1 \leq i \leq n$) un point P*i* d'un espace affine euclidien tel que :

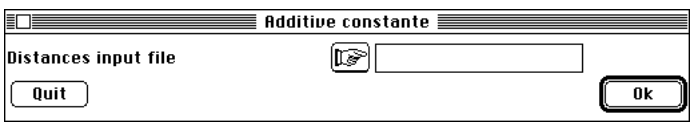
$$\left| P_i P_j \right|_{\mathbf{M}} = \delta_{ji}$$

Si c'est le cas on dit que la matrice de distance est euclidienne. L'intérêt de cette propriété est de pouvoir reconstruire un nuage de *n* points dans R^p (*p* est la dimension de la dissimilarité) dont les distances des points deux à deux sont exactement les dissimilarités observées et éventuellement de représenter ce nuage par projection (Analyse en Coordonnées Principales, PCO).

Si la matrice de distance est calculée à partir d'un tableau de données, on sait souvent si elle est euclidienne ou si elle ne l'est pas. Gower & Legendre (1986)[3] ont étudié un grand nombre de cas utilisés dans ce module. Si elle dérive directement de l'observation, on sait par la présente option si elle est ou non euclidienne.

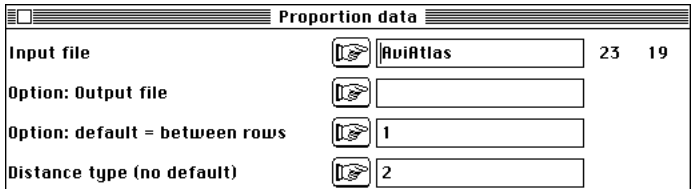
On peut toujours générer une matrice de distance euclidienne à partir d'une matrice de distances qui ne l'est pas en ajoutant une constante positive c à chacune des distances δ_{ij}. Si la matrice de distances n'est pas euclidienne, l'option calcule cette constante, la plus petite possible, avec la solution donnée par Cailliez (1983)[4].

L'option utilise une seule fenêtre de dialogue :



Nom du fichier binaire d'entrée contenant une matrice de distances.

Utiliser la carte Oiseaux de la pile de données. Calculer les distances entre espèces avec l'indice de chevauchement (Distances : Proportion data) :




```
Distance amongst frequency distributions
Input file: AviAtlas
It has 23 rows and 19 columns
Distances are computed among columns

Output file: AviAtlas_Fre2
It has 19 rows and 19 columns
d2 distances computed
Manly 1994 Multivariate statistical methods. A primer
2nd edition. Chapman & Hall 1994. formula 5.8 p. 68
----------------------------------------
```

L'option diagonalise la matrice **W** (notations de Drouet (1989 Chapitre V)[5]:

$$\mathbf{W} = -\frac{1}{2}\mathbf{Q}_1\Delta * \Delta\mathbf{Q}_1^t = \left\{-\frac{1}{2}\delta_{ij}^2\right\}_{..}$$

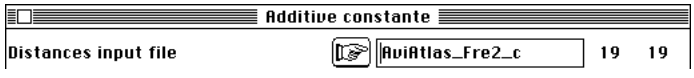
Le double point indique que la matrice est doublement centrée par ligne et par colonne. Le programme édite ses valeurs propres :

```
Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  1.680e+00|002  2.106e-01|003  7.366e-02|004  5.771e-02|
005  3.835e-02|006  1.071e-02|007  8.766e-03|008  3.536e-03|
009  9.177e-04|010 -4.329e-16|011 -1.634e-03|012 -2.090e-03|
013 -5.580e-03|014 -9.346e-03|015 -1.476e-02|016 -3.557e-02|
017 -4.864e-02|018 -7.984e-02|019 -2.137e-01|
```

La présence de valeurs propres négatives indique que cette matrice de distance n'est pas euclidienne. Une matrice 38-38 ($2n$-$2n$) est diagonalisée : sa première valeur propre est la constante recherchée ([4]) :

```
Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  3.727e+00|002  1.576e+00|003  1.279e+00|004  1.210e+00|
005  1.131e+00|006  1.108e+00|007  1.065e+00|008  1.053e+00|
009  1.039e+00|010  1.035e+00|011  1.031e+00|012  1.023e+00|
013  1.017e+00|014  1.008e+00|015  1.007e+00|016  1.004e+00|
017  1.003e+00|018  1.001e+00|019  1.000e+00|020 -2.683e-01|
021 -6.346e-01|022 -7.819e-01|023 -8.262e-01|024 -8.846e-01|
025 -9.027e-01|026 -9.394e-01|027 -9.497e-01|028 -9.623e-01|
029 -9.658e-01|030 -9.702e-01|031 -9.776e-01|032 -9.828e-01|
033 -9.919e-01|034 -9.933e-01|035 -9.963e-01|036 -9.967e-01|
037 -9.993e-01|038 -1.000e+00|

Input file: AviAtlas_Fre2
Output file: AviAtlas_Fre2_c
additive cte:  3.727e+00
```

Remarquer que la nouvelle matrice de distances est euclidienne :




```
Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  2.149e+01|002  1.057e+01|003  8.864e+00|004  8.433e+00|
005  7.873e+00|006  7.702e+00|007  7.391e+00|008  7.306e+00|
009  7.231e+00|010  7.178e+00|011  7.152e+00|012  7.109e+00|
```

```
013  7.069e+00|014  7.001e+00|015  6.987e+00|016  6.954e+00|
017  6.931e+00|018  6.923e+00|019  1.119e-14|

Input file: AviAtlas_Fre2_c
Euclidean distance found inside
Additive constante is not useful
```

1 Autem, M. & Bonhome, F. (1980) Eléments de systématique biochimique chez les Mugilidés (Poissons, Téléostéens) de Méditerranée. *Biochemical Systematics and Ecology* : 8, 305-308.

2 Agnese, J.F. (1989) *Différenciation génétique de plusieurs espèces de Siluriformes ouest-africains ayant un intérêt pour la pèche et l'aquaculture.* Thèse de Doctorat, Université des Sciences et Techniques du Languedoc, Montpellier. 1-194.

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

4 Cailliez, F. (1983) The analytical solution of the additive constant problem. *Psychometrika* : 48, 305-310.

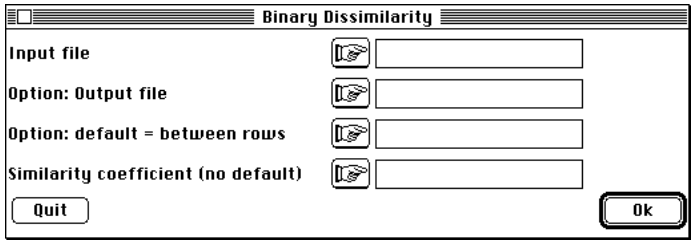
5 Drouet d'Aubigny, G. (1989) *L'analyse multidimensionelle des données de dissimilarité*. Thèse de doctorat, Université Grenoble 1. 1-485.

Distances : Binary Dissimilarity

Utilitaire de calcul de matrices de distance.

L'option calcule des matrices de distance à partir d'indices de similarités sur données binaires. On les utilise pour les tableaux floro-faunistiques en présence-absence.

L'option utilise une seule fenêtre de dialogue :

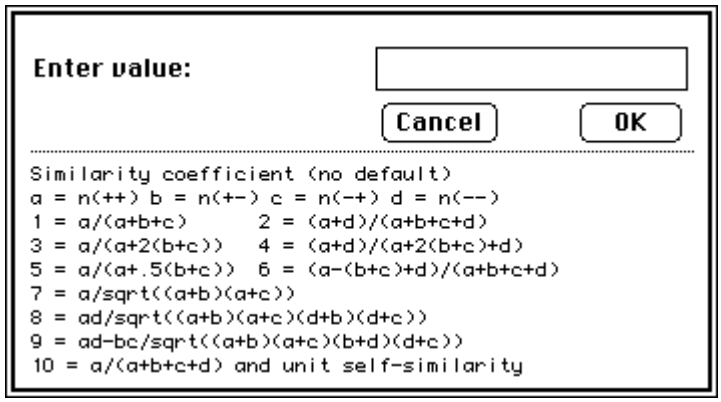


Nom du fichier binaire d'entrée.

Nom du fichier binaire de sortie (création). Par défaut, il est généré avec le nom du fichier d'entrée et des chaînes de caractères qui indiquent les options choisies.

Option de calcul. Utiliser 1 pour calculer les distances entre colonnes du tableau. Par défaut, les distances sont calculées entre lignes.

Option de choix de l'indice de dissimilarité. Il y a 10 options :



Deux objets (en écologie, lignes ou colonnes d'un tableau floro-faunistiques) sont comparés sur une liste de valeurs. Ces valeurs sont réduites en 0-1 (1 si la valeur est strictement positive, 0 sinon). Deux relevés sont ainsi comparés par la liste des espèces présentes, deux espèces sont comparées par la liste des relevés dans lesquels elles sont présentes. Ces listes ont la forme :

01100001010010...
01010001100010...

n est le nombre d'enregistrements, *a* est le nombre de concordance 11, *b* le nombre de concordance 10, *c* le nombre de concordance 01 et *d* le nombre de concordances 00. Ainsi deux espèces sont présentes ensemble dans un même relevé *a* fois, deux relevés possèdent *a* espèces en commun. Les deux objets définissent donc la table de contingence 2-2 :

	1	0	Tot
1	*a*	*b*	*a+b*
0	*c*	*d*	*c+d*
Tot	*a+c*	*b+d*	*n*

Les quatre nombres de la table définissent une similarité entre les deux objets. On peut utiliser :

$$S_1 = \frac{a}{a+b+c}$$ Indice de communauté de Jaccard

$$S_2 = \frac{a+d}{n}$$ Indice de Sokal & Michener

$$S_3 = \frac{a}{a+2(b+c)}$$ Indice de Sokal & Sneath

$$S_4 = \frac{a+d}{a+2(b+c)+d}$$ Indice de Rogers et Tanimoto

$$S_5 = \frac{2a}{2a+b+c}$$ Indice de Sorensen

$$S_6 = \frac{a-(b+c)+d}{n}$$ Indice de Gower & Legendre

$$S_7 = \frac{a}{\sqrt{(a+b)(a+c)}}$$ Indice de Ochiai

$$S_8 = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$ Indice de Sockal & Sneath

$$S_9 = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$ Phi de Pearson

$S_{10} = \frac{a}{n}$ avec l'unité si les deux objets sont identiques

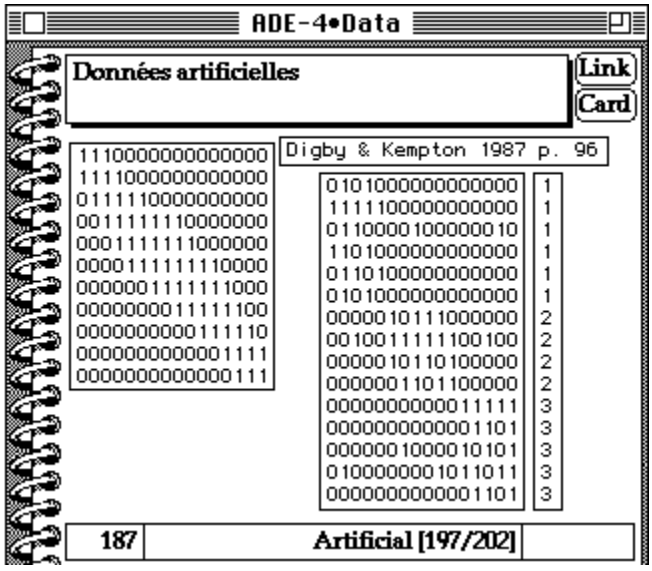
On trouvera les références d'origine dans [1].

