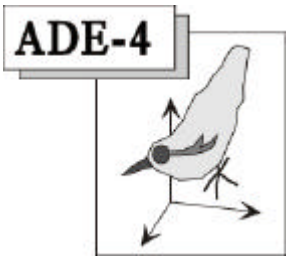


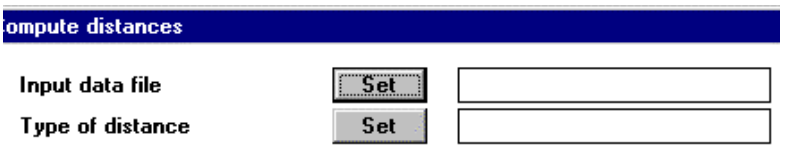
Clusters

Clusters : Compute distances

Type Utilitaire de calcul de matrices de distances.

Objet L'option calcule une matrice de distances entre objets.

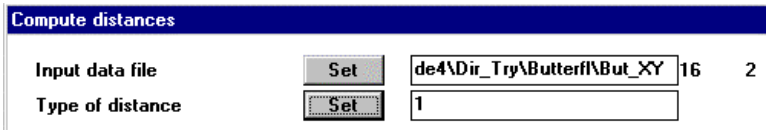
Dialogue L'option utilise une seule fenêtre de dialogue :



1) Nom du fichier binaire d'entrée.

2) Option de calcul des distances. Dans l'option type de distances, taper 1 pour calculer des distances euclidiennes avec la métrique canonique, taper 2 pour calculer des distances du khi2 et taper 3 pour obtenir la distance dérivée de l'indice de Jaccard.

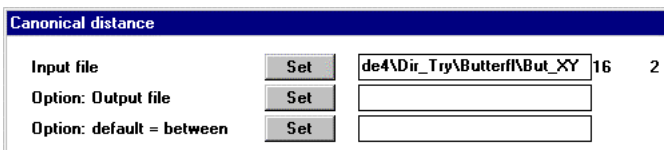
Exemple Utiliser la carte Butterfly[1,2] de la pile ADE-4•Data pour obtenir le fichier But_XY (16-2) :




```
Clusters: Compute distances
Input file: D:\Ade4\Dir_Try\Butterfl\But_XY
Number of rows: 16, columns: 2
The distance used is the Euclidean distance
Output file: D:\Ade4\Dir_Try\Butterfl\But_XY.dist
Number of rows: 16, columns: 16

--------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\BUTTERFL\But_XY.dist - 16 rows, 16 cols.
   1 |  0.0000  0.4150  0.4262  0.4423  0.4546  0.4535  0.4784  0.9790  1.0000  0.4603
0.6661  0.2388  0.4814  0.5040  0.6127  0.4765
   2 |  0.4150  0.0000  0.0125  0.0276  0.0396  0.0505  0.0813  0.6109  0.6306  0.0836
0.2669  0.2135  0.2013  0.1947  0.2994  0.2226
…
  14 |  0.5040  0.1947  0.1976  0.1933  0.1900  0.1540  0.1331  0.4751  0.4961  0.1145
0.2019  0.2652  0.0325  0.0000  0.1152  0.0582
  15 |  0.6127  0.2994  0.2996  0.2918  0.2854  0.2526  0.2251  0.3714  0.3926  0.2158
0.1842  0.3747  0.1318  0.1152  0.0000  0.1376
  16 |  0.4765  0.2226  0.2283  0.2277  0.2271  0.1908  0.1764  0.5086  0.5299  0.1524
0.2584  0.2401  0.0279  0.0582  0.1376  0.0000
```

Important On obtient exactement le même résultat en deux étapes avec :

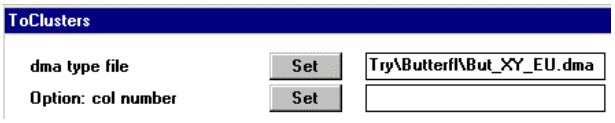
1) DMAUtil: Canonical distance :




```
Distance matrix computation
-------------------------------------------
Input file: D:\Ade4\Dir_Try\Butterfl\But_XY
It has 16 rows and 2 columns
```

```
Distances are computed among rows
-----------------------------------------
Canonical distances computed
Output file: D:\Ade4\Dir_Try\Butterfl\But_XY_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: D:\Ade4\Dir_Try\Butterfl\But_XY_EU.dma
  1 -> 16
  2 -> 1
  3 -> Classical metric on D:\Ade4\Dir_Try\Butterfl\But_XY
  4 -> TRUE
-----------------------------------------
```

1) DMAUtil: ToClusters :



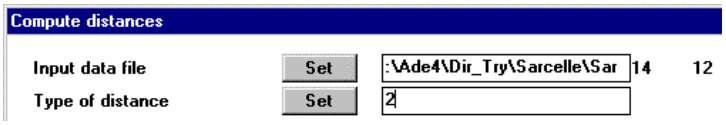

```
Output file : D:\Ade4\Dir_Try\Butterfl\But_XY_EU1.dist Row: 16 Col: 16
Tranformation: rescaling on [0,1] by y=(x-min)/(max-min)

-------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\BUTTERFL\But_XY_EU1.dist - 16 rows, 16 cols.
   1 |  0.0000  0.4150  0.4262  0.4423  0.4546  0.4535  0.4784  0.9790  1.0000  0.4603
0.6661  0.2388  0.4814  0.5040  0.6127  0.4765
   2 |  0.4150  0.0000  0.0125  0.0276  0.0396  0.0505  0.0813  0.6109  0.6306  0.0836
0.2669  0.2135  0.2013  0.1947  0.2994  0.2226
…
  15 |  0.6127  0.2994  0.2996  0.2918  0.2854  0.2526  0.2251  0.3714  0.3926  0.2158
0.1842  0.3747  0.1318  0.1152  0.0000  0.1376
  16 |  0.4765  0.2226  0.2283  0.2277  0.2271  0.1908  0.1764  0.5086  0.5299  0.1524
0.2584  0.2401  0.0279  0.0582  0.1376  0.0000
```

Comme indiqué dans le listing, les distances sont ramenées dans l'intervalle [0,1] par le changement de variable :

$$d_{ij} \mapsto \frac{d_{ij}}{\max_{i,j}\left(d_{ij}\right)} \quad (1)$$

Difficulté Le module clusters demande des fichiers d'extension **.dist**. Ces fichiers contiennent des matrices de distances modifiées par la changement de variables (1). Pour définir un dendogramme ou une partition sur la base d'une matrice de distance cette opération est neutre et garantit des bonnes conditions numériques. Pour obtenir la matrice de distance exacte, on prendra une des options de DMAUtil puis DMAUtil: ToClusters pour l'utiliser dans Clusters. DMAUtil permet en outre de calculer un très grand nombre de matrices de distances (voir la fiche du module) dont on a seulement 3 exemples dans la présente option.

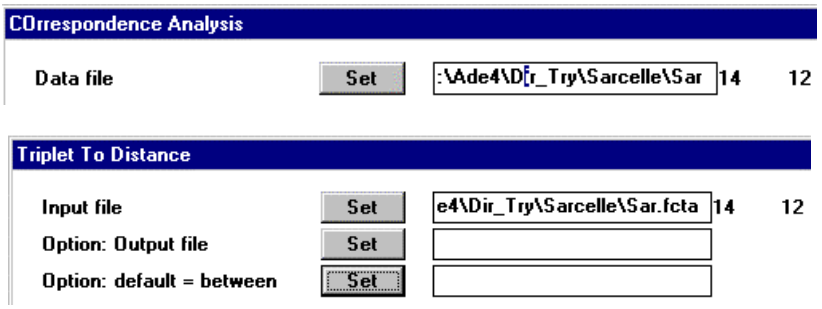
Lien Pour avoir le distance du khi2, utiliser la carte Sarcelles [3] de la pile ADE-4•Data pour obtenir le fichier Sar (14-12). Calculer la distance du Khi2 entre lignes :



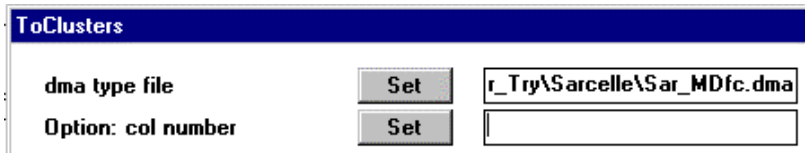

```
Clusters: Compute distances
Input file: D:\Ade4\Dir_Try\Sarcelle\Sar
Number of rows: 14, columns: 12
The distance used is the Chi2 distance
Output file: D:\Ade4\Dir_Try\Sarcelle\Sar.dist
Number of rows: 14, columns: 14

-------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\SARCELLE\Sar.dist - 14 rows, 14 cols.
```

```
   1 |  0.0000  0.2344  0.1020  0.2225  0.4256  0.3184  0.6141  0.4052  0.4489  0.5718
0.6216  0.6502  0.8267  0.6774
   2 |  0.2344  0.0000  0.2328  0.3721  0.2967  0.2338  0.6366  0.3124  0.4215  0.4532
0.5497  0.5804  0.8033  0.6264
   3 |  0.1020  0.2328  0.0000  0.2255  0.4420  0.2817  0.5895  0.3659  0.3905  0.5158
0.5790  0.6094  0.7974  0.6379
…
```

On a exactement le même résultat en trois étapes avec :



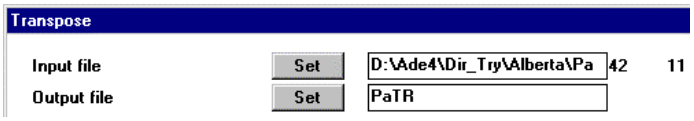

```
Computed distances use the diagonal metric and the centered table of the triplet
Output file: D:\Ade4\Dir_Try\Sarcelle\Sar_MDfc
It has 91 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: D:\Ade4\Dir_Try\Sarcelle\Sar_MDfc.dma
  1 -> 14
  2 -> 1
  3 -> Euclidean distance from triplet D:\Ade4\Dir_Try\Sarcelle\Sar.fcta
  4 -> TRUE
```



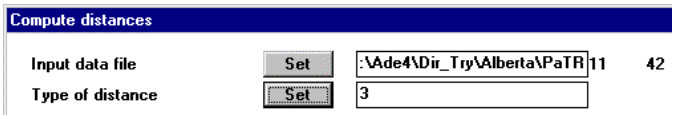

```
Output file : D:\Ade4\Dir_Try\Sarcelle\Sar_MDfc1.dist Row: 14 Col: 14
Tranformation: rescaling on [0,1] by y=(x-min)/(max-min)

--------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\SARCELLE\Sar_MDfc1.dist - 14 rows, 14 cols.
   1 |  0.0000  0.2344  0.1020  0.2225  0.4256  0.3184  0.6141  0.4052  0.4489  0.5718
0.6216  0.6502  0.8267  0.6774
   2 |  0.2344  0.0000  0.2328  0.3721  0.2967  0.2338  0.6366  0.3124  0.4215  0.4532
0.5497  0.5804  0.8033  0.6264
   3 |  0.1020  0.2328  0.0000  0.2255  0.4420  0.2817  0.5895  0.3659  0.3905  0.5158
0.5790  0.6094  0.7974  0.6379
…
```

On retrouve ici la première utilisation de l'analyse des correspondances dans la logique de la classification [4]. Cette procédure donne une matrice de distances entre les lignes (ou entre les colonnes) d'un tableau *pour tout type d'analyse du premier niveau*, en particulier pour des variables qualitatives (MCA: Multiple Correspondence Analysis), floues (MCA: Fuzzy Correspondence Analysis) ou des mélanges de type (MCA: Hill and Smith Analysis).