Ces indices sont tous inférieurs ou égaux à 1 et la distance associée est définie par :

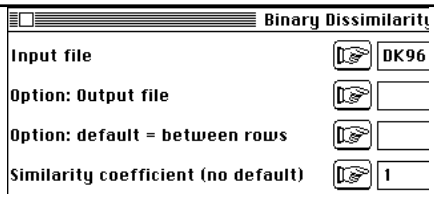
$$D_k = \sqrt{1 - S_k}$$

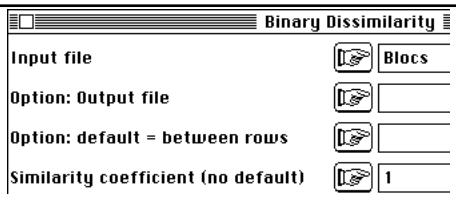
⚠ **Tous les indices retenus donnent des distances euclidiennes [2].**

Utiliser la carte Artificial qui contient deux tableaux artificiels dont les structures sont évidentes :



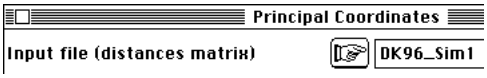
Pour chacun des deux, calculer la matrice de distances entre lignes :

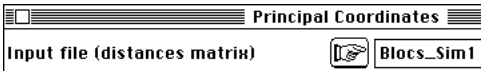


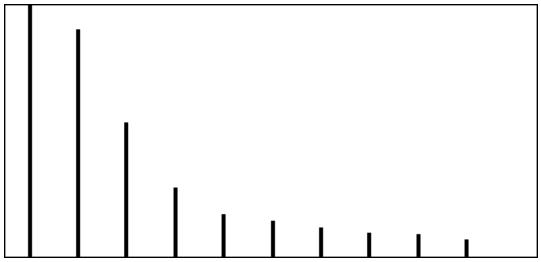


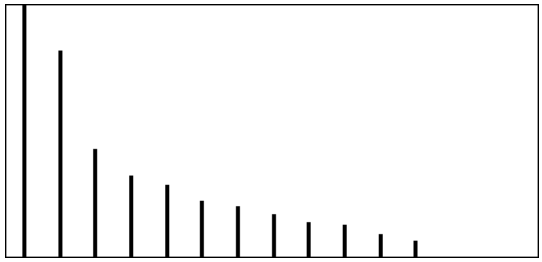

```
Euclidean distance matrix computation from dissimilarity coefficients
Gower J.C. & Legendre P. (1986)
Metric and Euclidean properties of dissimilarity coefficients
Journal of Classification, 3, 5-48

Table 2 p. 23
Input file: DK96
It has 11 rows and 16 columns
Distances are computed among rows

Output file: DK96_Sim1
It has 11 rows and 11 columns
JACCARD index (1901)
S3 coefficient of GOWER & LEGENDRE
Euclidean distance
Distances are computed by
s = a/(a+b+c) --> d = sqrt(1 - s)

Euclidean distance matrix computation from dissimilarity coefficients
Gower J.C. & Legendre P. (1986)
Metric and Euclidean properties of dissimilarity coefficients
Journal of Classification, 3, 5-48

Table 2 p. 23
Input file: Blocs
It has 15 rows and 16 columns
Distances are computed among rows

Output file: Blocs_Sim1
It has 15 rows and 15 columns
JACCARD index (1901)
S3 coefficient of GOWER & LEGENDRE
Euclidean distance
Distances are computed by
s = a/(a+b+c) --> d = sqrt(1 - s)
```

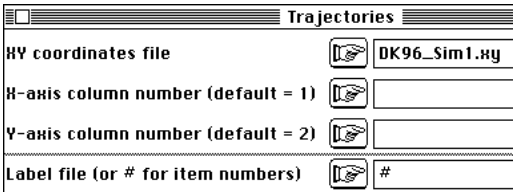
Faire l'analyse en coordonnées principales (Distances : Principal Coordinates). Comparer l'expression des structures fournies par chaque indice.

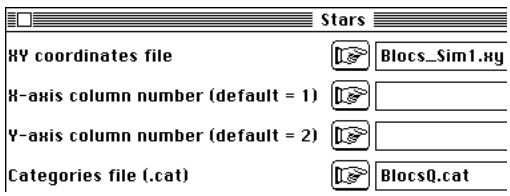




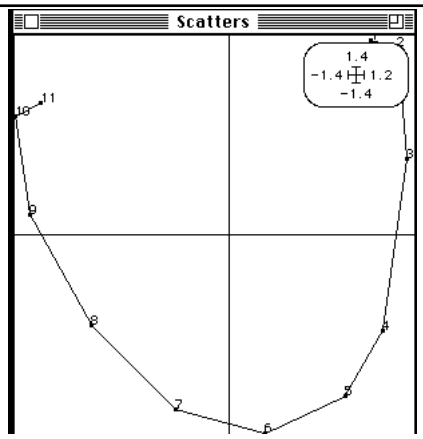


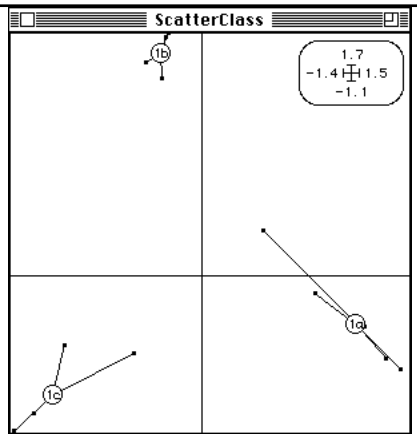




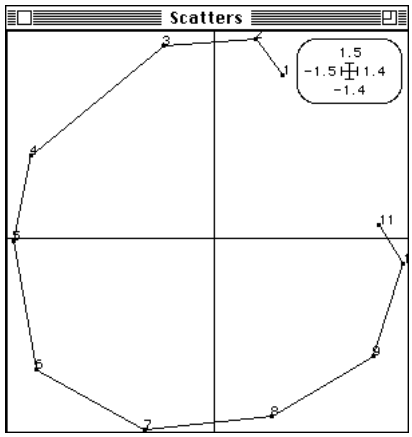


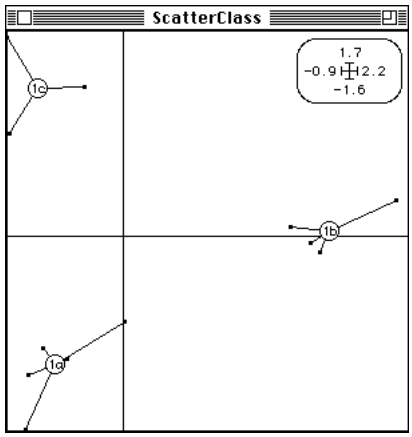
$$S_1 = \frac{a}{a+b+c}$$



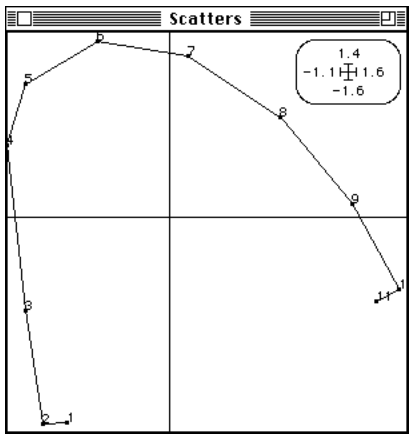


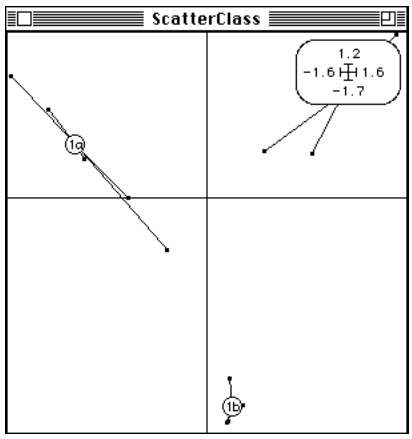
$$S_2 = \frac{a+d}{n}$$



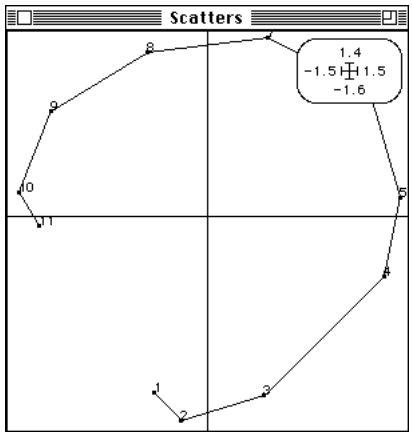


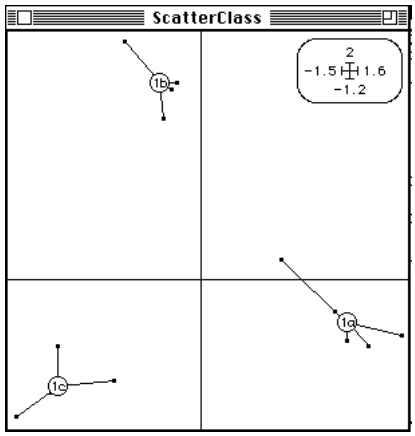
$$S_3 = \frac{a}{a+2(b+c)}$$





$$S_4 = \frac{a+d}{a+2(b+c)+d}$$





•••

La difficulté tient bien sûr au choix d'un indice.

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

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

Distances : Canonical distance

Utilitaire de manipulation de données.

L'objectif est de constituer une matrice de distances à partir d'un tableau de données (métrique canonique, voir [1] p. 57) :

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

L'option utilise une seule fenêtre de dialogue :

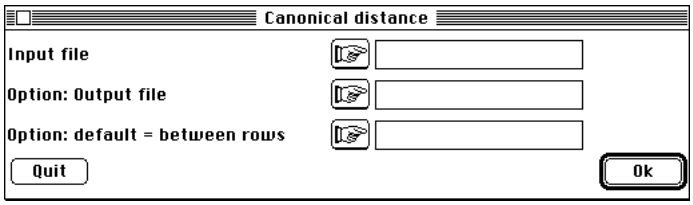


Tableau de données (fichier binaire) avec *n* lignes et *p* colonnes.

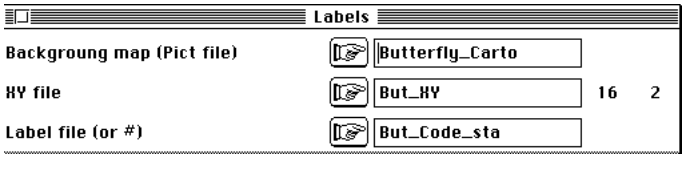
Nom du fichier à créer. Par défaut, il dérive du nom du fichier d'entrée.

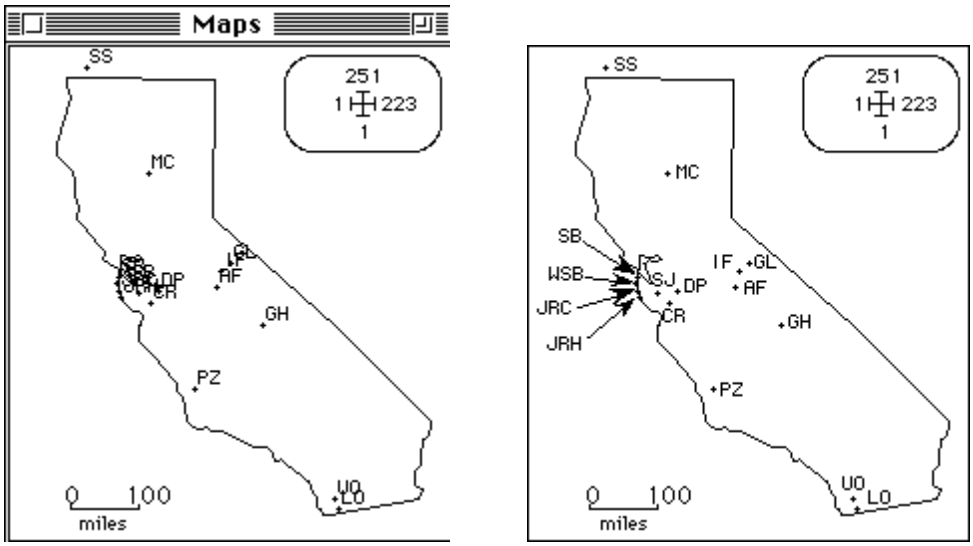
Par défaut le calcul porte sur les distances entre lignes et donne une matrice de sortie *n-n*. Taper 1 pour obtenir une matrice de distances entre colonnes : on obtient alors une matrice de sortie *p-p*.

Utiliser la carte Butterfly[1,2] de la pile ADE-4•Data pour obtenir les fichiers ButEnvir (16-4), But_Code_sta (16 étiquettes) et But_Label_Envir (4 étiquettes).

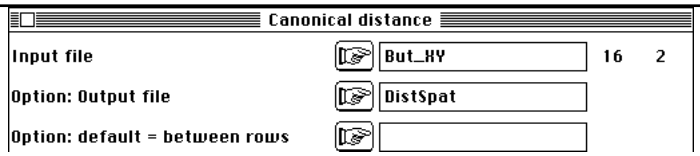
Créer avec la même carte But_XY (16-2), But_Biol (16-6) et But_Code_Biol (6 étiquettes).

Récupérer dans le dossier ADE/Files le fond de carte Butterfly_Carto et vérifier la position des étiquettes sur le fond (Maps : Labels arrangé dans MacDraw™) :





Calculer la matrice des distances spatiales entre stations :




```
Distance matrix computation
----------------------------------------
Input file: But_XY
It has 16 rows and 2 columns
Distances are computed among rows
----------------------------------------
Output file: DistSpat_EU
It has 16 rows and 16 columns
Euclidean distances computed
----------------------------------------
```

Continuer en utilisant le fiche de Distances : Triplet To Distance.