Difficulté Pour illustrer la distance de Jaccard, utiliser la carte Alberta [5]. Transposer le fichier PA :

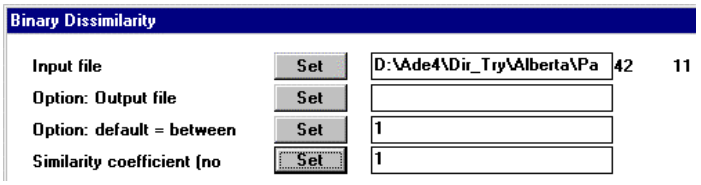


Calculer les distances entre espèces :

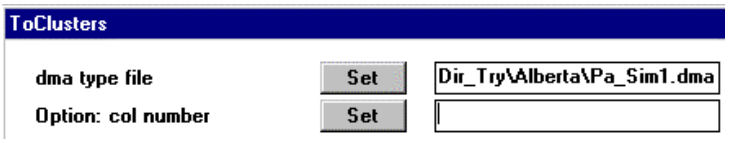



```
Clusters: Compute distances
Input file: D:\Ade4\Dir_Try\Alberta\PaTR
Number of rows: 11, columns: 42
The distance used is the Jaccard index
Output file: D:\Ade4\Dir_Try\Alberta\PaTR.dist
Number of rows: 11, columns: 11

-------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\ALBERTA\PaTR.dist - 11 rows, 11 cols.
   1 |  0.0000  0.2105  0.9000  0.8000  0.9615  0.8421  0.8333  0.9474  1.0000  1.0000  1.0000
   2 |  0.2105  0.0000  0.9310  0.8333  0.9583  0.8889  0.8824  0.9412  1.0000  1.0000  1.0000
   3 |  0.9000  0.9310  0.0000  0.5000  0.5000  1.0000  1.0000  1.0000  0.9333  1.0000  1.0000
…
```

On n'obtient pas exactement le même résultat en deux étapes avec :



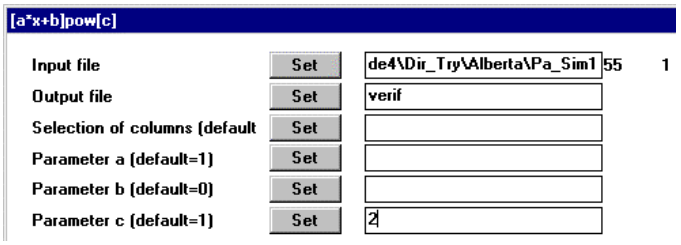

```
Input file: D:\Ade4\Dir_Try\Alberta\Pa
It has 42 rows and 11 columns
Distances are computed among columns

JACCARD index (1901)
S3 coefficient of GOWER & LEGENDRE
Euclidean distance
Distances are computed by
s = a/(a+b+c) --> d = sqrt(1 - s)
Output file: D:\Ade4\Dir_Try\Alberta\Pa_Sim1
It has 55 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: D:\Ade4\Dir_Try\Alberta\Pa_Sim1.dma
  1 -> 11
  2 -> 1
  3 -> JACCARD index on D:\Ade4\Dir_Try\Alberta\Pa
  4 -> TRUE
```



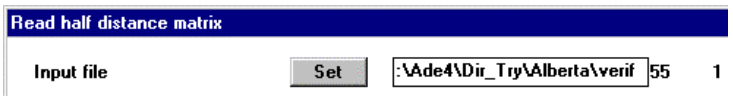

```
Output file : D:\Ade4\Dir_Try\Alberta\Pa_Sim11.dist Row: 11 Col: 11
Tranformation: rescaling on [0,1] by y=(x-min)/(max-min)

-------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\ALBERTA\Pa_Sim11.dist - 11 rows, 11 cols.
   1 |  0.0000  0.4588  0.9487  0.8944  0.9806  0.9177  0.9129  0.9733  1.0000  1.0000  1.0000
   2 |  0.4588  0.0000  0.9649  0.9129  0.9789  0.9428  0.9393  0.9701  1.0000  1.0000  1.0000
   3 |  0.9487  0.9649  0.0000  0.7071  0.7071  1.0000  1.0000  1.0000  0.9661  1.0000  1.0000
```

On tombe sur une des principales difficultés des méthodes basées sur les calcul de distances. Si d_{ij} est la distance entre les objets i et j, $\sqrt{d_{ij}}$ en est une autre, d_{ij}^2 en est une troisième, … Chacune d'entre elles fournit un dendrogramme pour chacun des algorithmes disponibles et le nombre de cas possibles est énorme. Dans cette illustration Clusters: Compute distances (option Jaccard) choisit de travailler sur $d_{ij} = 1 - s_{ij}$ où s désigne l'indice de similarité. DMAUtil: Binary Dissimilarity (option 1) choisit de travailler sur $d_{ij} = \sqrt{1 - s_{ij}}$. Du point de vue de la classification, il n'y a pas un choix plus justifié que

l'autre. Du point de vue de l'ordination, ce n'est pas vrai car la matrice des $d_{ij} = \sqrt{1 - s_{ij}}$ est euclidienne (voir DMAUtil et DMAUse) alors que $d_{ij} = 1 - s_{ij}$ ne l'est pas. L'une justifie une analyse en coordonnées principales (DMAUse: Principal Coordinates) et l'autre non. Pour passer d'un point de vue à l'autre, utiliser Bin-Bin: [a*x+b]pow[c] :

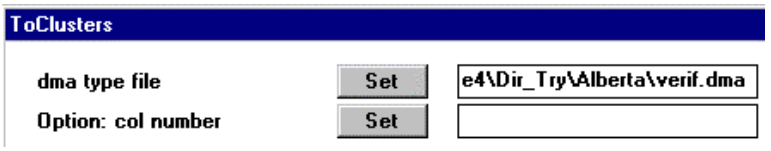


Pa_Sim1 contient la demi-matrice de distances écrite sur une seule colonne. On peut lui faire subir une transformation arbitraire (ici élever au carré), puis relire le résultat comme une demi-matrice de distances (DMAUtil: Read half distance matrix) :



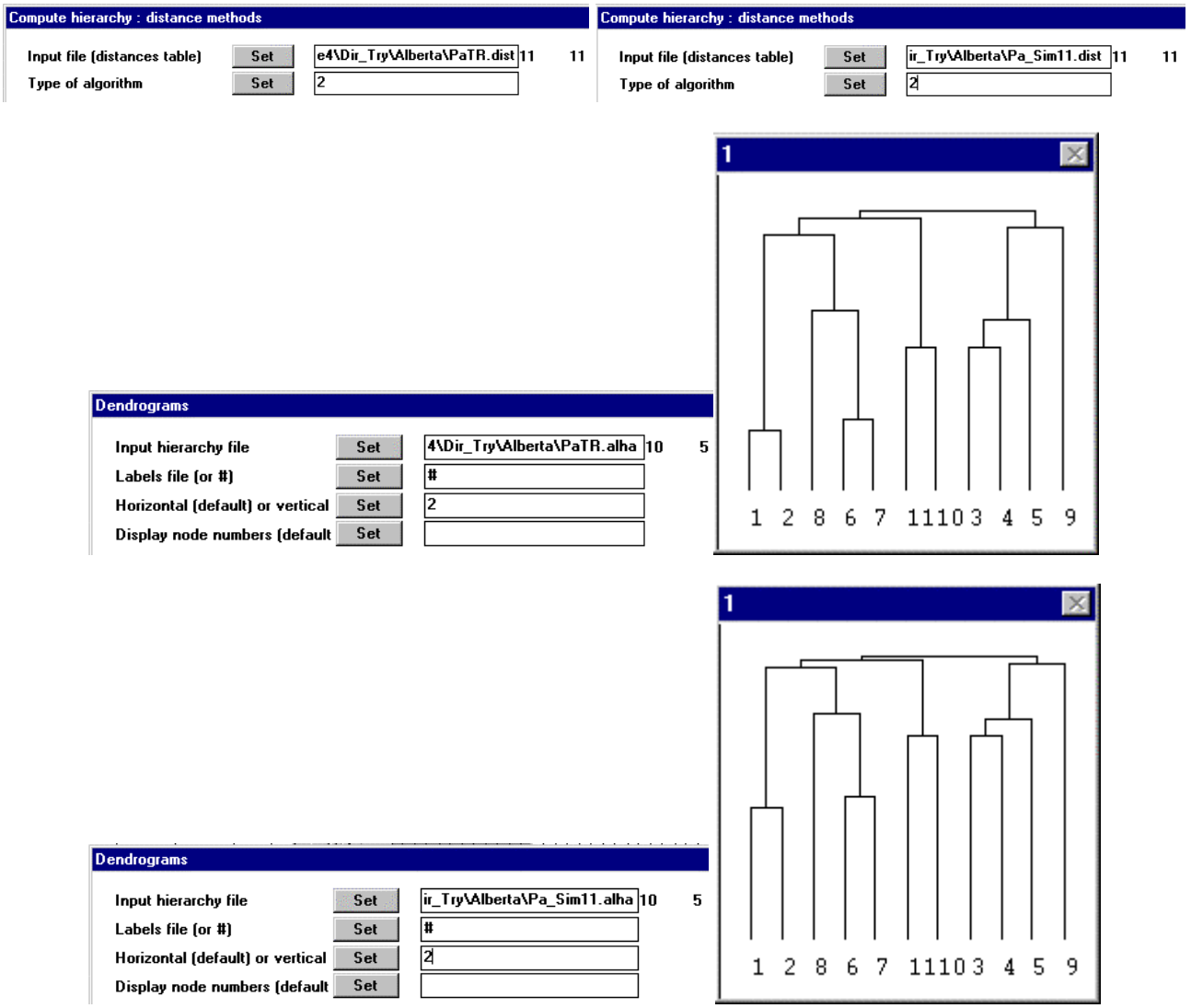

```
Input file: D:\Ade4\Dir_Try\Alberta\verif
D:\Ade4\Dir_Try\Alberta\verif is a binary file with 55 rows and 1 columns
55 is 11*(11-1)/2
Test of the euclidean property by diagonalization (theorem of GOWER)
 matrix number  1 ---> FALSE
Text file: D:\Ade4\Dir_Try\Alberta\verif.dma
  1 -> 11
  2 -> 1
  3 -> Input half distance matrix file D:\Ade4\Dir_Try\Alberta\verif
  4 -> FALSE
```

Observer que la nouvelle matrice n'est pas euclidienne conformément aux résultats classiques [6] et envoyer cette matrice à Clusters :

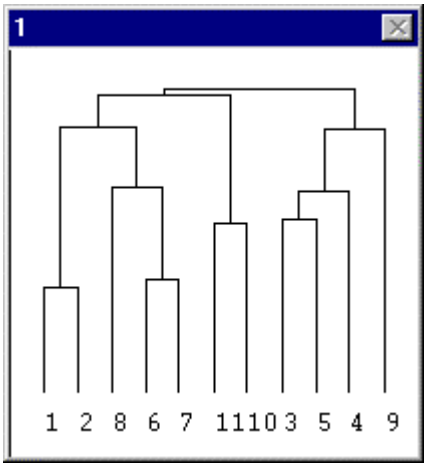



```
Output file : D:\Ade4\Dir_Try\Alberta\verif1.dist Row: 11 Col: 11
Tranformation: rescaling on [0,1] by y=(x-min)/(max-min)

------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\ALBERTA\verif1.dist - 11 rows, 11 cols.
   1 |  0.0000  0.2105  0.9000  0.8000  0.9615  0.8421  0.8333  0.9474  1.0000  1.0000
1.0000
   2 |  0.2105  0.0000  0.9310  0.8333  0.9583  0.8889  0.8824  0.9412  1.0000  1.0000
1.0000
   3 |  0.9000  0.9310  0.0000  0.5000  0.5000  1.0000  1.0000  1.0000  0.9333  1.0000
1.0000
…
```

On obtient exactement le résultat de la présente option. Cet exemple montre que ADE-4 permet de faire exactement ce qu'on veut, ce qui n'est pas forcément rassurant. En effet pour des données de présence-absence il y a 10 indices disponibles dans DMAUtil: Binary Dissimilarity, 2 manières (au moins) de les utiliser (simple ou au carré), 3 type de liens dans Clusters: Compute hierarchy : distance methods, soit déjà 60 dendrogrammes différents.



Il faut donc bien préciser ce qu'on fait. Ci-dessous, dendrogramme sur 11 espèces obtenu par le lien UGPMA sur la distance euclidienne ($d_{ij} = \sqrt{1 - s_{ij}}$) associée à l'indice d'Ochiai $s = a \Big/ \sqrt{(a+b)(a+c)}$ (a nombre de présences communes, b et c nombre de présences séparées) choisi parce que dans ce tableau il y a 10 lacs sans aucune espèce de poissons.



Références [1] 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.

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

[3] Lebreton, J.D. (1973) Etude des déplacements saisonniers des Sarcelles d'hiver, Anas c. crecca L., hivernant en Camargue à l'aide de l'analyse factorielle des correspondances. *Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris,* D : III, 277, 2417-2420.

[4] Roux, G. & Roux, M. (1967) A propos de quelques méthodes de classification en phytosociologie. *Revue de Statistique Appliquée* : XV, 2, 59-72.

[5] Robinson, C.L.K. & Tonn, W.M. (1989) Influence of environmental factors and piscivory in structuring fish assemblages of small Alberta lakes. *Canadian Journal of Fisheries and Aquatic Sciences* : 46, 81-89.

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

Clusters : Compute hierarchy : distance methods

Type Mise en œuvre des algorithmes de classification hiérarchique (**CAH**).

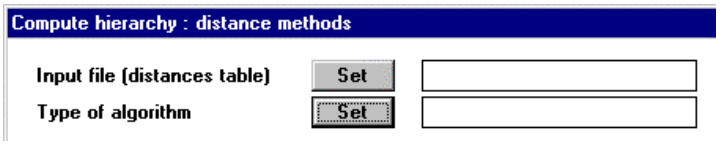
Objet L'option calcule une hiérarchie à partir d'une matrice de distances. Elle propose trois procédures. Une hiérarchie sur un ensemble fini A est un ensemble $\mathcal{A}$ de parties de A (les classes), c'est à dire un sous-ensemble de l'ensemble des parties de A telle que :

1 - A est un élément de $\mathcal{A}$ (la classe la plus grande contient tous les éléments) ;
2 - Si a est un élément de $\mathcal{A}$, {a} est un élément de $\mathcal{A}$ (les classes les plus petites contiennent un seul élément) ;
3 - Si H et K sont deux classes de $\mathcal{A}$, de trois choses l'une : ou bien elles sont sans éléments communs, ou bien H contient K, ou bien K contient H (deux classes sont soit disjointes soit emboîtées) [1].

La classification linnéenne est hiérarchique : un genre et une famille ou bien n'ont pas d'espèces en commun ou bien le genre appartient à la famille.

Un algorithme de classification hiérarchique est une méthode de construction d'une hiérarchie. Les classifications hiérarchiques ascendantes (**CAH**) partent de l'ensemble des classes à un seul élément et à chaque pas réunissent deux classes les plus ressemblantes jusqu'à obtenir la classe contenant tous les éléments. La notion de ressemblance ou différence de deux classes utilise une notion de proximité entre classes qui dérive de la distance entre éléments de départ [2]. Les choix d'une distance initiale, d'un principe de construction, d'un indice de proximité entre classes engendrent une multitude de possibilités qui exclue qu'une hiérarchie soit vraie ou fausse : elle peut être seulement utile. Voir les principes généraux dans [3] et une revue des logiciels dans [4]. La présentation faite en [5] est très efficace.