L'exemple présenté dans [1] (Cf. exercice p. 74) pose la question des relations biologique-environnemental sous contrainte spatiale ou encore celle des structures spatiales des caractères biologiques sous contrainte environnementale. Une stratégie consiste à amener les données dans la logique spatiale et donc de constituer des matrices de distances, l'autre consiste à amener l'espace dans la logique des données et donc de constituer des tableaux qui expriment la position spatiale. Le présent module permet d'aborder les deux stratégies.

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

[2] 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.

Distances : Genetic distance

Utilitaire de calcul de matrices de distances.

Cette option calcule des matrices de distances à partir de tableaux de fréquences alléliques. Les lignes du tableau sont des populations. Les colonnes sont regroupées par bloc, chaque bloc étant un locus. Les colonnes d'un même bloc sont les allèles de ce locus. Les cellules du tableau contiennent des nombres positifs ou nuls indiquant la fréquence de rencontre dans la population de chaque allèle de chaque locus (structure de variables floues). Les données sont soit des effectifs (carte de données Chrysichtys **1**), soit des fréquences exprimées en pour cent (carte de données Sicile **2** ou Chevaine **3**) ou en valeur. Dans tous les cas les données sont ramenées en fréquence par blocs (somme unité).

Soit **A** un tableau de fréquences alléliques avec *t* lignes (populations) et *m* colonnes (allèles). Soit *v* le nombre de loci. Le locus *j* a *m*(*j*) allèles.

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

Pour la $i^{\text{ème}}$ ligne et la $k^{\text{ème}}$ modalité de la variable *j*, on note la valeur a_{ij}^{k} ($1 \le i \le t$, $1 \le j \le v$, and $1 \le k \le m(j)$), la valeur du tableau des données brutes. Soit :

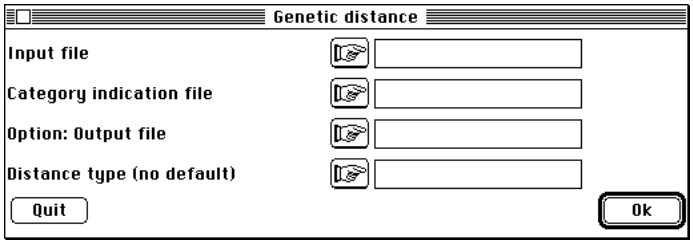
$$a_{ij}^{\bullet} = \sum_{k=1}^{m(j)} a_{ij}^{k} \text{ and } p_{ij}^{k} = \frac{a_{ij}^{k}}{a_{ij}^{\bullet}}$$

Soit le tableau $\mathbf{P} = \left[p_{ij}^{k} \right]$ et les paramètres :

$$p_{ij}^{\bullet} = \sum_{k=1}^{m(j)} p_{ij}^{k} = 1,\ p_{i\bullet}^{\bullet} = \sum_{j=1}^{v} p_{ij}^{\bullet} = v, \text{ and } p_{\bullet\bullet}^{\bullet} = \sum_{i=1}^{t} p_{i\bullet}^{\bullet} = tv$$

L'option calcule des matrices de distances entre populations utilisant les fréquences p_{ij}^{k}.

L'option utilise une seule fenêtre de dialogue :

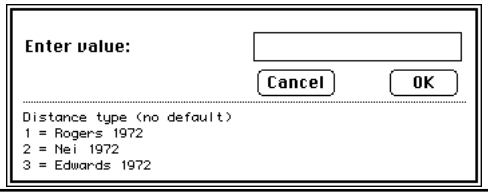


Nom du fichier binaire d'entrée.

Nom du fichier binaire d'indicateur de blocs..

Nom du fichier binaire de sortie. Par défaut, il dérive du nom de fichier d'entrée et du type de distance utilisée.

Type de distance utilisée.



Il y a 3 options :

1 — Distance de Rogers [4] (Voir [5]) :

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

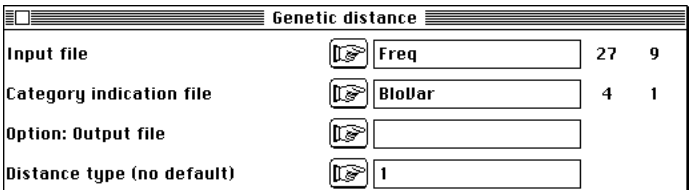
2 — Distance de Nei [6] (Voir [5]) :

$$D_2(a,b) = -Ln\left(\frac{\sum_{k=1}^{v}\sum_{j=1}^{m(k)} p_{aj}^k p_{bj}^k}{\sqrt{\sum_{k=1}^{v}\sum_{j=1}^{m(k)}\left(p_{aj}^k\right)^2}\sqrt{\sum_{k=1}^{v}\sum_{j=1}^{m(k)}\left(p_{bj}^k\right)^2}}\right)$$

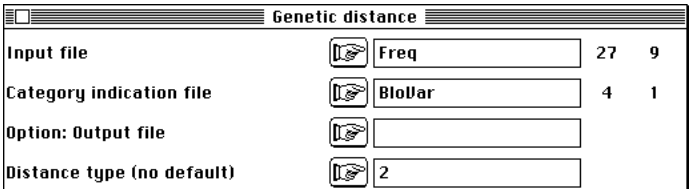
3 — Distance de Edwards [7] (Voir [8]) :

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

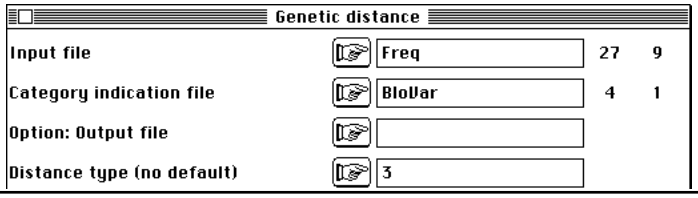
Utiliser la carte Chevaine :




```
Distance amongst multiple frequency distributions
Input file: Freq
It has 27 rows and 9 columns
Bloc indicator: BloVar
Distances are computed among rows

Output file: Freq_Gen1
It has 27 rows and 27 columns
Euclidean distance
d3=mean(sqrt(0.5*Sum(p(i)-q(i)^2)))
Rogers 1972 in Avise 1994 p. 95
```

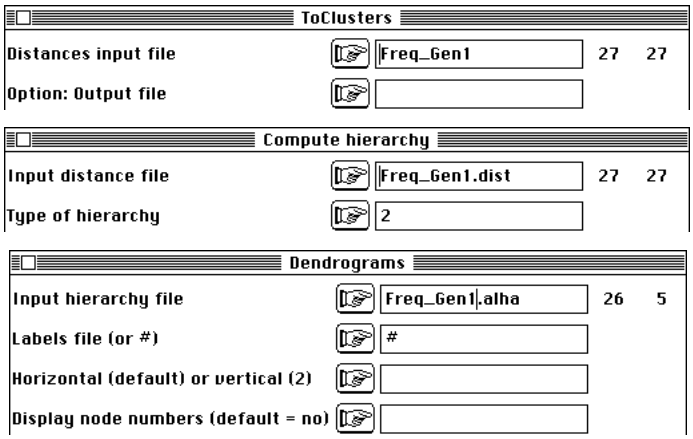


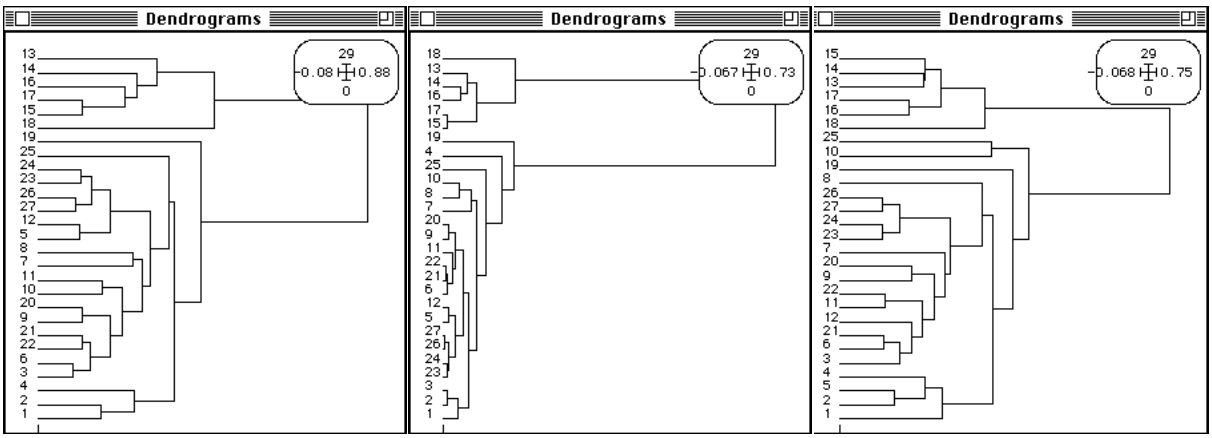

```
Output file: Freq_Gen2
It has 27 rows and 27 columns
d2=-ln( (Sum(p(i)q(i)) / sqrt( Sum(p(i)*p(i))*Sum(q(i)*q(i)) ) )
Nei 1972 in Avise 1994 p. 95
```



```
Output file: Freq_Gen3
It has 27 rows and 27 columns
d3 = sqrt[ Mean (1 - (Sum(sqrt(p(i)q(i))))]
Euclidean distance
Edwards 1971 in Hartl & Clark 1989 multi locus extension
```

Pour chacune des trois distances, utiliser Distances : ToClusters, Clusters : Compute hierarchy et Dendrograms : Dendrograms :





1 Agnese, J.F. (1989) *Différenciation génétique de plusieurs espèces de Siluriformes ouest-africains ayant un intérêt pour la pèche et l'aquaculture*. Thèse de Doctorat, Université des Sciences et Techniques du Languedoc, Montpellier. 1-194.

2 Pigliucci, M. & Barbujani, G. (1991) Geographical patterns of gene frequencies in Italian populatons of Ornithogalum montanum (Liliaceae). *Genetical research*, Cambridge : 58, 95-104.

3 Guinand, B., Bouvet, Y. & Brohon, B. (1996) Spatial aspects of genetic differentiation of the European chub in the Rhone River basin. *Journal of Fish Biology* : 49, 714-726.

4 Rogers, J.S. (1972) Measures of genetic similarity and genetic distances. *Studies in Genetics*, Univ. Texas Publ. 7213: 145-153.

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

6 Nei, M. (1972) Genetic distances between populations. *American Naturalist* : 106. 283-292.

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

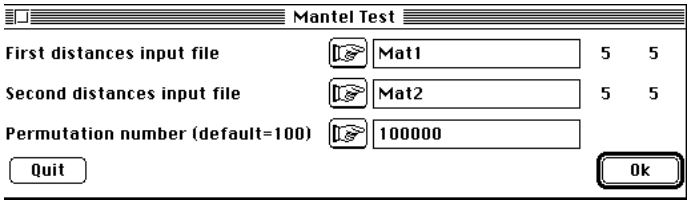
8 Hartl, D.L. & Clark, A.G. (1989) *Principles of population genetics*. Sinauer Associates, Sunderland, Massachussetts. 1-682 (p. 303).

Distances : Mantel Test

Test de permutation pour mesurer la corrélation entre deux matrices de distances.

La statistique utilisée est celle de Mantel[1] mais la randomisation est faite à la machine[2]. Utiliser p. 70-75 de cet ouvrage pour une approche aisée du principe de ce test statistique.

L'option utilise une seule fenêtre de dialogue :



Nom du fichier de la première matrice de distances.

Nom du fichier de la seconde matrice de distances.

Nombre de permutations aléatoires.

Obtenir les fichiers Mat1 (5-5) et Mat2 (5-5) sur la carte Mantel de la pile ADE-4•Data. C'est l'exemple proposé dans [2] p. 73.

```
Correlation between two distance matrices

r index :  9.543e-01
Permutation test (Manly 1994 p. 73)
Test on Z value (formula 5.9 p. 70)
number of random permutattions: 100000   Observed: 9.597700
Histogramm:  minimum = 5.737100, maximum = 9.659451
number of simulations X<Obs: 98334 (frequency: 0.983340)
number of simulations X>=Obs: 1666 (frequency: 0.016660)

    |***********
    |*********************************
    |************************************************
    |*********************************
    |*******************
    |****
    |******
    |
    |********
    |**************
    |**************
    |***********
    |******
    |******
    |****
    |
    |
    |****
    |******
•->|******
```

Voir la fiche Distances : Triplet To Distance Matrix pour un exemple proposé dans [2] d'usage de ce test .

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

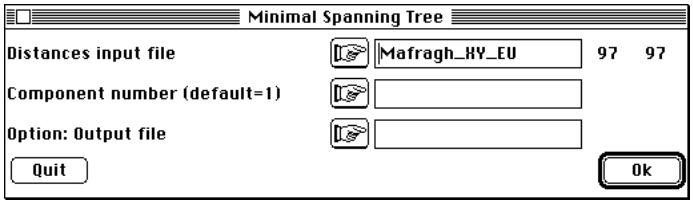
[2] Manly, B.F. (1994) *Multivariate Statistical Methods. A primer*. Second edition. Chapman & Hall, London. 1-215.

Distances : Minimal Spanning Tree

Utilitaire de création d'un graphe de voisinage (arbre de longueur minimum)

Une matrice de distances *n*-*n* définit les distances entre *n* points. On cherche le graphe de voisinage arborescent (connexe sans boucle avec *n*-1 arêtes) qui relie tous les points et dont la somme des distances entre points reliés est minimum. On obtient l'arbre de longueur minimal à une composante (algorithme utilisé dans [1]). En s'interdisant d'employer les arêtes déjà utilisées, on recommence la même opération : on obtient l'arbre de longueur minimale à deux composantes orthogonales[2],...Il est intéressant de représenter l'arbre de longueur minimale sur un plan factoriel pour repérer les déformations associées aux projections [3] (p. 396). L'option permet également de passer d'une matrice de distances à un graphe de voisinage et d'une logique à l'autre.

L'option utilise une seule fenêtre de dialogue :

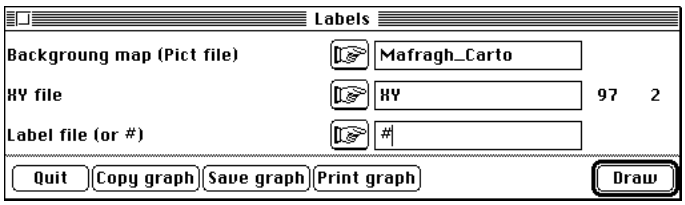


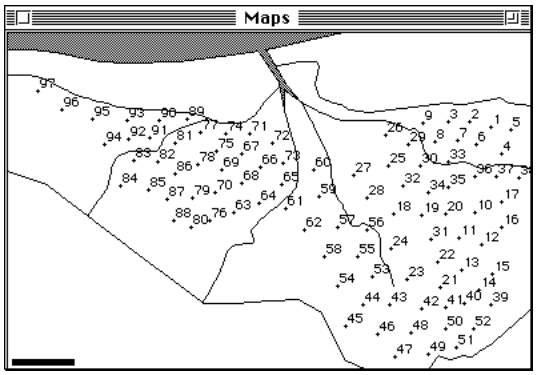
Nom du fichier de la matrice de distances.

Nombre de composantes (par défaut c'est une).

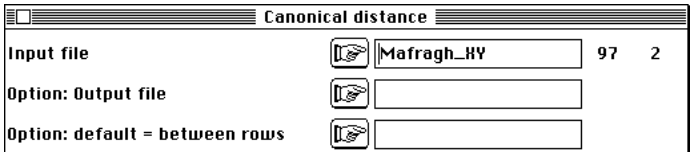
Nom du fichier de sortie (par défaut il dérive du nom de la matrice de distances).

Utiliser la carte Mafragh+3 de la pile ADE-4•Data pour obtenir le fichier de coordonnées Mafragh_XY (97-2) et étiqueter le fond de carte Mafrag_Carto (Maps) :





Calculer la matrice de distances spatiales entre points :

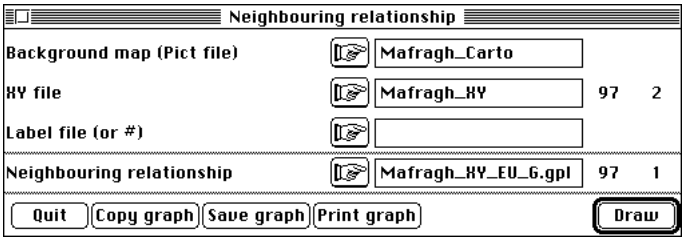


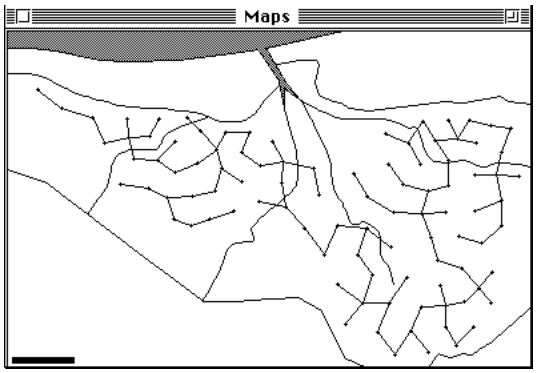
```
Distance matrix computation
-------------------------------------------
Input file: Mafragh_XY
It has 97 rows and 2 columns
Distances are computed among rows
-------------------------------------------
Output file: Mafragh_XY_EU
It has 97 rows and 97 columns
Euclidean distances computed
-------------------------------------------
```

Calculer l'arbre de longueur minimale à une composantes orthogonale :

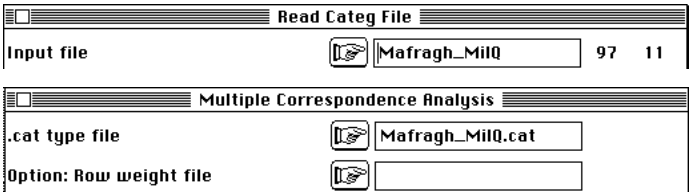
```
Neighboorhing relationship from Minimal Spanning Tree
Input file (distances matrix): Mafragh_XY_EU
Rank: 1
Neighboorhing relationship in text file: Mafragh_XY_EU_G
It contains graph matrix (LEBART's M) with 97 rows and columns
Neighboorhing weights in binary file: Mafragh_XY_EU_G.gpl
It contains  97 rows and 1 column
```

Représenter ce graphe sur le fond de carte (Maps) :

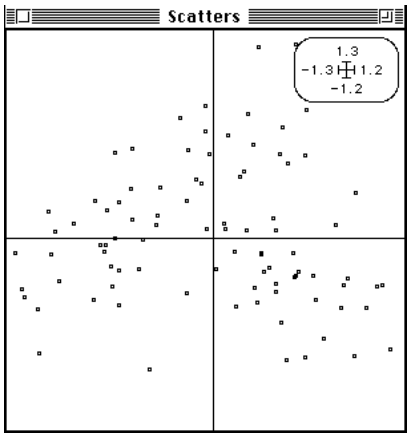


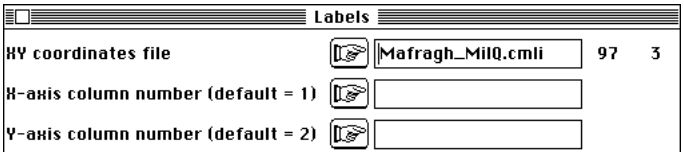


Faire l'ACM du tableau Mil issu de la carte Mafragh+2 avec l'enchaînement :

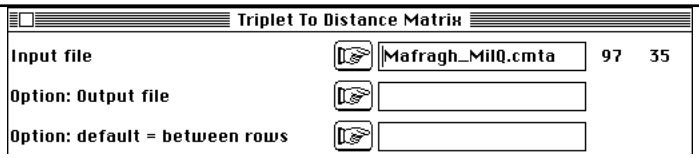


Tracer la carte factorielle (Scatters) :

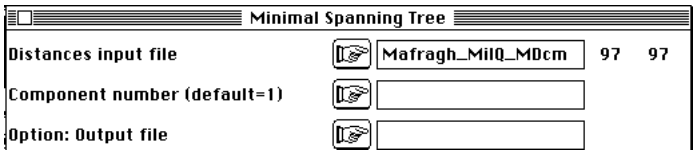




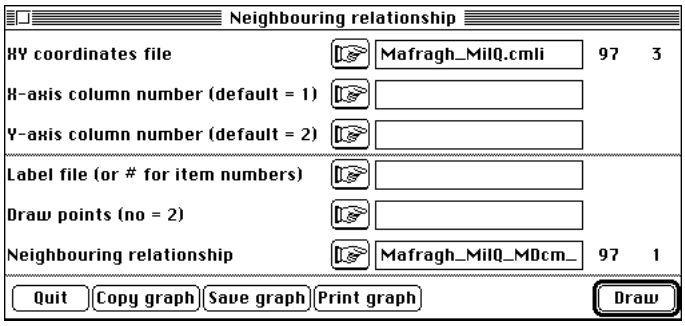
Calculer les distances entre points au sens de l'analyse :

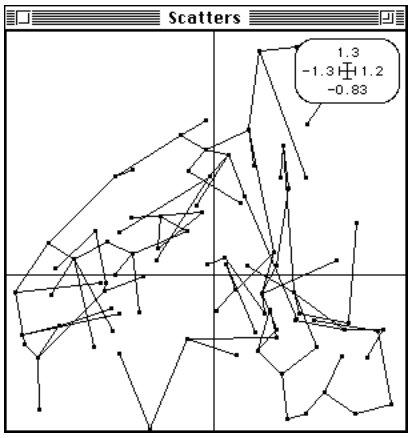


Calculer l'arbre de longueur minimal au sens de cette distance :



Représenter cet arbre sur la carte factorielle (Scatters):





Cette opération oriente fortement la lecture du plan factoriel.

[1] Kevin, V. & Whitney, M. (1972) Algorithm 422. Minimal Spanning Tree [H]. *Communications of the Association for Computing Machinery* : 15, 4, 273-274.

[2] Friedman, J.H. & Rafsky, L.C. (1979) Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* : 7, 697-717.

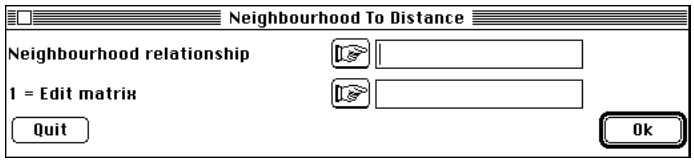
[3] Lebart, L., Morineau, A. & Fenelon, J.P. (1982) *Traitement des données statistiques. Méthodes et Programmes*. Dunod, 2° édition, Paris. 1-518.

Distances : Neighbourhood To Distance

Utilitaire de calcul de la distance associée à un graphe de voisinages.

Si **M** est la matrice d'un graphe connexe, on peut simplement définir la distance entre deux sommets comme le nombre d'arêtes minimum d'un chemin reliant ces deux sommets ([1] p. 191).

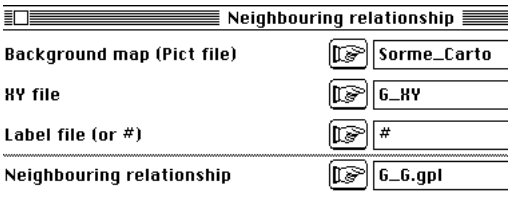
L'option utilise une seule fenêtre de dialogue :

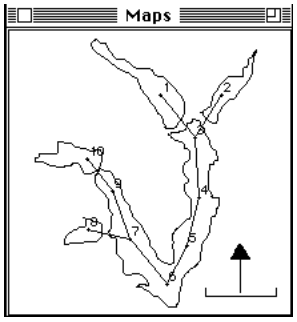


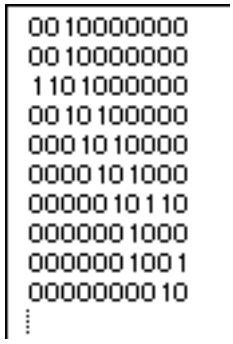
Nom du fichier .gpl d'accès au graphe de voisinage.

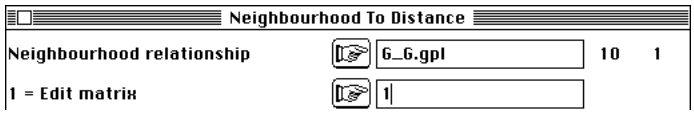
Option d'édition : taper 1 pour éditer la matrice de distances.

Utiliser la carte Sorme. Digitaliser les 10 stations sur le fond de carte (Digit : Digitize), vérifier les positions (Maps : Labels), observer la taille de la carte (MapUtil : GetPictSize), envoyer l'information au logiciel Cabri-Graph (NGUtil : To_Cabri_Graph), Tracer le graphe de voisinage à la souris et renvoyer l'information (NGUtil : From_Cabri_Graph), vérifier (Maps : Neighbouring Relationship) :










```
Multiscale neighbourhood relation from neighboorhood graph
Input file (neighbourhood graph): G_G.gpl
Multiscale neighbourhood graph in binary file: G_G.gms
Aij = k <=> arcs of shortest length between i and j have k edges
 -  -  -  -  -  -  -  -  -  -
  •  2  1  2  3  4  5  6  6  7
  2  •  1  2  3  4  5  6  6  7
  1  1  •  1  2  3  4  5  5  6
  2  2  1  •  1  2  3  4  4  5
  3  3  2  1  •  1  2  3  3  4
  4  4  3  2  1  •  1  2  2  3
  5  5  4  3  2  1  •  1  1  2
  6  6  5  4  3  2  1  •  2  3
  6  6  5  4  3  2  1  2  •  1
  7  7  6  5  4  3  2  3  1  •
 -  -  -  -  -  -  -  -  -  -
```

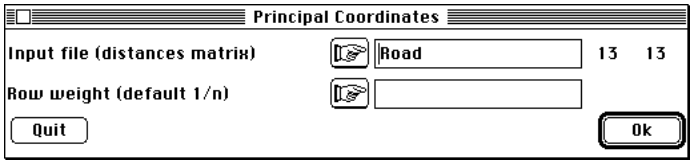

[1] Berge, C. (1967) *Théorie des graphes et ses applications.* Dunod, Paris. 1-269.