Dialogue L'option utilise une seule fenêtre de dialogue :



1) Matrices de distances dans un fichier **.dist** (depuis Clusters: Compute distances ou DMAUtil: ToClusters).

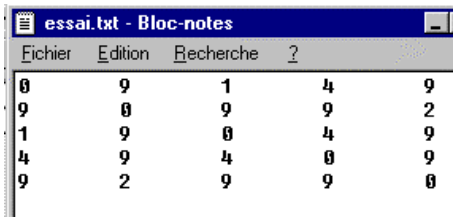
2) Type de lien. A chaque étape, les deux classes de distance minimum sont réunies. Le lien définit la distances entre deux parties à partir de la distance entre deux individus.

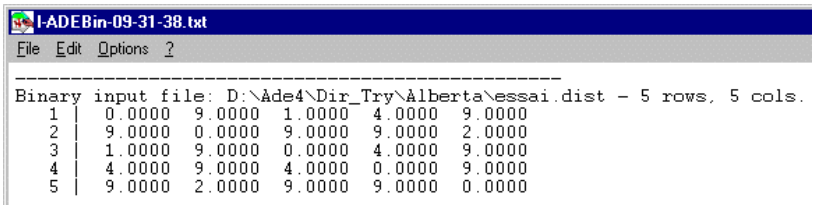
1 - **CAH** : Méthode du lien simple. A chaque pas, la distance entre deux classes est définie par la plus petite distance entre deux points de chacune des deux classes (saut minimal ou *single linkage* [5] p. 156 ou lien minimum [1] p. 57).

2 - **CAH** : Méthode du lien moyen ou UPGMA. A chaque pas, la distance entre deux classes est définie par la moyenne de la distance entre deux points de chacune des deux classes (distance moyenne [5] p. 156).

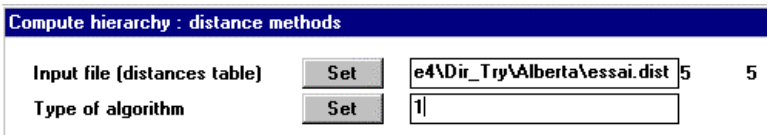
3 - **CAH** : Méthode du lien complet. A chaque pas, la distance entre deux classes est définie par la distance maximum entre deux points de chacune des deux classes (saut maximal ou diamètre [5] p. 156, distance du lien maximum [1] p. 57).

Exemple Saisir la matrice de distances proposée en [5] (tableau 2.2 1 p. 160) :





et sauvegarder en texte. Passer en binaire dans Essai.dist (5-5).




```
Clusters: Compute hierarchy
Distance file: D:\Ade4\Dir_Try\Alberta\essai.dist
Number of rows: 5, columns: 5
Output file: D:\Ade4\Dir_Try\Alberta\essai.slha
Number of rows: 4, columns: 5
Hierarchy algorithm used : single link
```

La hiérarchie est définie par :

```
-------------------------------------------------
Binary input file: D:\Ade4\Dir_Try\Alberta\essai.slha - 4 rows, 5 cols.
   1 |  6.0000  1.0000  3.0000  2.0000  1.0000
   2 |  7.0000  2.0000  5.0000  2.0000  2.0000
   3 |  8.0000  6.0000  4.0000  3.0000  4.0000
   4 |  9.0000  8.0000  7.0000  5.0000  9.0000
```

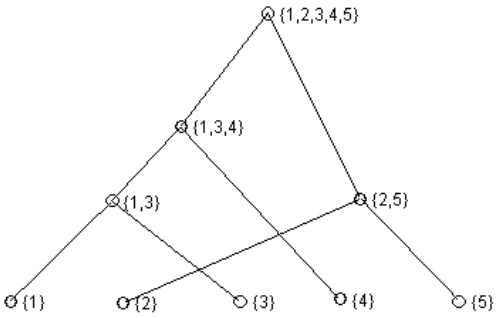
Le fichier de sortie contient *n*-1 lignes si la matrice de distances porte sur *n* points. Il comporte 5 colonnes, respectivement le numéro de la classe (à partir de *n* + 1), le numéro de l'aîné, le numéro du cadet, le nombre d'éléments et l'indice de niveau. La signification de ces objets est la suivante.

Les 5 premières parties de la hiérarchie de partitions sont les 5 classes à un élément, soit {1}, {2}, ... {5}. Les suivantes sont numérotées 6, ..., 9 dans la première colonne du fichier de sortie. La classe 6 a pour aîné la classe 1 (colonne 2) et pour cadet la classe 3 (colonne 3). La nomination aîné-cadet exprime simplement que la classe 6 a été obtenue par regroupement des classes 1 et 3. Elle a donc 2 éléments (colonne 4) et est la partie {1,3}.

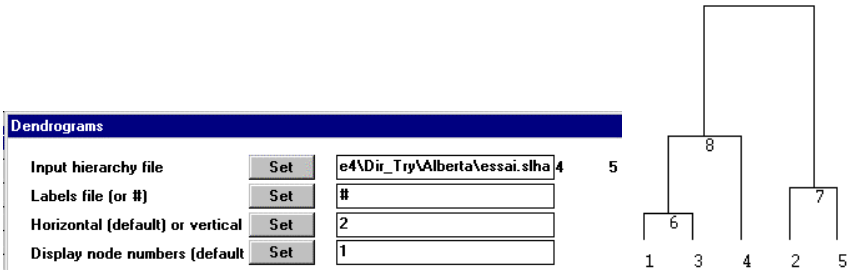
La classe 7 a pour aîné la classe 2 et pour cadet la classe 5. Elle a donc 2 éléments et vaut {2,5}.

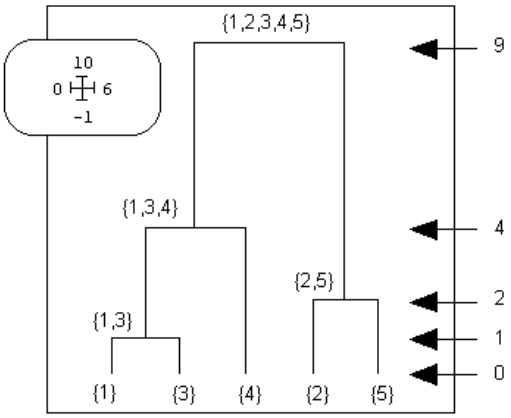
La classe 8 a pour aîné 6 et pour cadet 4, elle a donc 3 éléments et vaut {1, 3, 4}

La classe 9 a pour aîné 8 et pour cadet 7 : elle a 5 élément et vaut {1, 2, 3, 4, 5}. C'est l'ensemble tout entier. On a donc la hiérarchie de parties :



Le dessin automatique de la hiérarchie est le dendrogramme :





On a une hiérarchie indicée si à toute partie H de la hiérarchie on peut associer une valeur numérique v(H) positive ou nulle. La hiérarchie est indicée de façon naturelle par la valeur de la distance correspondant à chaque étape (δ p. 160). L'indice vaut 0 pour une classe élémentaire à un seul élément. Au niveau 1, on a agrégé les classes {1} et {3} distantes de 1 = d(1,3). Les distances au pas suivant (lien simple) sont :

	{1,3}	{2}	{4}	{5}
{1,3}	0	9	4	9
{2}	9	0	9	2
{4}	4	9	0	9
{5}	9	2	9	0

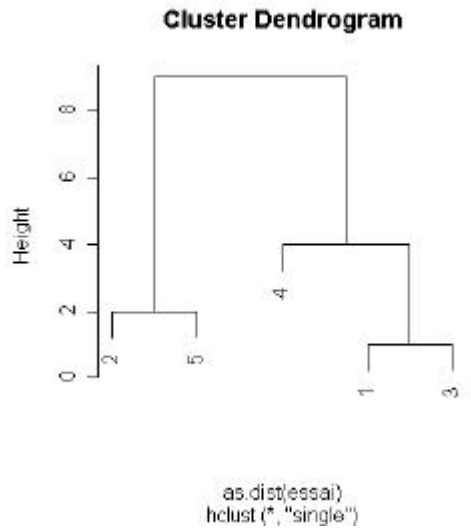
L'indice prend alors la valeur 2 et conduit à la matrice de distances :

	{1,3}	{2,5}	{4}
{1,3}	0	9	4
{2,5}	9	0	9
{4}	4	9	0

L'indice prend la valeur 4 et la distance entre {1,3,4} et {2,5} vaut 9 dernier niveau de l'indice.

Lien Dans R (Voir http://pbil.univ-lyon1.fr/R) :

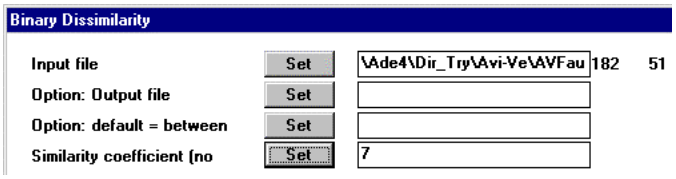
```
[1] "Dossier de travail = D:\\Ade4\\Dir_Try\\Alberta"
> read.table("essai.txt")
  V1 V2 V3 V4 V5
1  0  9  1  4  9
2  9  0  9  9  2
3  1  9  0  4  9
4  4  9  4  0  9
5  9  2  9  9  0
> essai_read.table("essai.txt")
> plot.hclust(hclust(as.dist(essai),met="single"))
```



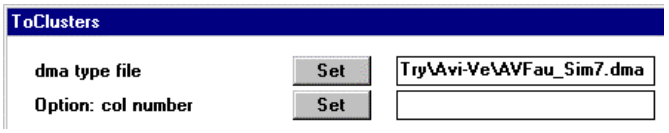

```
> ?hclust
```

A number of different clustering methods are provided. Ward's minimum variance method aims at finding compact, spherical clusters. The complete linkage method finds similar clusters. The single linkage method (which is closely related to the minimal spanning tree) adopts a `friends of friends' clustering strategy. The other methods can be regarded as aiming for clusters with characteristics somewhere between the single and complete link methods.

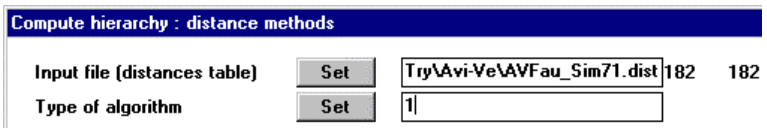
Exemple Utiliser la carte Avi-Ve [6]. Calculer la distance basée sur l'indice d'Ochiai entre les 182 relevés :



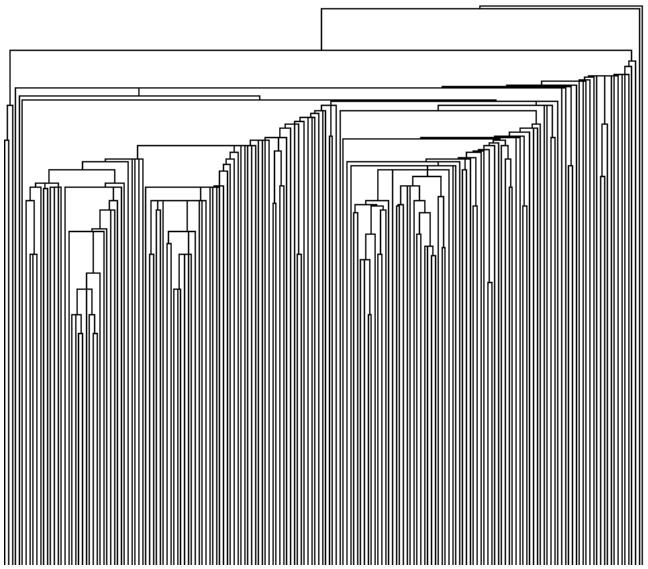

```
S12 coefficient of GOWER & LEGENDRE
OCHIAI (1957)
Euclidean distance
Distances are computed by
s = a/sqrt((a+b)(a+c)) --> d = sqrt(1 - s)
Output file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim7
It has 16471 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim7.dma
  1 -> 182
  2 -> 1
  3 -> S12 index of GOWER & LEGENDRE on D:\Ade4\Dir_Try\Avi-Ve\AVFau
  4 -> TRUE
```



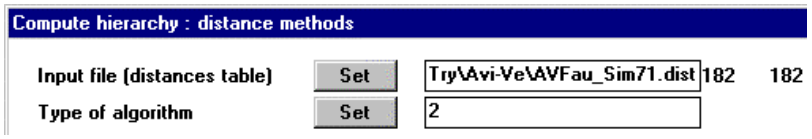
```
Output file : D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.dist Row: 182 Col: 182
Tranformation: rescaling on [0,1] by y=(x-min)/(max-min)
```



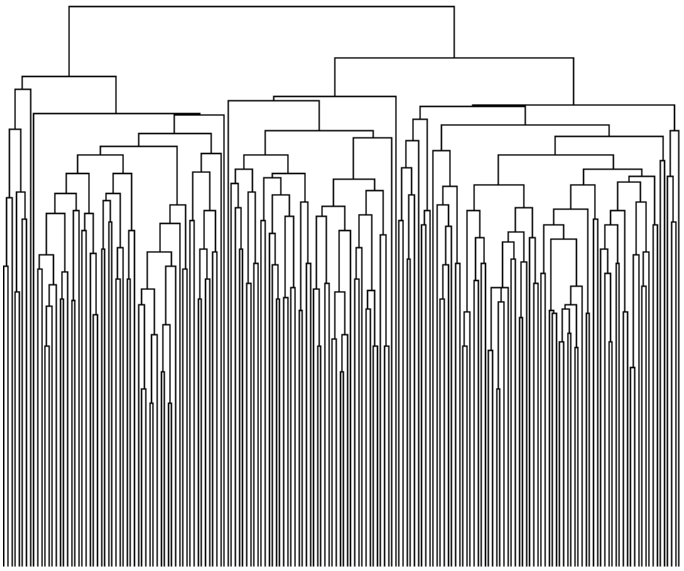

```
Clusters: Compute hierarchy
Distance file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.dist
Number of rows: 182, columns: 182
Output file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.slha sl pour single link
Number of rows: 181, columns: 5
Hierarchy algorithm used : single link
```



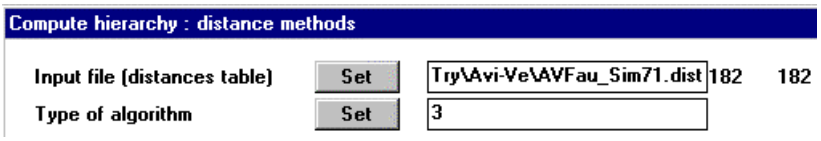
`*friends of friends' clustering strategy : effet de chaîne de voisin en voisin* (5 p. 167)



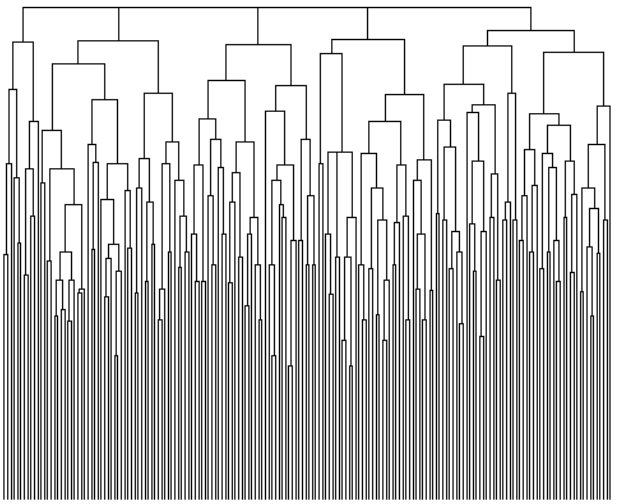

```
Clusters: Compute hierarchy
Distance file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.dist
Number of rows: 182, columns: 182
Output file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.alha al pour average link
Number of rows: 181, columns: 5
Hierarchy algorithm used : average link (UPGMA)
```



characteristics between the single and complete link methods : compromis du lien moyen

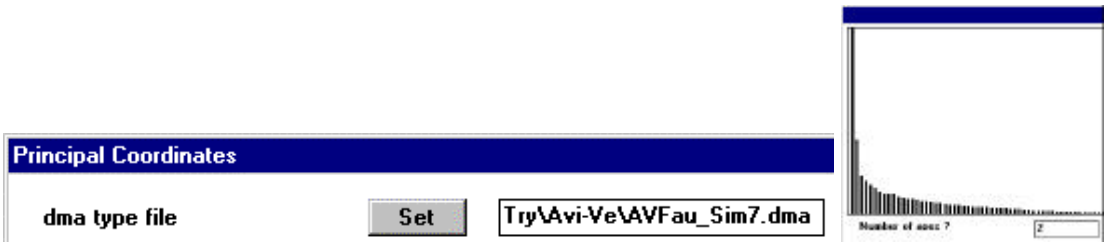



```
Clusters: Compute hierarchy
Distance file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.dist
Number of rows: 182, columns: 182
Output file: D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.clha cl pour complete link
Number of rows: 181, columns: 5
Hierarchy algorithm used : complete link
```

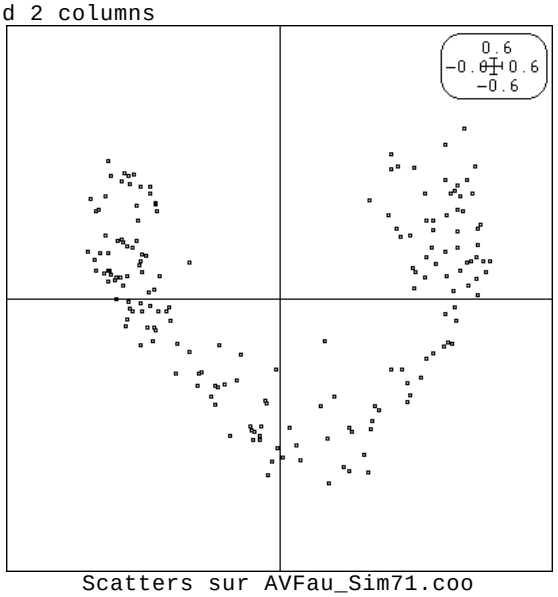


complete linkage method finds similar clusters : groupement régulier du lien complet

L'ordination correspondante (DMAUtil) :

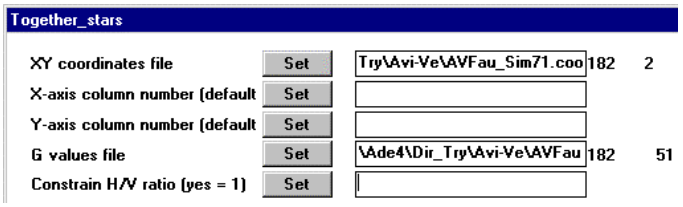


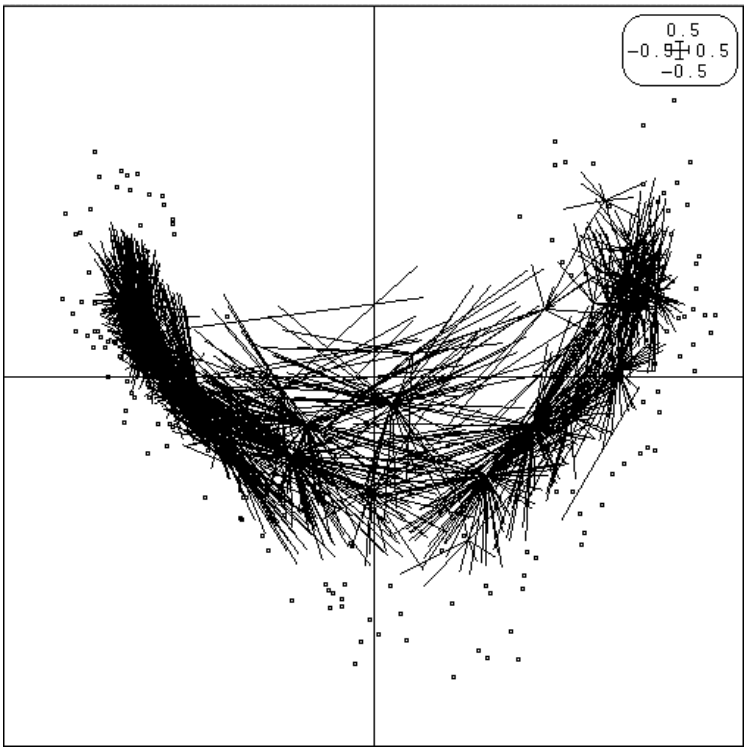

```
File  D:\Ade4\Dir_Try\Avi-Ve\AVFau_Sim71.coo  contains  the  principal  coordinates
(norm=sqrt(lambda))
--- It has 182 rows and 2 columns
```



Scatters sur AVFau_Sim71.coo

Pour voir les espèces qui "replient" le gradient (ScatterDistri: Together_stars) :





Les structures écologiques sont des objets complexes qui méritent ordination ET classification.

Références

1 Diday, E., Lemaire, J., Pouget, J. & Testu, F. (1982) *Elements d'analyse de données*. Dunod, Paris. 1-462. (Ch. 2, Classification automatique, p. 74).

2 Rouanet, H. & Le Roux, B. (1993) *Analyse des données multidimensionnelles*. Dunod, paris. 1-310. (Ch. V-3 Classification ascendante hiérarchique p. 120).

3 Roux, M. (1985) *Algorithmes de classification*. Masson, Paris. -.
Roux, M. (1991) Basic procedures in hierarchical cluster analysis. In : *Applied Multivariate Analysis in SAR and Environmental Studies*. Devillers, J. & Karcher, W. (Ed.) Kluwer Academic Publishers, Dordrecht, The Netherlands. 115-136.

4 Blashfield, R.K., Aldenderfer, M.S. & Morey, L.C. (1982) Cluster Analysis Software. In : *Handbook of Statistics*, Vol. 2. Krishnaiah, P.R. & Kanal, L.N. (Eds.) North Holland Publishing Company, Amsterdam. 245-266.

5 Lebart, L., Morineau, A. & Piron, M. (1995) *Statistique exploratoire multidimensionnelle*. Dunod, Paris. 1-439. Section 2.2 Classification hiérarchique.

6 Prodon, R. & Lebreton, J.D. (1981) Breeding avifauna of a Mediterranean succession : the holm oak and cork oak series in the eastern Pyrénées. 1 : Analysis and modelling of the structure gradient. *Oï kos* : 37, 21-38.

Clusters : Compute hierarchy : Ward method

Type Calcul de hiérarchie de parties basée sur l'inertie. Méthode de CAH partant d'un tableau alors que Clusters: Compute hierarchy : distance methods part d'une matrice de distances.

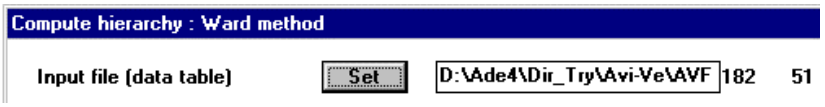
Objet Méthode de Ward ou méthode ascendante sur le moment d'ordre 2. A un pas donné, chaque classe est remplacée par son centre de gravité. L'inertie du nuage des centre de gravité est l'inertie inter-classe et les deux classes réunies sont choisies pour une diminution minimum de cette variance inter-classe (Voir [1] p. 79 , [2] p. 122 et suivantes, [3] critère d'agrégation selon la variance p. 167 et suivantes, et [4]).

L'option part du tableau lui-même et utilise implicitement la métrique canonique (distance euclidienne). Elle peut être mise en œuvre sur tout type de distance en utilisant comme tableau celui des coordonnées factorielles d'une analyse de base ou celui de l'analyse en coordonnées principales d'une matrice de distances euclidienne quelconque.

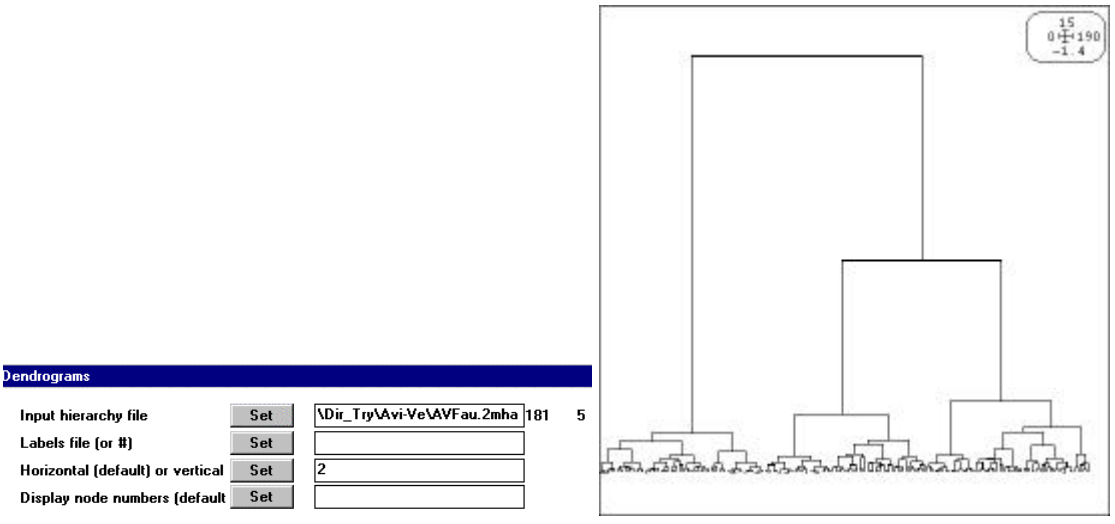
Dialogue L'option utilise une seule fenêtre de dialogue :

1) Tableau d'entrée pour classification hiérarchique des lignes (transposer pour étudier les colonnes).

Exemple Utiliser l'exemple mis en place dans Clusters: Compute hierarchy : distance methods.