Distances : Principal Coordinates Analysis

Méthode d'ordination utilisant une matrice de distance (Gower, J.C., 1966, Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* : 53, 325-338).

L'analyse permet d'obtenir des représentations euclidiennes (cartes factorielles représentant des configurations d'objets) à partir d'une matrice de distances entre ces objets. Elle est présentée dans [1] (§ 3.5, p. 83 et suivantes) ou [2] (§ 12.3, p. 190 et suivantes). L'abréviation la plus utilisée est PCO. La matrice de distances entre objets est transformée par $d(i,j) \mapsto -d^2(i,j)/2$, doublement centrée et diagonalisée. Les composantes des vecteurs propres sont les coordonnées des objets. L'opération est totalement valide si et seulement si les valeurs propres sont toutes positives ou nulles [3]. Elle est simplement utile dans le cas contraire. La présentation et la discussion de Digby & Kempton ([1]) est recommandée avant l'emploi de cette technique.

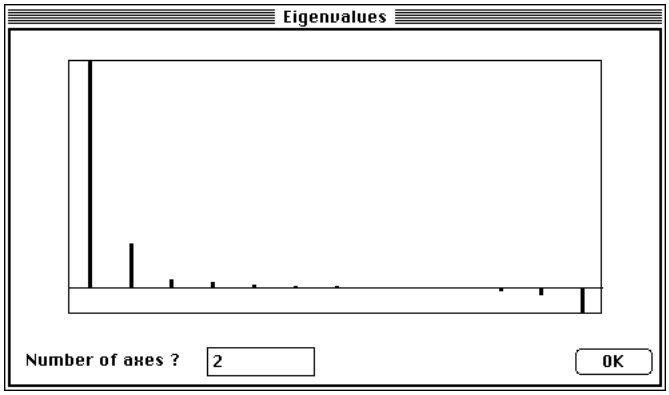
L'option utilise une seule fenêtre de dialogue :



Nom de fichier contenant une matrice de distances (*n*-*n*) entre *n* objets.

Nom de fichier contenant une pondération des *n* objets (par défaut, pondération uniforme).

Utiliser la carte Zealand de la pile ADE-4•Data pour obtenir le fichier Road (13-13) contenant une matrice des distances routières entre 13 villes ([2], tableau 11.1) et le code des villes.



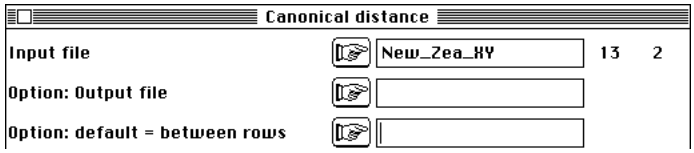
On obtient :

```
File Road.pp contains the matrix aij - ai. -a.j + a..
with aij = -d2ij/2
--- It has 13 rows and 13 columns

Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  5.154e+04|002  9.858e+03|003  1.864e+03|004  1.154e+03|
005  4.882e+02|006  2.970e+02|007  1.055e+02|008 -1.749e-12|
009 -5.116e+01|010 -2.132e+02|011 -6.403e+02|012 -1.718e+03|
013 -5.979e+03|

File Road.vp contains the eigenvalues
--- It has 13 rows and 1 column

File Road.xy1 contains the principal coordinates (norm=1)
--- It has 13 rows and 2 columns
```

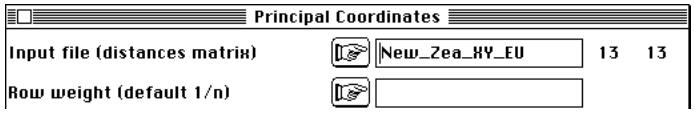
```
File :Road.xy1
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.863e+00| 1.456e+00|
|   2|-1.755e+00| 2.406e+00|
|----|----------|----------|

File Road.xy contains the principal coordinates  (norm=sqrt(lambda))
--- It has 13 rows and 2 columns
File :Road.xy
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-4.230e+02| 3.306e+02|
|   2|-1.743e+02| 2.389e+02|
|----|----------|----------|
```

Le double centrage impose la valeur propre nulle et le centrage des composantes des vecteurs propres. Il existe des valeurs propres négatives, ce qui indique que les distances entre villes ne sont pas des distances euclidiennes.

Digitaliser le fond de carte New_Zealand_Digit (dossier ADE/Files) ou utiliser le fichier New_Zea_XY et calculer les distances ordinaires entre les villes (“à vol d’oiseau”) :

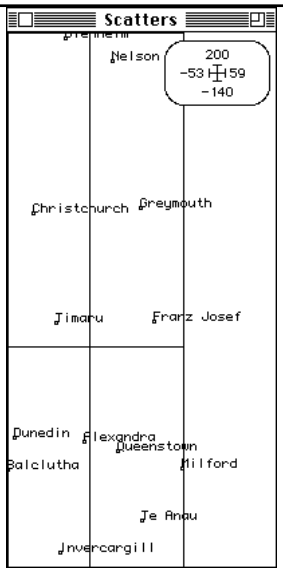


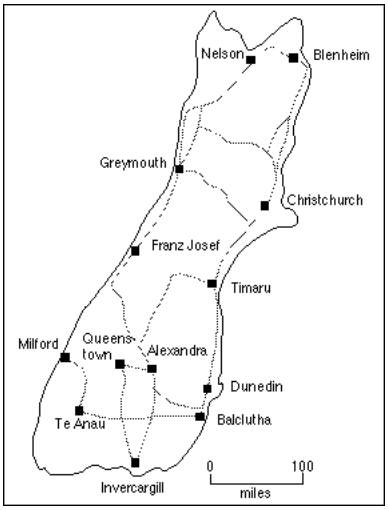
Exécuter le même analyse sur la nouvelle matrice de distance :

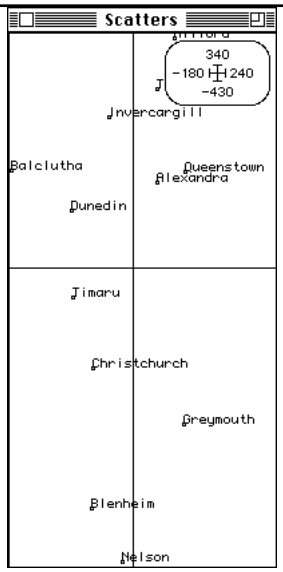



```
File New_Zea_XY_EU.pp contains the matrix aij - ai. -a.j + a..
with aij = -d2ij/2
--- It has 13 rows and 13 columns

Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  1.073e+04|002  1.166e+03|003  4.172e-04|004  2.099e-04|
005  1.259e-04|006  5.280e-05|007  1.525e-05|008 -9.249e-13|
009 -3.586e-05|010 -6.786e-05|011 -8.623e-05|012 -1.995e-04|
013 -3.920e-04|

File New_Zea_XY_EU.vp contains the eigenvalues
--- It has 13 rows and 1 column

File New_Zea_XY_EU.xy1 contains the principal coordinates (norm=1)
--- It has 13 rows and 2 columns
File :New_Zea_XY_EU.xy1
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.262e+00| 1.892e+00|
|   2|-1.528e+00| 1.713e+00|
|----|----------|----------|

File New_Zea_XY_EU.xy contains the principal coordinates
(norm=sqrt(lambda))
--- It has 13 rows and 2 columns
File :New_Zea_XY_EU.xy
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.308e+02| 1.960e+02|
|   2|-5.220e+01| 5.849e+01|
|----|----------|----------|
```

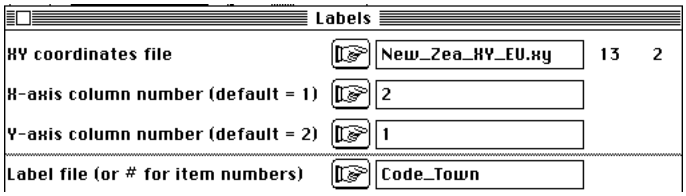
Noter qu’il n’y a que deux valeurs propres non nulles et positives (distance euclidienne dans un espace de dimension deux). Représenter les cartes des individus des deux analyses (Scatters) et comparer avec la carte initiale (ci-dessous, au milieu) :



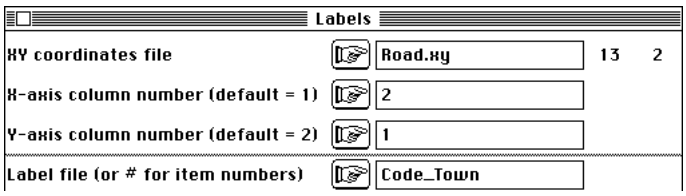




Ci-dessus à gauche :



Ci-dessus à droite :



On n'oubliera pas que le signe d'un vecteur propre n'a aucune signification (si **v** est un vecteur propre normé -**v** l'est aussi, et on trouve l'un ou l'autre au hasard). Dans le fichier .xy on a les vecteurs propres normés. Dans .xy1 on a les vecteurs normés à la valeur propre (coordonnées de variance lambda).

L'analyse en coordonnées principales est une forme élémentaire des méthodes de positionnement multiple ("*Multidimensional scaling*"). Appliquée à une matrice de distances euclidiennes (Distances : Triplet To Distance Matrix), elle donne les cartes des analyses ordinaires.

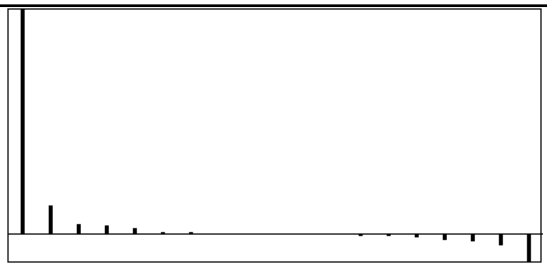
Appliquée à une matrice de distances spatiales (à deux dimensions Distances : Canonical distance, elle redonne la carte à une rotation près.

Quand on dispose d'un vrai tableau de données, on peut calculer deux matrices de distances entre lignes et entre colonnes, ce qui conduit à deux typologies par ce procédé. Seule une analyse linéaire de la première couche donne les deux approches simultanées et coordonnées.

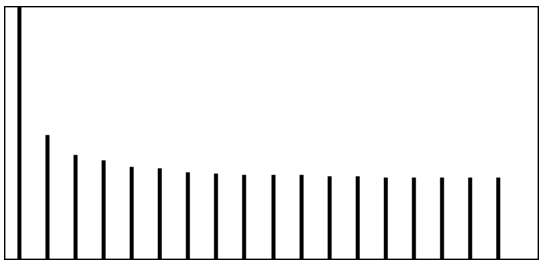
Il faut bien identifier le fait que, en partant de la matrice de distance (PCO), l'origine des cartes factorielles obtenues fait partie des résultats de la démarche ([4]) et non des données de départ comme dans le cas d'une ACP.

Utiliser l'exemple mise en place dans Distances : Additive constante :

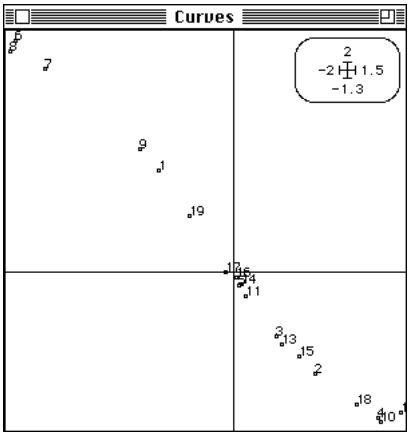
La matrice n'étant pas celle d'une distance euclidienne, il y a des valeurs propres négatives :

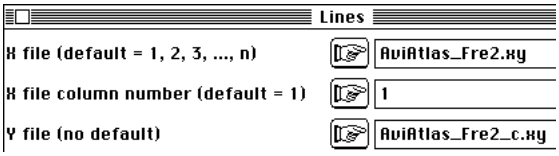


La matrice étant celle d'une distance euclidienne, il n'y a pas de valeurs propres négatives :



Comparer les deux premiers vecteurs propres (Curves : Lines) :





Ils sont très proches sans être identiques.