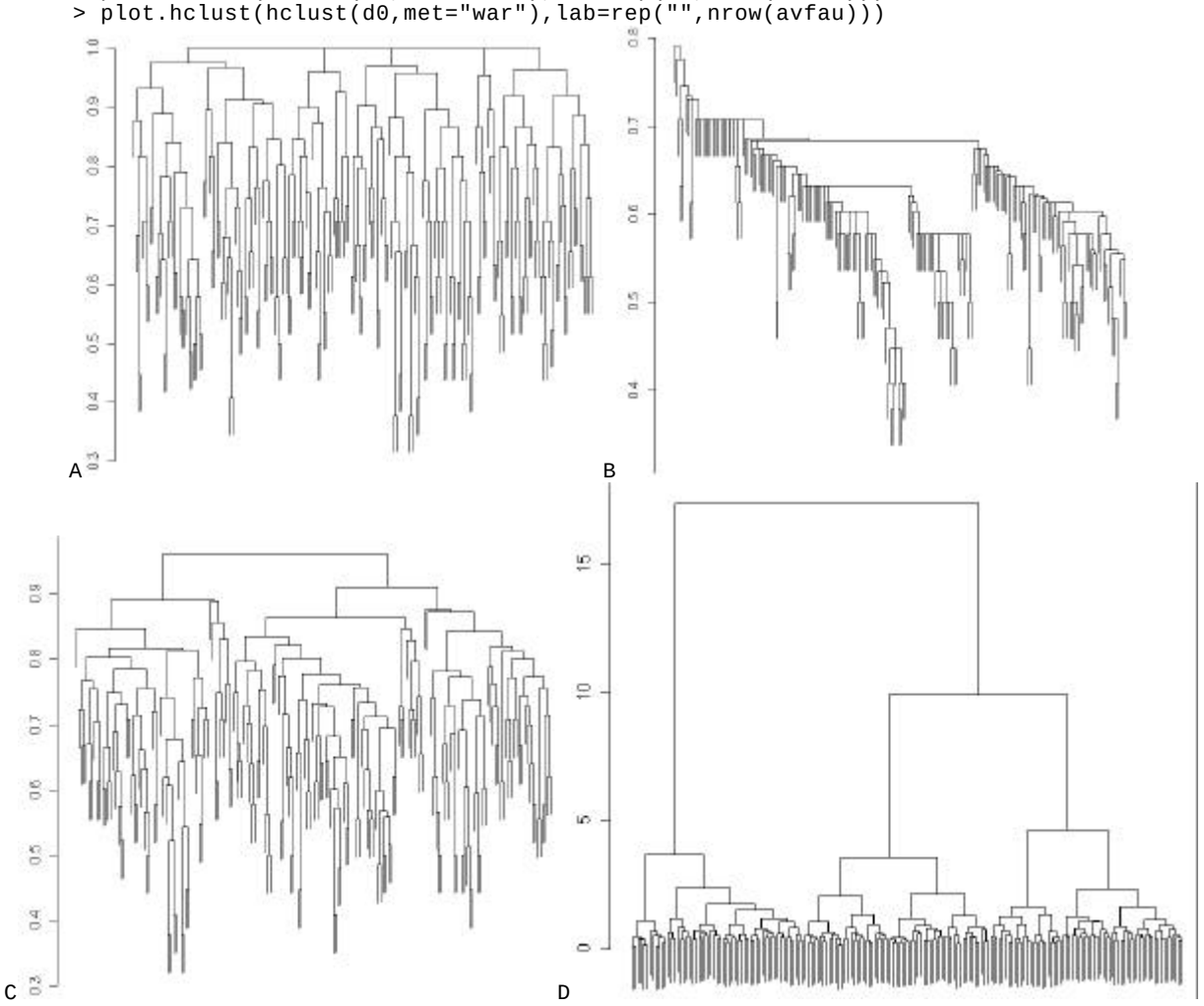

```
Clusters: Compute hierarchy
Data file: D:\Ade4\Dir_Try\Avi-Ve\AVFau
Number of rows: 182, columns: 51
Output file: D:\Ade4\Dir_Try\Avi-Ve\AVFau.2mha 2m pour moment d'ordre 2
Number of rows: 181, columns: 5
Hierarchy algorithm used : second order moment (Ward's method)
```



La différence avec les autres est très importante.

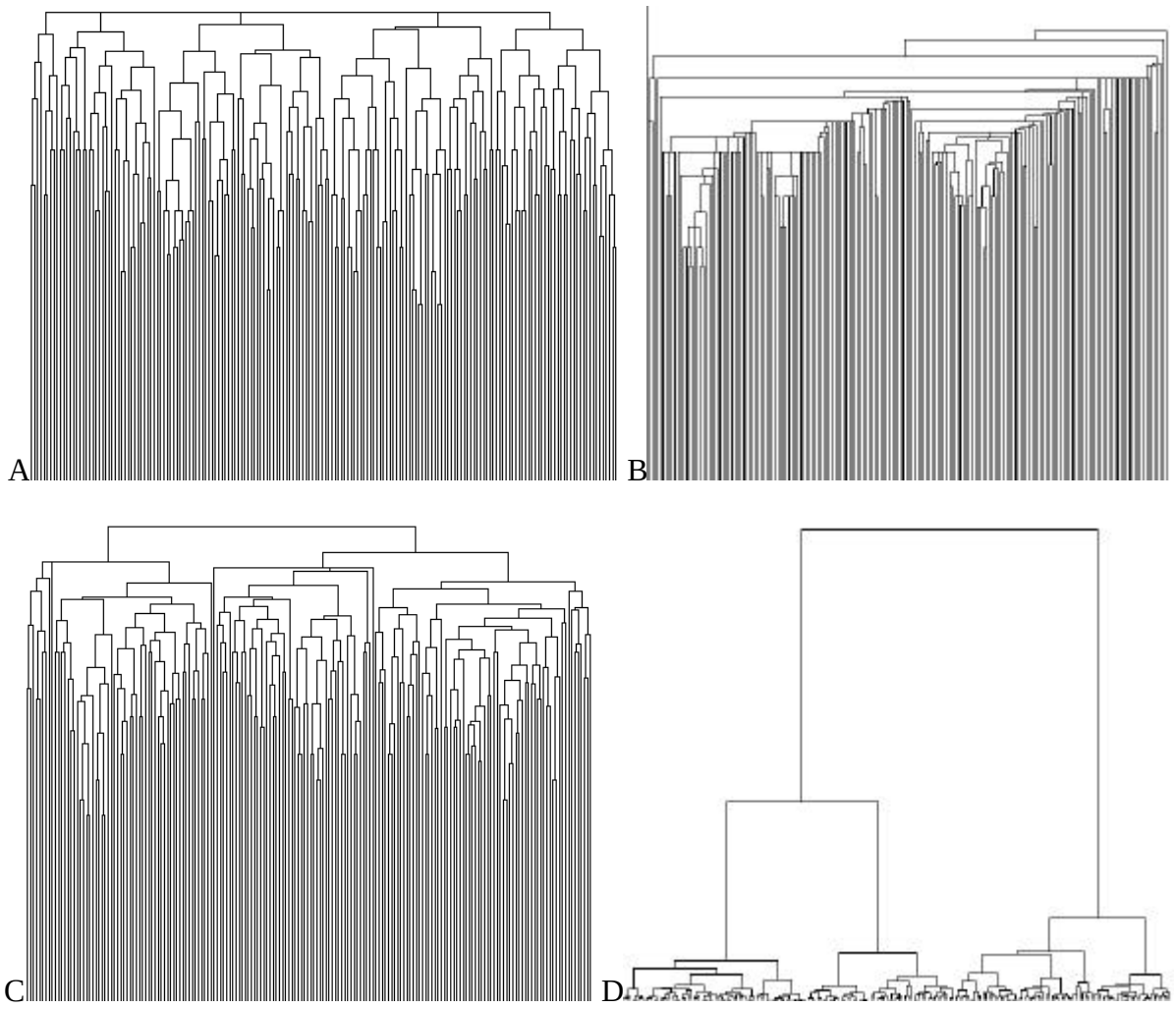
Comparaison On peut vérifier dans R, le rôle considérable du lien utilisé :

```
> d0_as.dist(sqrt(as.matrix(dist(avfau,met="bin"))))
> plot.hclust(hclust(d0,met="comp"),lab=rep("",nrow(avfau)))
> plot.hclust(hclust(d0,met="sing"),lab=rep("",nrow(avfau)))
> plot.hclust(hclust(d0,met="ave"),lab=rep("",nrow(avfau)))
> plot.hclust(hclust(d0,met="war"),lab=rep("",nrow(avfau)))
```



A Lien complet, B Lien simple, C Lien moyen, D Critère de Ward

Comparer avec les modules d'ADE-4. Matrice de distances depuis AVFau avec DMAUtil: Binary Dissimilarity option 1, Clusters: Compute hierarchy : distance methods option 1,2,3, DMAUse: Principal Coordinates et Clusters: Compute hierarchy : Ward method sur le fichier coo :



A Lien complet, B Lien simple, C Lien moyen, D Critère de Ward

Le procédé graphique est différent mais le résultat est globalement le même.

Références

1 Diday, E., Lemaire, J., Pouget, J. & Testu, F. (1982) *Elements d'analyse de données*. Dunod, Paris. 1-462. (Ch. 2, Classification automatique, p. 74).

2 Rouanet, H. & Le Roux, B. (1993) *Analyse des données multidimensionnelles*. Dunod, paris. 1-310. (Ch. V-3 Classification ascendante hiérarchique p. 120).

3 Lebart, L., Morineau, A. & Piron, M. (1995) *Statistique exploratoire multidimensionnelle*. Dunod, Paris. 1-439. Section 2.2 Classification hiérarchique.

4 Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* : 58, 238-244.

Clusters : Compute hierarchy : divisive algorithm

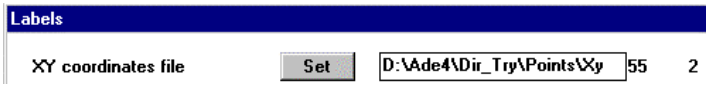
Type Calcul d'une hiérarchie de parties sur matrice de distances. Les classifications hiérarchiques descendantes (**CDH**) partent de la classe contenant tous les éléments et à chaque pas divisent une classe en deux parties les plus différentes jusqu'à obtenir la totalité des classes à un seul élément.

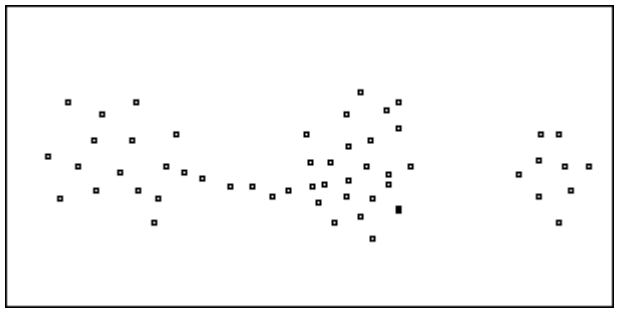
Objet Méthode descendante du moment d'ordre 2.

Dialogue L'option utilise une seule fenêtre de dialogue :

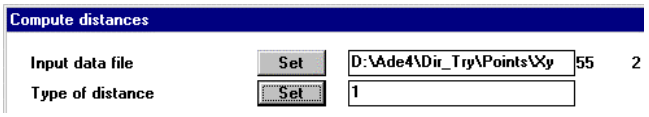
1) Matrice de distances entre objets dans un fichier .dist.

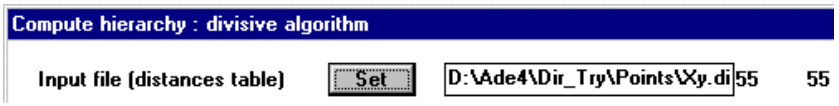
Exemple Utiliser la carte Points (Lebart, L., Morineau, A. & Piron, M. (1995) Statistique exploratoire multidimensionnelle. Dunod, Paris. p.167). Dans Scatters :



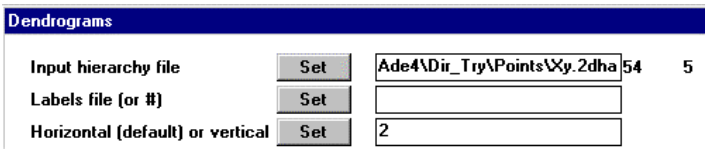


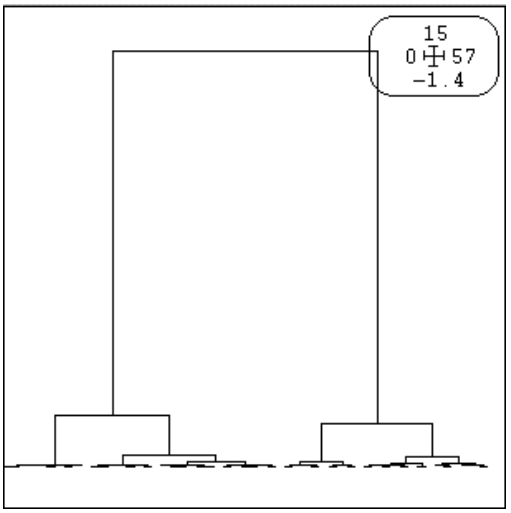
Calculer la matrice de distances entre points :



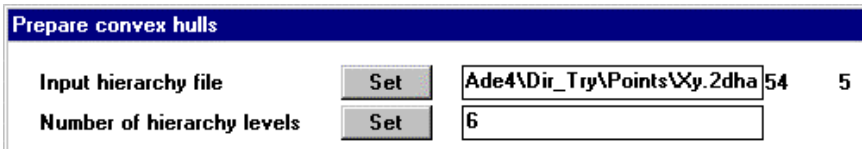



```
Clusters: Compute hierarchy
Distance file: D:\Ade4\Dir_Try\Points\Xy.dist
Number of rows: 55, columns: 55
Output file: D:\Ade4\Dir_Try\Points\Xy.2dha
Number of rows: 54, columns: 5
Hierarchy algorithm used : second order moment (divisive algorithm)
```





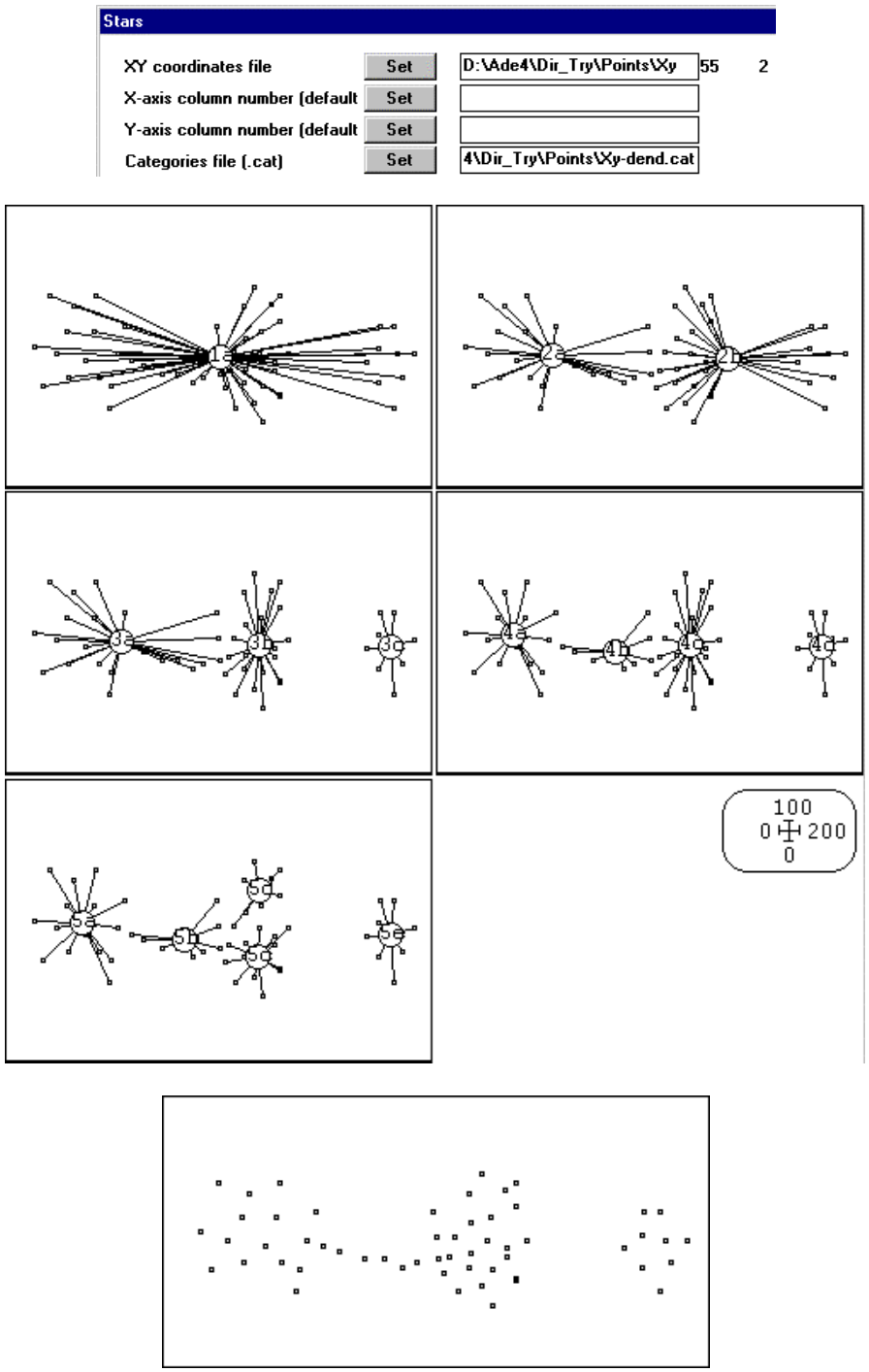
Cette méthode propose de voir dans les données 4 classes. Pour les représenter :




```
Clusters: Compute dendrogram
Hierarchy file: D:\Ade4\Dir_Try\Points\Xy.2dha
Number of rows: 54, columns: 5
Output file: D:\Ade4\Dir_Try\Points\Xy-dend
Number of rows: 55, columns: 6
************************************
*  Description of a coded matrix  *
************************************
Categorical variables: file D:\Ade4\Dir_Try\Points\Xy-dend
Rows: 55, Variables: 6, Categories: 20, Missing data: 0

Description of categories:
-----------------------------------------------
Variable number 1 has 1 categories    Une seule partie contenant tous les points
-----------------------------------------------
[  1]Category:   1 Num:   55 Freq.:         1

Variable number 2 has 2 categories    La partie unique est divisée en 2 (A et B)
-----------------------------------------------
[  2]Category:   1 Num:   24 Freq.:     0.436
[  3]Category:   2 Num:   31 Freq.:     0.564

Variable number 3 has 3 categories    B donne C et D et la partition A, C, D
-----------------------------------------------
[  4]Category:   1 Num:   24 Freq.:     0.436
[  5]Category:   2 Num:   22 Freq.:       0.4
[  6]Category:   3 Num:    9 Freq.:     0.164

Variable number 4 has 4 categories    A donne E et F et la partition E, F, C, D
-----------------------------------------------
[  7]Category:   1 Num:   15 Freq.:     0.273
[  8]Category:   2 Num:    9 Freq.:     0.164
[  9]Category:   3 Num:   22 Freq.:       0.4
[ 10]Category:   4 Num:    9 Freq.:     0.164

Variable number 5 has 5 categories    etc
-----------------------------------------------
[ 11]Category:   1 Num:   15 Freq.:     0.273
[ 12]Category:   2 Num:    9 Freq.:     0.164
[ 13]Category:   3 Num:   14 Freq.:     0.255
[ 14]Category:   4 Num:    8 Freq.:     0.145
[ 15]Category:   5 Num:    9 Freq.:     0.164

…
-----------------------------------------------
Auxiliary binary output file D:\Ade4\Dir_Try\Points\Xy-dendModa: Indicator vector of
modalities
It contains variable number for each modality
```

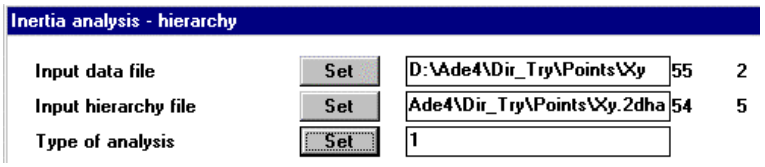
```
It has 20 rows (modalities) and one column

Auxiliary ASCII output file D:\Ade4\Dir_Try\Points\Xy-dend.123: labels (two characters) for
20 modalities
It contains one label for each modality
It has 20 rows (modalities) and labels 1a,1b, ..., 2a, 2b, ...
Variable number 1,2, ..., A, ..., Z,+, Modality number a,b, ..., z,+
------------------------------------------------
```

Les partitions successives sont des variables qualitatives rangées en colonnes dans un fichier de variables qualitatives accompagné du fichier auxiliaire .cat. Pour représenter les classes sur le nuage de points (ScatterClass: Stars) :



En bas, le nuage de départ. En haut, la partition la moins fine (1 seule classe), sa décomposition en deux classes, la décomposition de l'une des classes en deux classes pour minimiser la diminution de la variance inter, etc

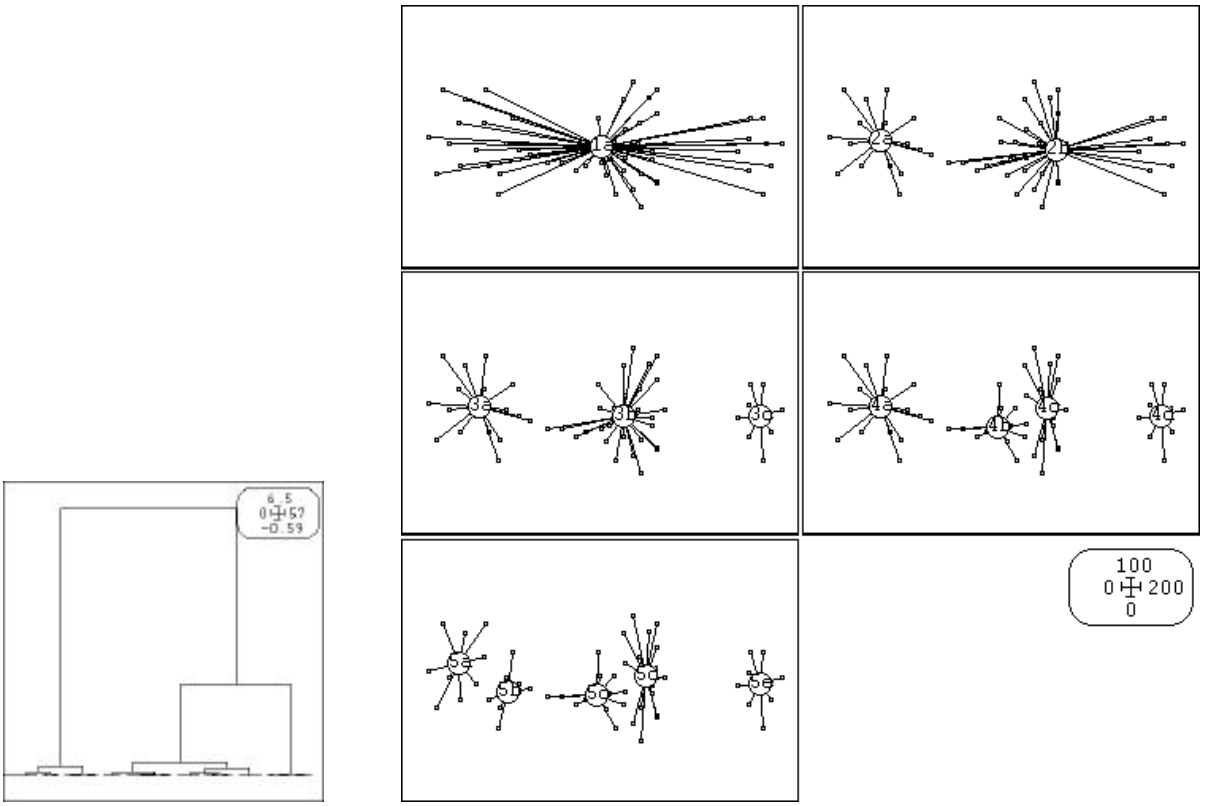
Pour connaître le rôle de chacune des variables dans les niveaux d'une hiérarchie (Clusters: Inertia analysis - hierarchy) :




```
Clusters: Inertia analysis
Interpretation of a hierarchy - Continuous variables
Input file : D:\Ade4\Dir_Try\Points\Xy
Hierarchy file : D:\Ade4\Dir_Try\Points\Xy.2dha

Contribution of variables to the nodes of the hierarchy:

Variables :         1      2
--------------------
Node # 56 :         4     96
Node # 57 :         0    100
Node # 58 :        64     36
…
Node # 103 :       36     64
Node # 104 :       99      1
Node # 105 :        0    100
Node # 106 :        1     99
Node # 107 :       99      1
Node # 108 :      100      0
Node # 109 :      100      0
```

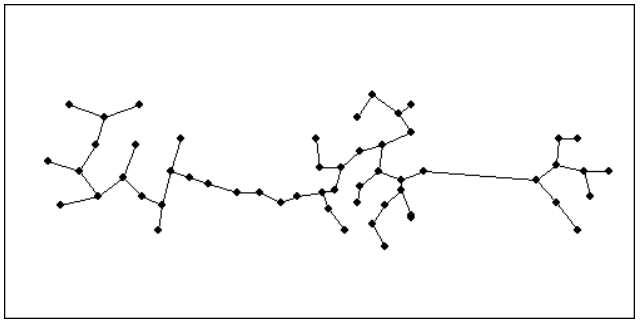
Les trois premiers decoupage se font sur la variable 1 (x) et donne 4 classes.

La classification ascendante (Clusters: Compute hierarchy : Ward method) donne avec les mêmes enchaînements :

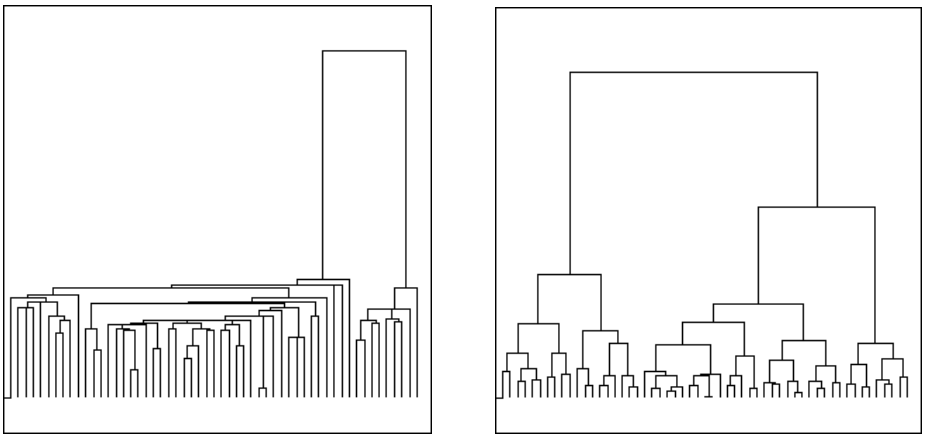


La méthode propose trois classes. Quand on monte en regroupant et quand on descend en divisant, on ne passe pas aux mêmes endroits comme dans toutes les méthodes pas à pas simples.

Lien Pour représenter l'arbre de longueur minimale; utiliser DMAUtil: Canonical distance, NGStat: Minimal Spanning Tree et Scatters: Neighbouring relationship :



Ce point de vue est compatible avec celui du lien simple (saut minimum, à gauche) qui voit deux classes (l'exemple est proposé par ses auteurs pour illustrer l'effet de chaîne) :



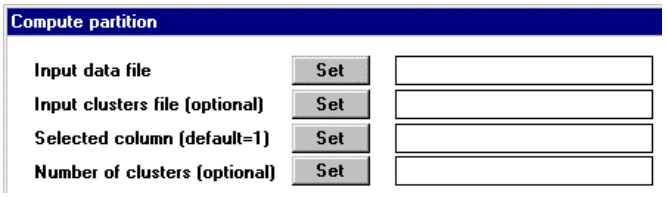
A gauche dendrogramme du lien simple, à droite celui du lien complet.