L'option diagonalise la matrice $\mathbf{WD}_p$ (notations de Drouet (1989 p. 167)[4] avec :

$$\mathbf{W} = -\frac{1}{2}\mathbf{Q}_1\Delta * \Delta\mathbf{Q}_1^t = \left[-\frac{1}{2}\delta_{ij}^2\right]$$

Nous avons suivi l'auteur en introduisant une pondération arbitraire sur les objets (lignes et colonnes) de la matrice de distances. Le double centrage se fait donc avec cette pondération et les codes numériques sont centrés, normés et non corrélés pour cette pondération. On peut hésiter entre la diagonalisation de $\mathbf{WD}_p$ et celle de $\mathbf{WD}_p\mathbf{WD}_p$ préférée par Drouet (op. cit. p. 168). Le module utilise la première, l'introduction de la constante additive permettant de se débarrasser des valeurs propres négatives si cela semble utile. La comparaison des résultats entre AFC des matrices de voisinages ([5]) et diagonalisation de l'opérateur de Moran (NGStat : Moran EigenVectors) ([6]) milite plutôt pour la diagonalisation de $\mathbf{WD}_p$ et l'utilisation des opérateurs non positifs ([7] [8]).

L'analyse en coordonnées principales (PCO) est un moyen simple et efficace de ramener une matrice de distance dans la logique des tableaux de données comme NGStat : Moran EigenVectors ramène un graphe de voisinage dans la logique des tableaux de données.

Toute analyse de tableaux définit des matrices de distances (Distances : Triplet To Distance) et il y a dans ce module de nombreuses options qui définissent des matrices de distances à partir de tableaux de données.

On peut enfin passer de graphe de voisinages à matrice de distance (Distances : Neighbourhood To Distance) et de matrices de distance à graphes de voisinages (Distances : Minimal Spanning Tree). Il y a donc un ensemble très vaste de possibilités d'interaction entre tableaux, graphes de voisinages et matrices de distances. Il ne faut pas confondre l'intérêt fondamental des matrices de distances dans les disciplines qui les mesurent (comme dans le contrôle des produits alimentaires par des jurys de dégustation) et le jeu qui consiste à les créer avec des tableaux de données comme c'est souvent le cas en écologie.

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

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

[3] Young, G. & Householder, A.S. (1938) Discussions of a set of points in terms of their mutual distances. *Psychometrika* : 3, 19-22.

[4] Drouet d'Aubigny, G. (1989) *L'analyse multidimensionelle des données de dissimilarité.* Thèse de doctorat, Université Grenoble 1. 1-485.

[5] Lebart, L. (1984) Correspondence analysis of graph structure. Bulletin technique du CESIA, Paris : 2, 1-2, 5-19.

[6] Thioulouse, J., Chessel, D. & Champely, S. (1995) Multivariate analysis of spatial patterns: a unified approach to local and global structures. Environmental and Ecological Statistics : 2, 1-14.

[7] Torre, F. & Chessel, D. (1994) Co-structure de deux tableaux totalement appariés. *Revue de Statistique Appliquée* : 43, 109-121.

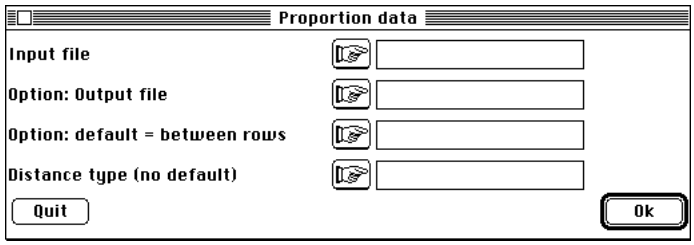
[8] Chessel, D. & Sabatier, R. (1993) Couplage de triplets statistiques et graphes de voisinage. In : *Biométrie et Données spatio-temporelles*. Asselain, B. & Coll. (Ed.) Société Française de Biométrie, ENSA, Rennes. 28-37.

Distances : Proportion data

Utilitaire de calcul de matrices de distances.

Un tableau **X** ne contient que des nombres positifs ou nuls et est considéré comme définissant des distributions de fréquences par ligne ou par colonne. L'option calcule des matrices de distances entre ces profils.

L'option utilise une seule fenêtre de dialogue :

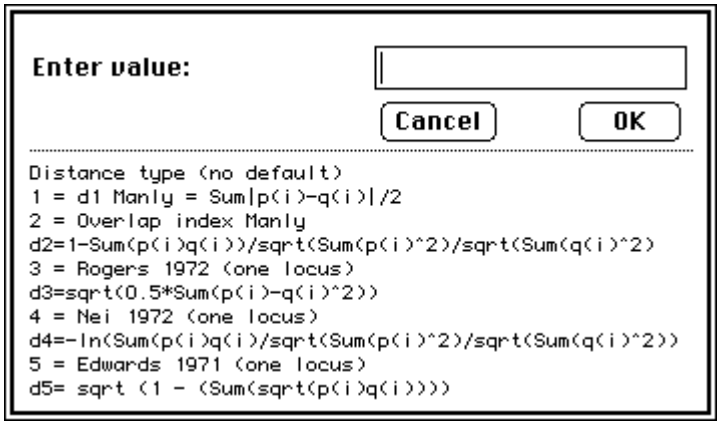


Nom du fichier binaire d'entrée.

Nom du fichier binaire de sortie (création). Par défaut, il dérive du nom du fichier d'entrée et de l'option de calcul choisi.

Option de calcul. Utiliser 1 pour calculer les distances entre colonnes du tableau. Par défaut, les distances sont calculées entre lignes.

Option de choix de l'indice de distance. Il y a 5 options :



1 — Distance d_1 (voir **1** formule (5.7) p. 68).

$$d_1 = \sum_i |p_i - q_i| / 2$$

2 — Distance d_2 (indice de chevauchement de niche, voir **1** formule (5.8) p. 68).

$$d_2 = 1 - \sum_i p_i q_i \Big/ \left\{ \sqrt{\sum_i p_i^2 \sum_i q_i^2} \right\}$$

3 — Distance de Rogers (Voir Distances : Genetic distance) :

$$d_3 = \frac{1}{2} \sum_i (p_i - q_i)^2$$

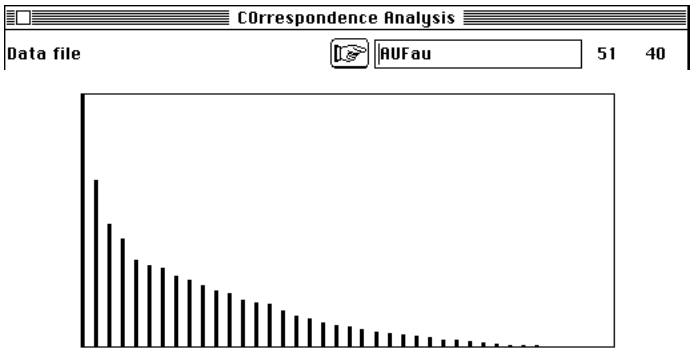
4 — Distance de Nei (Voir Distances : Genetic distance) :

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

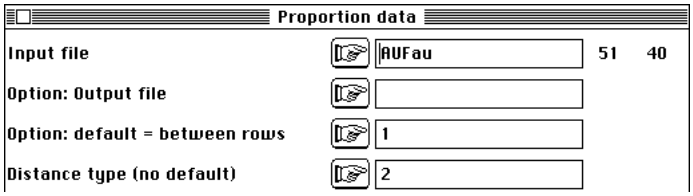
5 — Distance de Edwards (Voir Distances : Genetic distance) :

$$d_5 = \sqrt{1 - \sum_i \sqrt{p_i q_i}}$$

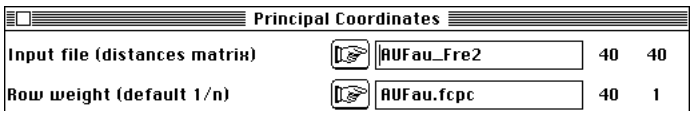
Utiliser la carte AviUrba. Faire l'analyse des correspondances du tableau faunistique :

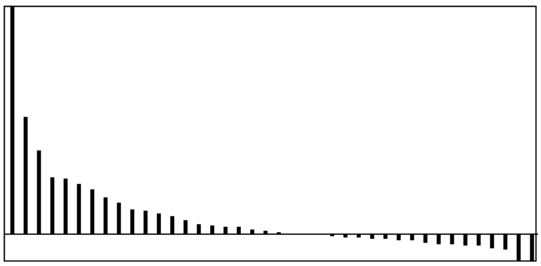


Garder quatre facteurs. Calculer la distance en terme de chevauchements de niche entre espèces :



Faire l'analyse en coordonnées principales avec la même pondération :



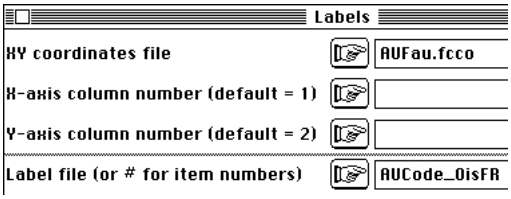


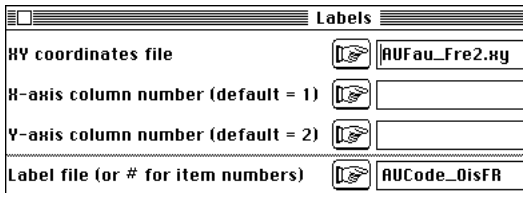
Garder trois axes.

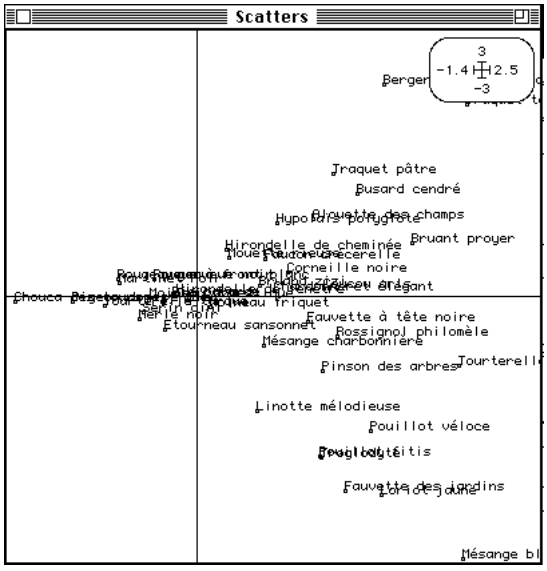
```
File AUFau_Fre2.pp contains the matrix aij - ai. -a.j + a..
with aij = -d2ij/2
--- It has 40 rows and 40 columns

Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  6.478e-02|002  3.315e-02|003  2.352e-02|004  1.593e-02|
005  1.572e-02|006  1.415e-02|007  1.270e-02|008  1.030e-02|
009  8.920e-03|010  6.851e-03|011  6.684e-03|012  5.568e-03|
013  5.135e-03|014  3.967e-03|015  2.494e-03|016  2.189e-03|
017  1.891e-03|018  1.752e-03|019  1.041e-03|020  7.224e-04|
021  3.692e-04|022 -2.175e-17|023 -2.724e-05|024 -1.274e-04|
```

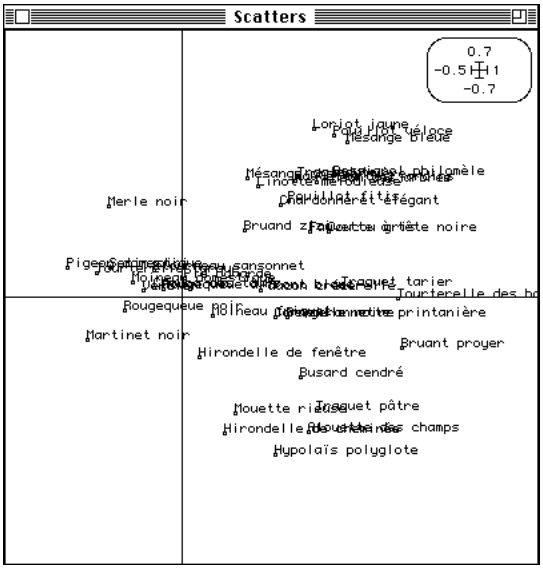
```
025 -2.787e-04|026 -6.114e-04|027 -7.294e-04|028 -1.067e-03|
029 -1.125e-03|030 -1.406e-03|031 -1.476e-03|032 -2.443e-03|
033 -2.667e-03|034 -2.824e-03|035 -2.907e-03|036 -3.141e-03|
037 -3.656e-03|038 -4.173e-03|039 -7.181e-03|040 -7.827e-03|

File AUFau_Fre2.vp contains the eigenvalues
--- It has 40 rows and 1 column

File AUFau_Fre2.xy1 contains the principal coordinates (norm=1)
--- It has 40 rows and 3 columns
File :AUFau_Fre2.xy1
|Col.|   Mini    |   Maxi    |
|----|-----------|-----------|
|   1|-1.272e+00| 2.438e+00|
|   2|-2.287e+00| 2.416e+00|
|   3|-1.884e+00| 2.826e+00|
|----|-----------|-----------|

File AUFau_Fre2.xy contains the principal coordinates  (norm=sqrt(lambda))
--- It has 40 rows and 3 columns
File :AUFau_Fre2.xy
|Col.|   Mini    |   Maxi    |
|----|-----------|-----------|
|   1|-3.236e-01| 6.204e-01|
|   2|-4.163e-01| 4.399e-01|
|   3|-2.889e-01| 4.333e-01|
|----|-----------|-----------|
```

Comparer les deux ordinations.