Clusters : Compute partition

Type Méthode de classification dite autour des centres mobiles.

Objet On peut trouver une description de l'intérêt et du principe de la méthode dans [1]. Elle fournit une partition dont on a fixé à l'avance *p* le nombre de classes. Elle prend en entrée une partition initiale en *p* classes qui, si elle n'est pas précisée par l'utilisateur, est générée aléatoirement. Elle propose en sortie une partition en *p* classes. Le principe général est qu'à chaque pas on calcule le centre de gravités des classes et on réaffecte chaque point au centre de classe le plus proche. La méthode converge vers un optimum local qui dépend du point de départ.

Dialogue L'option utilise une seule fenêtre de dialogue :



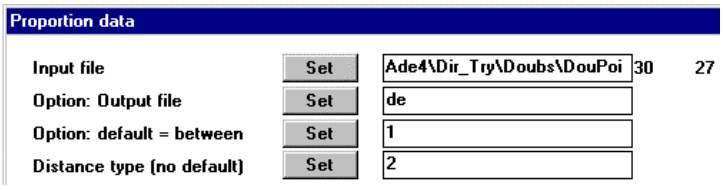
1) Nom du fichier binaire d'entrée.

2) Option : nom du fichier binaire contenant la partition initiale. Le nombre de classe de la partition initiale définit dans ce cas le nombre de classes de la partition finale. Si ce nom de fichier est omis, la partition initiale sera aléatoire mais on doit indiquer le nombre de classes dans la dernière boîte de dialogue.

3) Numéro de la colonne sélectionnée dans le fichier précédent s'il existe. Par défaut, c'est la première.

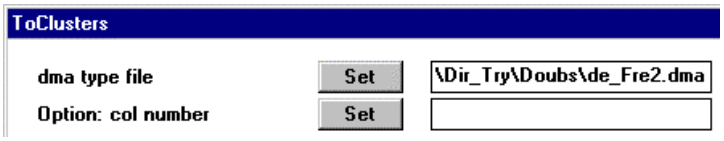
4) Option : nombre de classes désirées. Par défaut, c'est le nombre de classes de la partition initiale. Si celle-ci est omise, il est impérativement précisé ici.

Exemple Utiliser le carte Doubs [1]. Calculer la distance entre espèces avec l'indice de chevauchement de niche (DMAUtil: Proportion data) :

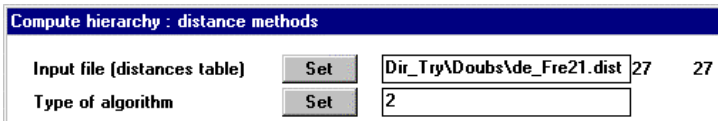



```
d2 distances computed
Manly 1994 Multivariate statistical methods. A primer
2nd edition. Chapman & Hall 1994. formula 5.8 p. 68
Test of the euclidean property by diagonalization (theorem of GOWER)
Output file: de_Fre2
It has 351 rows and 1 columns
d(2,1), d(3,1), d(3,2),..., d(n,1), d(n,2), ... d(n,n-1)
Text file: de_Fre2.dma
  1 -> 27
  2 -> 1
  3 -> d2 of MANLY on D:\Ade4\Dir_Try\Doubs\DouPoi
  4 -> FALSE
```

Préparer la matrice .dist pour le présent module (Voir Clusters: Compute distances) :

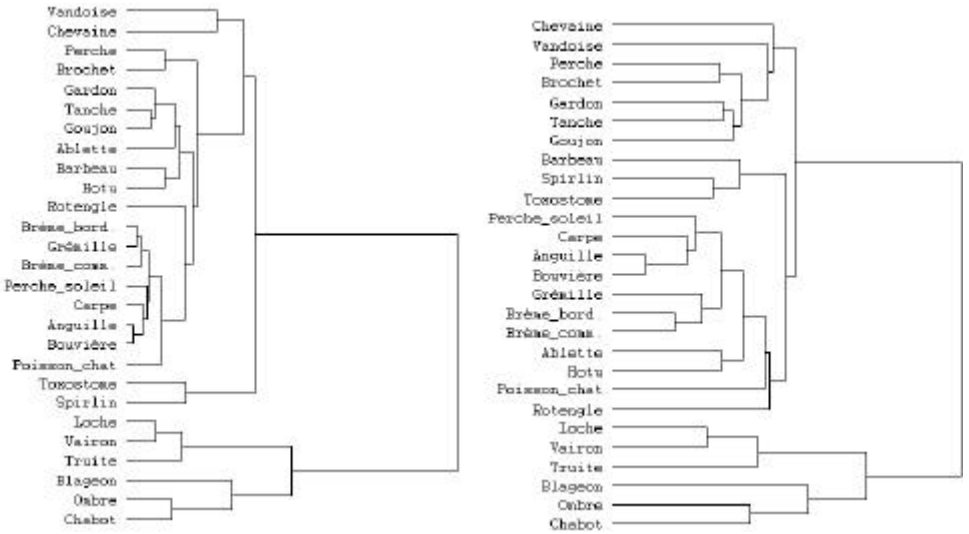



```
Output file : D:\Ade4\Dir_Try\Doubs\de_Fre21.dist Row: 27 Col: 27
Tranformation: rescaling on [0,1] by y=(x-min)/(max-min)
```



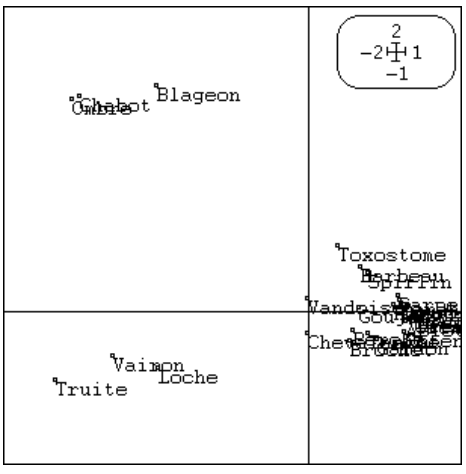

```
Clusters: Compute hierarchy
Distance file: D:\Ade4\Dir_Try\Doubs\de_Fre21.dist
Number of rows: 27, columns: 27
Output file: D:\Ade4\Dir_Try\Doubs\de_Fre21.alha
Number of rows: 26, columns: 5
Hierarchy algorithm used : average link (UPGMA)
```

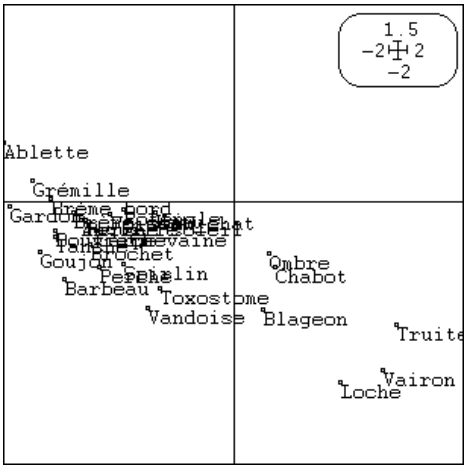
Tracer le dendrogramme (à gauche) :



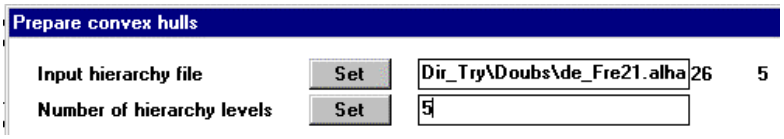
On reconnaît la séparation salmonidés/cyprinidés au niveau 1, puis Truite/Vairon/Loche contre Ombre/Blageon/Chabot au niveau 2.

Essayer l'option 5 (Distance d'Edwards) dans DMAUtil: Proportion data et refaire le dendrogramme pour le lien moyen (à droite). On reconnaît les mêmes éléments principaux mais le reste est ambigu. Faire l'analyse des correspondances du tableau (plan 1-2 sur Doupoi.fcco, les espèces sont en colonnes). Faire de même avec une ACP centrée :

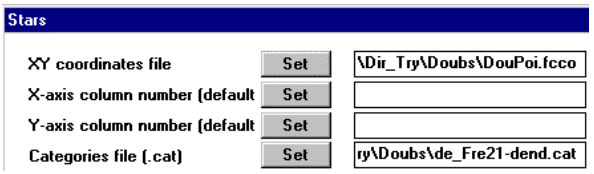


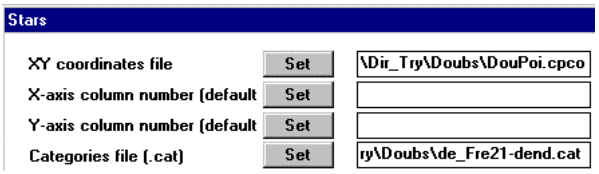


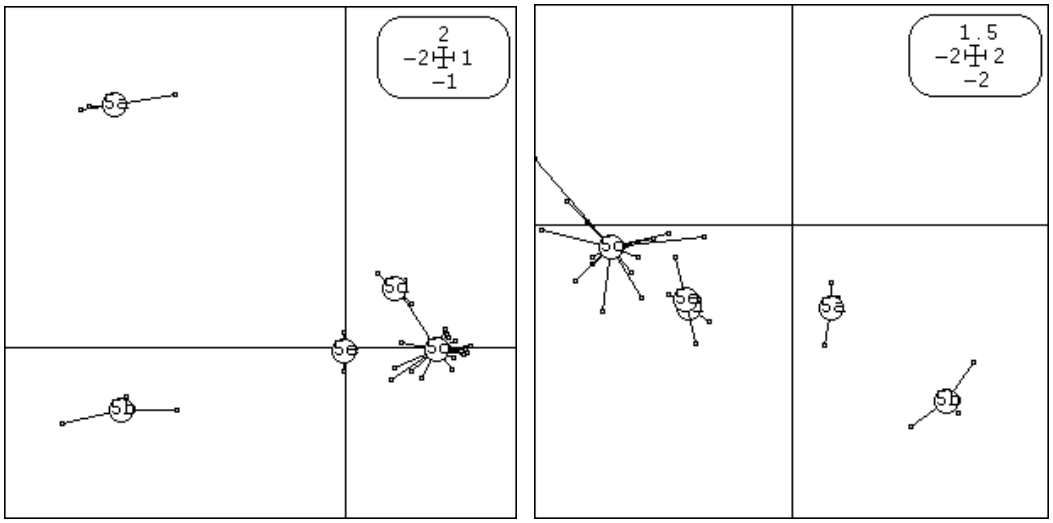
Pour reporter sur une carte factorielle le niveau d'une hiérarchie, passer par Clusters: Prepare convex hulls :



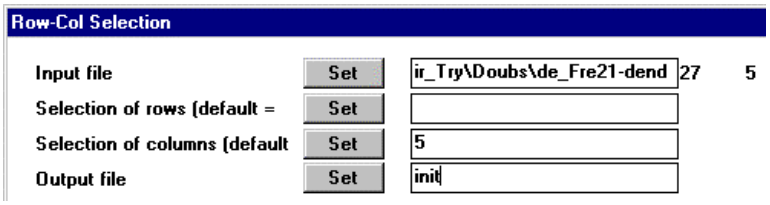

```
Categorical variables: file D:\Ade4\Dir_Try\Doubs\de_Fre21-dend
Rows: 27, Variables: 5, Categories: 15, Missing data: 0

Description of categories:
-----------------------------------------------
Variable number 1 has 1 categories
-----------------------------------------------
[  1]Category:   1 Num:   27 Freq.:         1

Variable number 2 has 2 categories
-----------------------------------------------
[  2]Category:   1 Num:    6 Freq.:     0.222
[  3]Category:   2 Num:   21 Freq.:     0.778

…
Variable number 5 has 5 categories
-----------------------------------------------
[ 11]Category:   1 Num:    3 Freq.:     0.111
[ 12]Category:   2 Num:    3 Freq.:     0.111
[ 13]Category:   3 Num:   17 Freq.:      0.63
[ 14]Category:   4 Num:    2 Freq.:    0.0741
[ 15]Category:   5 Num:    2 Freq.:    0.0741
```



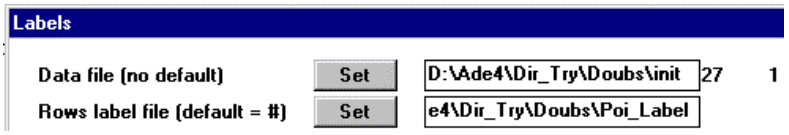




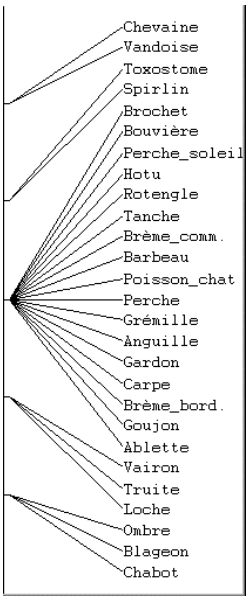
Les points de vue sont très différents mais cohérents. La partition à 5 classes de la hiérarchie est un bon point de départ pour une classification. Conserver cette variable dans un fichier init :