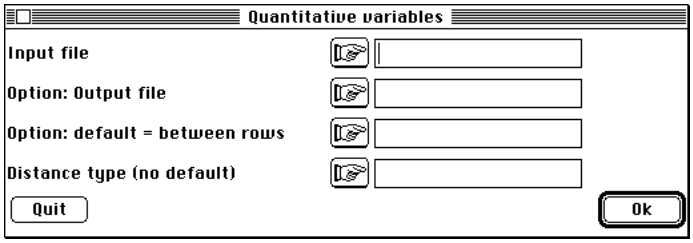
[1] Manly, B.F. (1994) *Multivariate Statistical Methods. A primer. Second edition*. Chapman & Hall, London. 1-215.

Distances : Quantitative variables

Utilitaire de calcul de matrices de distances pour variable quantitative.

Un tableau **X**, avec *n* lignes et *p* colonnes, contient des valeurs quantitatives $\mathbf{X} = \left[x_{ij}\right]$. L'option calcule une matrice des distances entre lignes ou entre colonnes avec un critère au choix parmi 7 possibilités choisies dans **1**.

L'option utilise une seule fenêtre de dialogue :

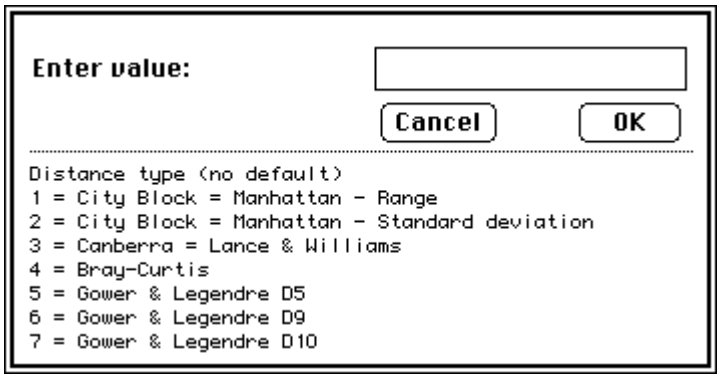


Nom du fichier binaire d'entrée.

Nom du fichier binaire de sortie (création). Par défaut ce nom dérive de celui du fichier d'entrée et de l'option choisie.

Option de calcul. Utiliser 1 pour calculer les distances entre colonnes du tableau. Par défaut, les distances sont calculées entre lignes.

Option de choix de l'indice de dissimilarité. Il y a 7 options :



1 — Distance de Manhattan, ou city block, ou de Gower 1971a (référence p. 20 dans **2**), ou D_3 de Gower & Legendre **1** (distance non euclidienne) :

$$d_1(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{\left|x_{ik} - x_{jk}\right|}{r_k} \quad \text{avec} \quad r_k = \underset{i=1}{\overset{n}{Max}}\left(x_{ik}\right) - \underset{i=1}{\overset{n}{Min}}\left(x_{ik}\right)$$

2 — Distance de Manhattan, ou de Cain & Harrison (référence p. 20 dans **2**), ou D_3 de Gower & Legendre **1** (distance non euclidienne) :

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

3 — Distance de Canberra, ou de Lance & Williams (référence p. 20 dans **2**), ou D_7 de Gower & Legendre **1** (distance non euclidienne) :

$$d_3(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{\left|x_{ik} - x_{jk}\right|}{\left|x_{ik}\right| + \left|x_{jk}\right|}$$

4 — Distance de Bray-Curtis ou de Odum (références p. 20 dans **2**), ou D_8 de Gower & Legendre **1** (distance non euclidienne) :

$$d_4(i,j) = \frac{\frac{1}{p}\sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right|}{\sum_{k=1}^{p}\left(x_{ik} + x_{jk}\right)}$$

5 — Distance D_5 de Gower & Legendre **1** pour données positives seulement (distance euclidienne) :

$$d_5(i,j) = \frac{1}{p}\sum_{k=1}^{p}\frac{\left(x_{ik} - x_{jk}\right)^2}{\left(x_{ik} + x_{jk}\right)^2}$$

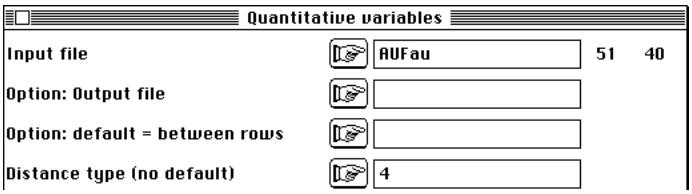
6 — Distance D_9 de Gower & Legendre **1** pour données positives seulement (distance non euclidienne) :

$$d_6(i,j) = \frac{\sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right|}{\sum_{k=1}^{p}\underset{i,j}{Max}\left(x_{ik}, x_{jk}\right)}$$

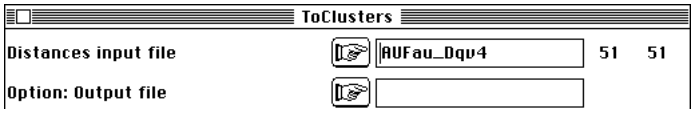
7 — Distance D_{10} de Gower & Legendre **1** pour données positives seulement (distance non euclidienne) :

$$d_7(i,j) = \frac{1}{p}\sum_{k=1}^{p}\left(1 - \frac{\underset{i,j}{Min}\left(x_{ik}, x_{jk}\right)}{\underset{i,j}{Max}\left(x_{ik}, x_{jk}\right)}\right)$$

Utiliser la carte AviUrba :

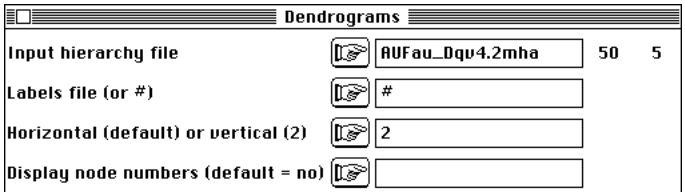


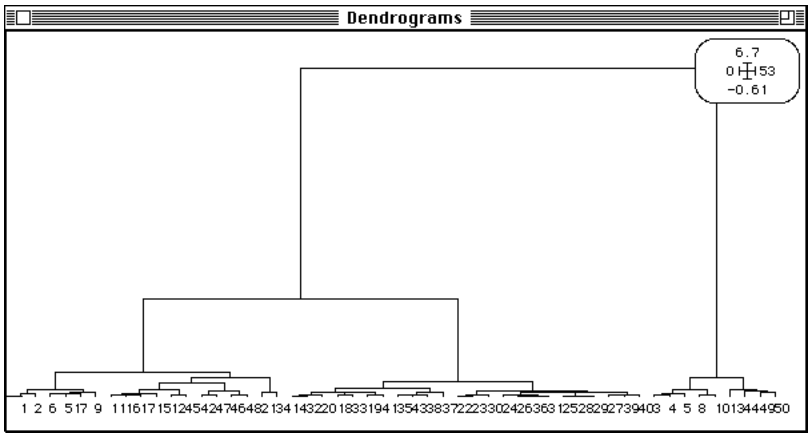

```
Distance matrix computation from dissimilarity coefficients
Dissimilarity coefficients amongst quantitative variables
Gower J.C. & Legendre P. (1986)
Metric and Euclidean properties of dissimilarity coefficients
Journal of Classification, 3, 5-48
Table 3 p. 27

Input file: AUFau
It has 51 rows and 40 columns
Distances are computed among rows

Output file: AUFau_Dqv4
It has 51 rows and 51 columns
Bray-Curtis
D8 coefficient of GOWER & LEGENDRE
Non Euclidean distance
Distances are computed by
dij = (1/p)Sum|xik-xjk|/(Sum(xik+xjk) 1<=k<=p
```



Dans Clusters :

Dans Dendrograms :





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

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

Distances : ToClusters

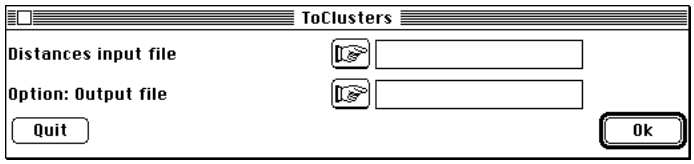
Utilitaire d'interface entre modules.

Le module Distances permet de calculer des distances entre lignes ou colonnes des tableaux ou des triplets. Certaines option du module Clusters utilise des matrices de distances dans des fichiers du type ---.dist. L'option crée ces fichiers en assurant la transformation de valeurs :

$$d_{ij} \mapsto \frac{d_{ij} - min_{i,j}\left(d_{ij}\right)}{max_{i,j}\left(d_{ij}\right) - min_{i,j}\left(d_{ij}\right)}$$

Les distances sont simplement ramenées dans l'intervalle [0,1] pour des questions numériques.

L'option utilise une seule fenêtre de dialogue :



Nom du fichier binaire d'entrée (matrice de distances).

Nom du fichier binaire de sortie (création d'un fichier ---.dist). Par défaut on utilise le nom de fichier d'entrée.

Toute matrice de distances calculées par les options :

1 — Distances : Binary Dissimilarity (distances basées sur une dissimilarité binaire, 10 options)
2 — Distances : Quantitative variables (distances basées sur des variables quantitatives, 7 options)
3 — Distances : Proportion data (distances basée sur des dissimilarités entre distributions de fréquences, 5 options)
4 — Distances : Genetic distance (distances génétiques pour fréquence alléliques, 3 options)
5 — Distances : Triplet To Distance (distances euclidiennes associées à tout triplet statistique, sur les lignes et/ou les colonnes , *x* options)
6 — Distances : Neighbourhood To Distance (distance basée sur un graphe de voisinage, 1 option)
7 — Distances : Canonical distance (distance de la géométrie ordinaire, 1 option)

fournit une partition ou une hiérarchie et un dendrogramme en passant par la présente option.

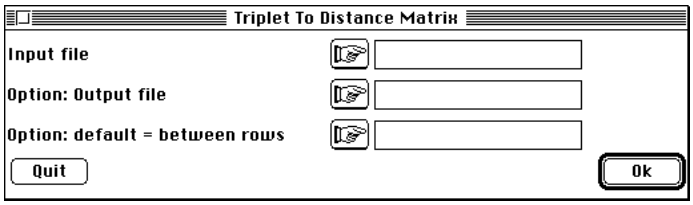
Voir Clusters : Compute distances.

Distances : Triplet To Distance

Utilitaire de création de matrices de distances à partir de triplets statistiques quelconques.

Un triplet statistique définit deux métriques diagonales qui donnent deux matrices de distances euclidiennes, respectivement entre les lignes et entre les colonnes du tableau transformé.

L'option utilise une seule fenêtre de dialogue :

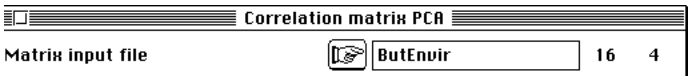


Fichier binaire du type ---.##ta, ## définissant le type de l'analyse. *n* est le nombre de lignes et *p* est le nombre de colonnes. Tous les types d'analyses de la première couche conviennent.

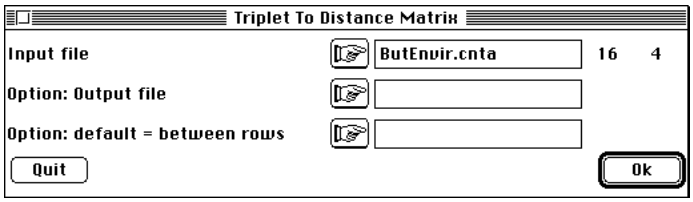
Nom du fichier de sortie. Par défaut, il est défini à partir du fichier d'entrée.

Par défaut le calcul porte sur les distances entre lignes et donne une matrice de sortie *n-n*. Taper 1 pour obtenir une matrice de distances entre colonnes : on obtient alors une matrice de sortie *p-p*.

Utiliser les données mises en place dans la fiche de Distances : Canonical distance. Faire l'ACP normée du tableau ButEnvir (PCA) :

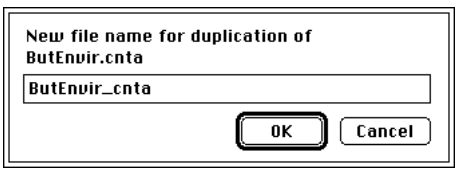


Calculer la matrice de distances entre lignes :

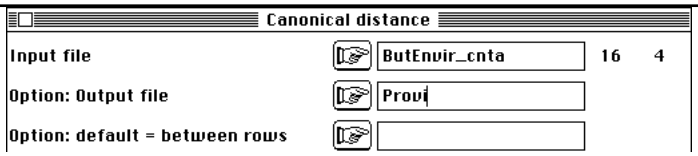



```
Distance matrix computation from a statistical triplet
------------------------------------------
Input file: ButEnvir.cnta
It has 16 rows and 4 columns
Distances are computed among rows
------------------------------------------
Output file: ButEnvir_MDcn
It has 16 rows and 16 columns
Computed distances use the diagonal metric and the centered table of the triplet
------------------------------------------
```

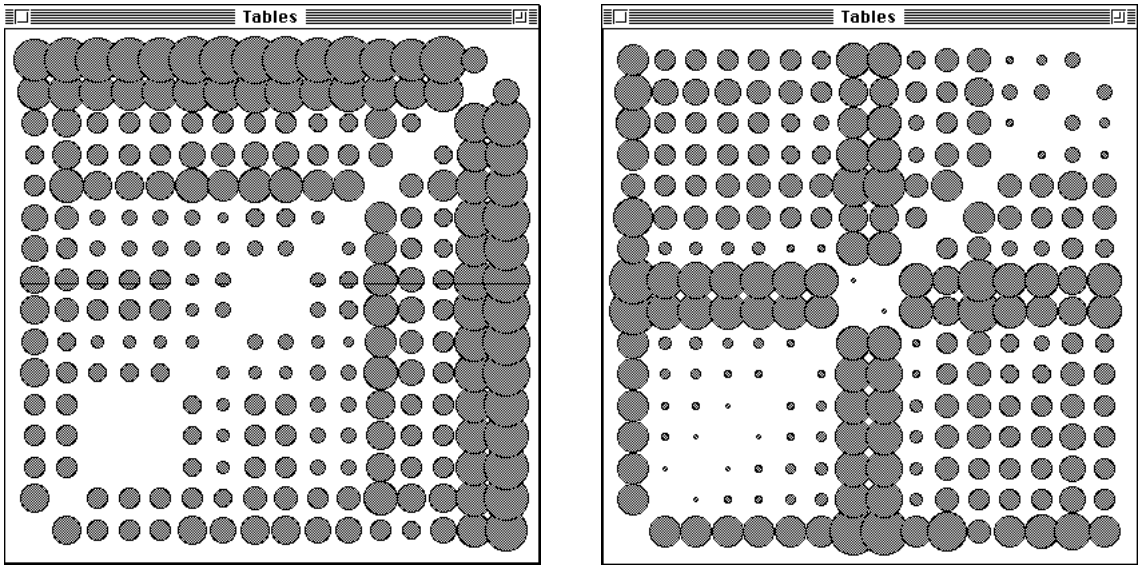
Dupliquer le fichier normalisé (pour éliminer le point d'extension) :



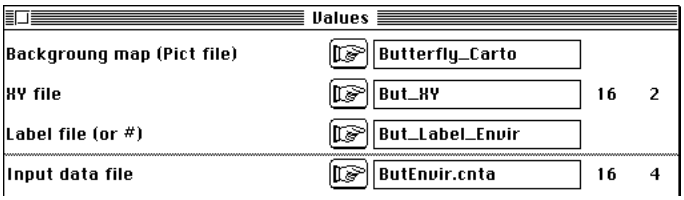
La pondération des colonnes étant unitaire dans une ACP normée vérifier qu'on obtient le même résultat avec l'option :

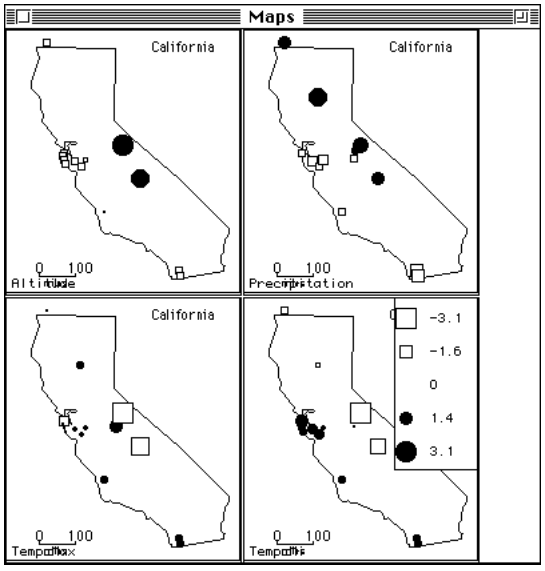


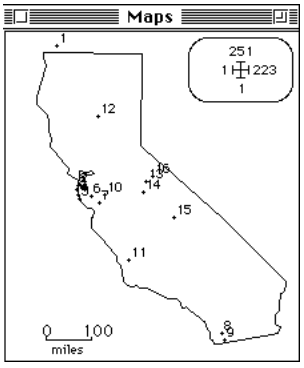
Les fichiers Provi_EU (distances euclidiennes classiques sur tableau normalisé) et ButEnvir_MDcn (distances entre lignes associées au triplet de l'ACP normée) ont le même contenu. Représenter par Tables : Values par la matrice de distances environnementales (à gauche) et la matrice de distances spatiales (à droite) :



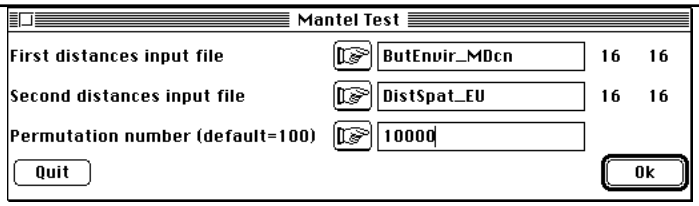
Interpréter celle de gauche par l'originalité des stations d'altitude (15 et 16) et celle de droite par l'éloignement considérable des stations Nord (1) et Sud (8 et 9) :





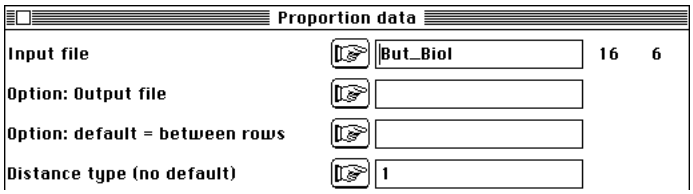


Tester la corrélation entre matrice de distance spatiale (Distances : Canonical Distance Matrix) et matrice de distance environnementale (ci-dessus) pat Distances : Mantel Test :



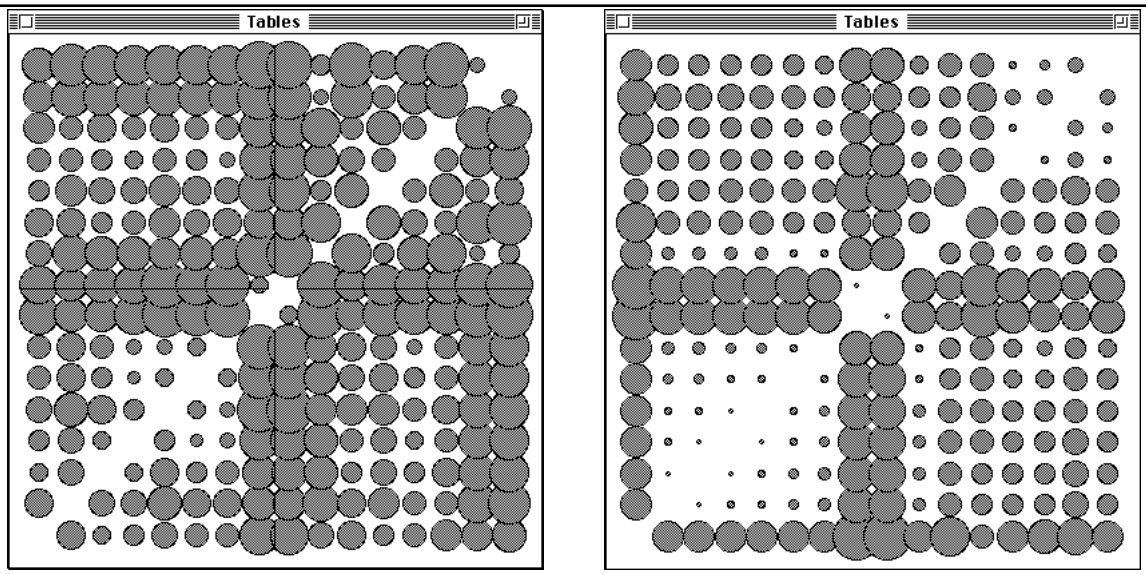

```
Correlation between two distance matrices
----------------------------------------
First input file: ButEnvir_MDcn
It has 16 rows and 16 columns
Second input file: DistSpa_EU
It has 16 rows and 16 columns
----------------------------------------
r index :  1.008e-01
Permutation test (Manly 1994 p. 73)
Test on the Z value (formula 5.9 p. 70)
number of random permutations: 10000   Observed: 21915.744141
Histogram:  minimum = 17261.015625, maximum = 29460.966797
number of simulations X<Obs: 7253 (frequency: 0.725300)
number of simulations X>=Obs: 2747 (frequency: 0.274700)

    |******
    |******************************
    |************************************************
    |*********************************************
    |************************************
    |*************************
    |***********************
•-> |***************
    |*************
    |****************
    |******************
    |************
    |*******
    |****
    |***
    |**
    |*
    |*
    |*
    |
```

On trouve logiquement une absence complète de signification. Calculer alors les distances génétiques entre populations :

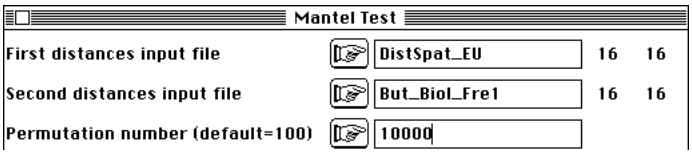



```
Distance amongst frequency distributions
Input file: But_Biol
It has 16 rows and 16 columns
Distances are computed among rows

Output file: But_Biol_Fre1
It has 16 rows and 16 columns
d1 distances computed
Manly 1994 Multivariate statistical methods. A primer
2nd edition. Chapman & Hall 1994. formula 5.7 p. 68
----------------------------------------
```

On répète la même opération avec la matrice des distances génétiques (à gauche) et spatiales (à droite) :

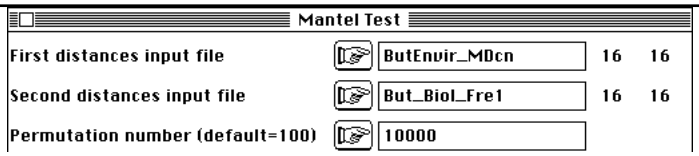


Tester la corrélation entre matrice de distance spatiale et matrice de distances génétiques (Distances : Mantel Test) :




```
Correlation between two distance matrices
-----------------------------------------
First input file: DistSpat_EU
It has 16 rows and 16 columns
-----------------------------------------
Second input file: But_Biol_Fre1
It has 16 rows and 16 columns
-----------------------------------------
r index :  4.683e-01
Permutation test (Manly 1994 p. 73)
Test on Z value (formula 5.9 p. 70)
number of random permutations: 10000   Observed:  3.348e+03
Histogram:  minimum =  2.488e+03, maximum =  3.600e+03
number of simulations X<Obs: 9932 (frequency:  9.932e-01)
number of simulations X>=Obs: 68 (frequency:  6.800e-03)

    |*
    |***********
    |****************************
    |*******************************************
    |************************************************
    |*********************************************
    |*******************************************
    |************************************************
    |************************************
    |*****************************
    |********************
    |***************
    |**********
    |****
    |**
•-> |**
    |*
    |
    |
    |
```

La liaison est très forte. Tester la corrélation entre la matrice de distances environnementales et la matrice de distances biologiques :




```
Correlation between two distance matrices
----------------------------------------
First input file: ButEnvir_MDcn
It has 16 rows and 16 columns
----------------------------------------
Second input file: But_Biol_Fre1
It has 16 rows and 16 columns
----------------------------------------
r index :  4.354e-01
Permutation test (Manly 1994 p. 73)
Test on Z value (formula 5.9 p. 70)
number of random permutations: 10000   Observed:  1.041e+02
Histogram:  minimum =  7.827e+01, maximum =  1.138e+02
number of simulations X<Obs: 9861 (frequency:  9.861e-01)
number of simulations X>=Obs: 139 (frequency:  1.390e-02)

    |***
    |******************
    |************************************
    |*************************************************
    |*************************************************
    |*********************************************
    |**********************************
    |************************
    |**********************
    |***********************
    |*****************
    |************
    |*********
    |*****
•-> |**
    |*
    |*
    |
    |
    |
```

La liaison est encore significative. La matrice des distances génétiques comporte deux composantes, l'une spatiale liée au stations Nord et Sud, l'autre environnementale liée aux stations d'altitude. La représentation graphique des matrices de distances et l'approche inférentielle (test de Mantel) sont ici fortement cohérentes. L'exercice est formulé par Manly[1] (p. 74) : la solution apportée modifie sensiblement la question 4 (p. 75).

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