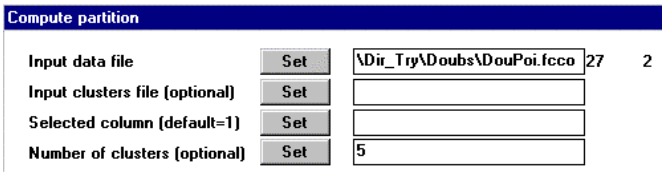
Dans Graph1D



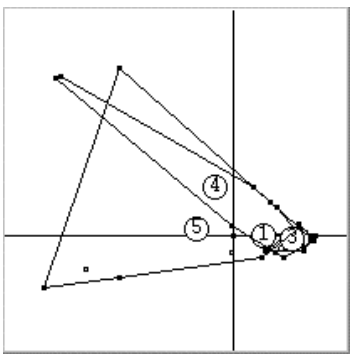

```
Binary input file: D:\ADE4\DIR_TRY\DOUBS\init
   1 |  1.0000
   2 |  2.0000
   3 |  2.0000
   4 |  2.0000
   5 |  1.0000
   6 |  1.0000
   7 |  3.0000
   8 |  4.0000
   9 |  5.0000
  10 |  5.0000
  11 |  3.0000
  12 |  4.0000
  13 |  3.0000
  14 |  3.0000
  15 |  3.0000
  16 |  3.0000
  17 |  3.0000
  18 |  3.0000
  19 |  3.0000
  20 |  3.0000
  21 |  3.0000
  22 |  3.0000
  23 |  3.0000
  24 |  3.0000
  25 |  3.0000
  26 |  3.0000
  27 |  3.0000
```



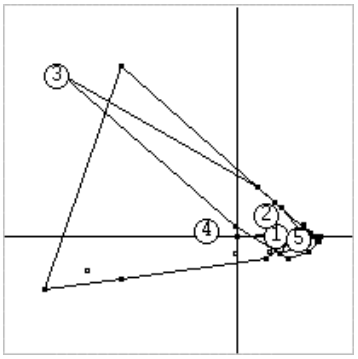
Important La procédure de classification part d'une répartition des objets en *p* classes. Il vaut toujours mieux partir d'une information préalable plutôt qu'au hasard. Par exemple :



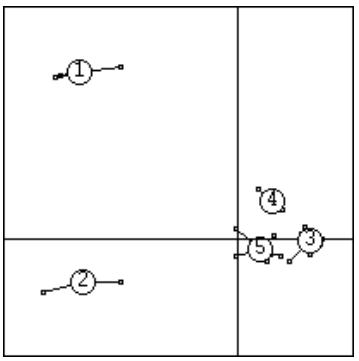

```
Final partition - List of elements in each cluster :
1: 22, 27, 28, 29, 30,
2: 19, 20, 21, 26,
3: 15, 16, 17, 18,
4: 1, 5, 8, 9, 23, 24, 25,
5: 2, 3, 4, 6, 7, 10, 11, 12, 13, 14 -> Essai 1
```



Essai 1



Essai 2



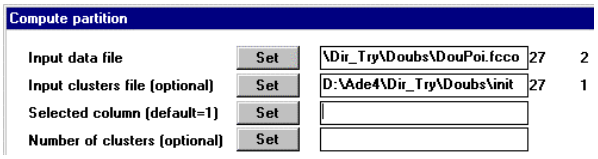
Essai 3

Si on recommence avec les mêmes paramètres, on converge vers un autre minimum local :

```
Final partition - List of elements in each cluster :
1: 21, 22, 27, 28, 29, 30,
2: 1, 8, 9, 16, 17, 18, 23, 24, 25,
3: 5,
4: 2, 3, 4, 6, 7, 10, 11, 12, 13, 14, 15,
```

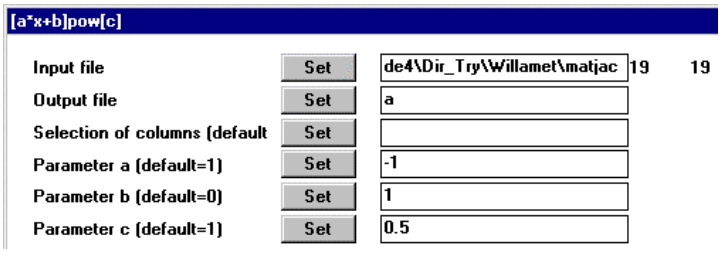
```
5: 19, 20, 26, -> Essai 2
```

Si on recommence avec une bonne partition initiale, on obtient un tout autre résultat :




```
Final partition - List of elements in each cluster :
1: 1, 5, 6,
2: 2, 3, 4,
3: 7, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27,
4: 8, 11, 12,
5: 9, 10, 13, 14, 15, 20, -> Essai 3
```

Lien Utiliser la carte Willamet [3]. Elle donne une matrice de similarité de Jaccard (matjac). La transformer en matrice de distances par $s \mapsto d = \sqrt{1-s}$:

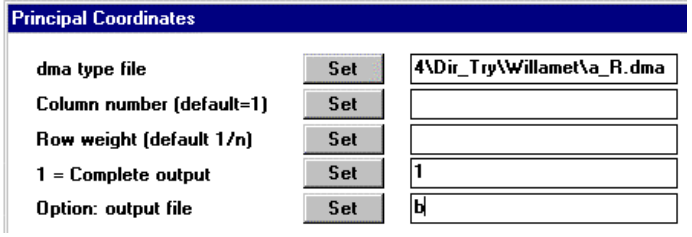


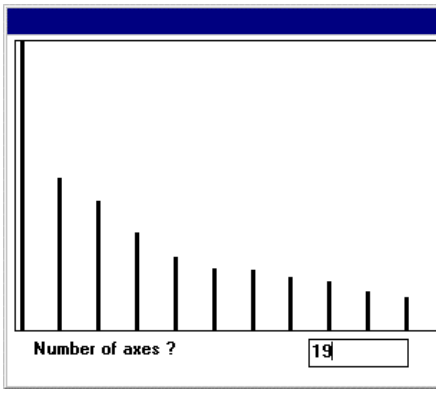
Lire le tableau a comme matrice de distances (DMAUtil) :


```
Test of the euclidean property by diagonalization (theorem of GOWER)
Output file: D:\Ade4\Dir_Try\Willamet\a_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: D:\Ade4\Dir_Try\Willamet\a_R.dma
  1 -> 19
  2 -> 1
  3 -> Input distance file D:\Ade4\Dir_Try\Willamet\a
  4 -> TRUE
```

On trouve une matrice de distances euclidiennes. Si la précision numérique n'avait pas été suffisante pour obtenir ce résultat (à cause des formats d'édition de l'article cité) on aurait utilisé DMAUtil: Quasi_Euclidean. Ce n'est pas nécessaire.

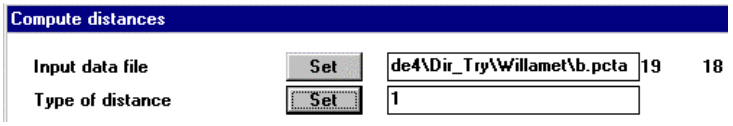
On peut donc faire une analyse en coordonnées principales (DMAUse) :





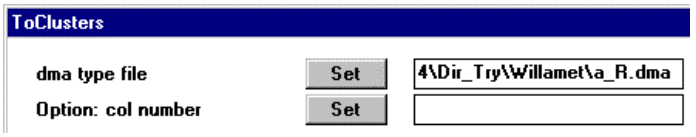
```
File b.pcta contains the principal coordinates  (norm=sqrt(lambda))
--- It has 19 rows and 18 columns
```

On garde à dessein **tous les axes**. Les distances euclidiennes euclidienne entre les lignes du tableau b.pcta sont exactement les distances de Jaccard. Calculer ces distances avec Clusters :




```
-------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\WILLAMET\b.pcta.dist - 19 rows, 19 cols.
   1 |  0.0000  0.6956  0.7332  0.8032  0.8032  0.8860  0.8614  0.8980  0.8980  0.8860
0.8980  0.9616  0.9275  0.8860  0.9217  1.0000  0.9783  0.9727  0.9333
   2 |  0.6956  0.0000  0.4864  0.5957  0.7332  0.7332  0.7332  0.7690  0.8488  0.7829
0.8488  0.9217  0.8860  0.8488  0.8738  0.9217  0.8738  0.8980  0.8676
…
```

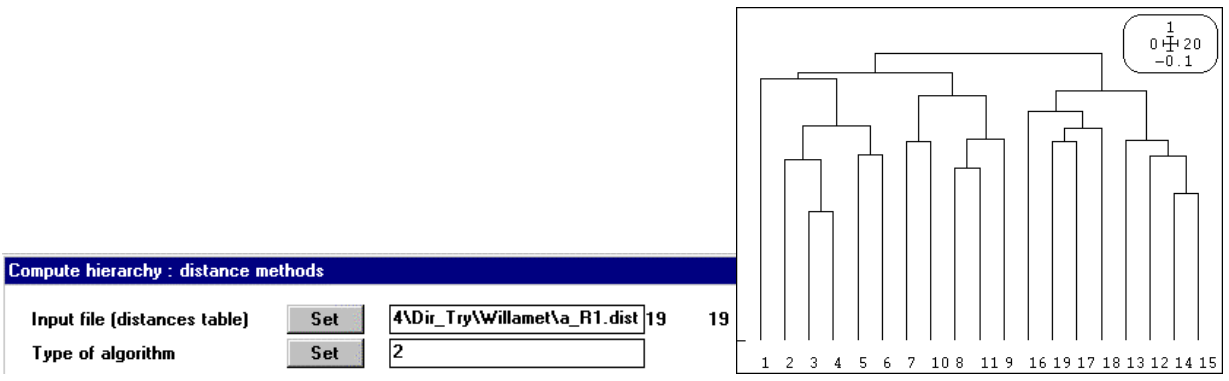
Envoyer la matrice observée à Clusters (DMAUtil) :



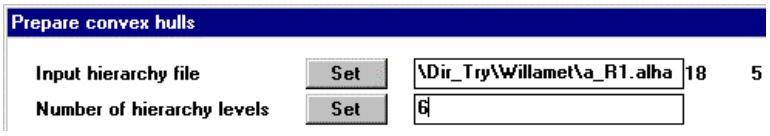

```
Output file : D:\Ade4\Dir_Try\Willamet\a_R1.dist Row: 19 Col: 19
Tranformation: rescaling on [0,1] by y=(x-min)/(max-min)
-------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\WILLAMET\a_R1.dist - 19 rows, 19 cols.
   1 |  0.0000  0.6956  0.7332  0.8032  0.8032  0.8860  0.8614  0.8980  0.8980  0.8860
0.8980  0.9616  0.9275  0.8860  0.9217  1.0000  0.9783  0.9727  0.9333
   2 |  0.6956  0.0000  0.4864  0.5957  0.7332  0.7332  0.7332  0.7690  0.8488  0.7829
0.8488  0.9217  0.8860  0.8488  0.8738  0.9217  0.8738  0.8980  0.8676
…
```

Nous avons donc fabriqué un tableau artificiel qui reproduit exactement les distances entre individus de la matrice de départ. Ainsi tout tableau donne une matrice de distance et toute matrice de distances (euclidienne) donne un tableau qui ont les mêmes propriétés.

Faire un lien moyen sur la matrice de distances :

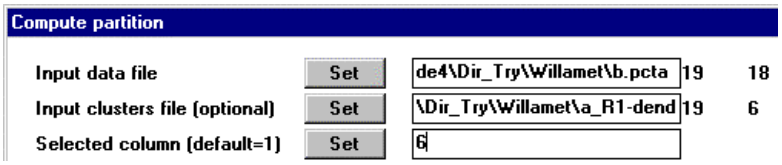


Garder 6 classes provisoires :



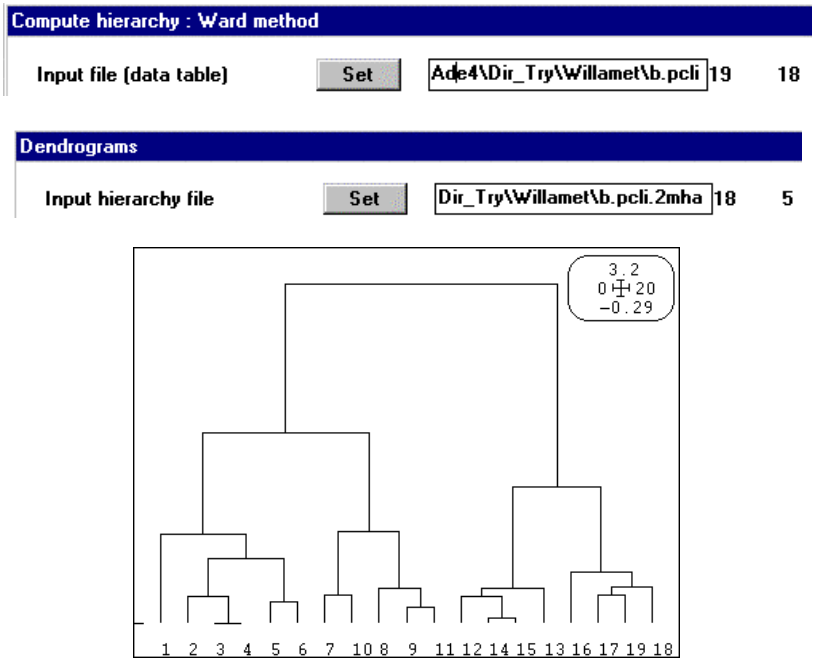

```
Categorical variables: file D:\Ade4\Dir_Try\Willamet\a_R1-dend
Rows: 19, Variables: 6, Categories: 21, Missing data: 0
```

Faire une classification des sites :

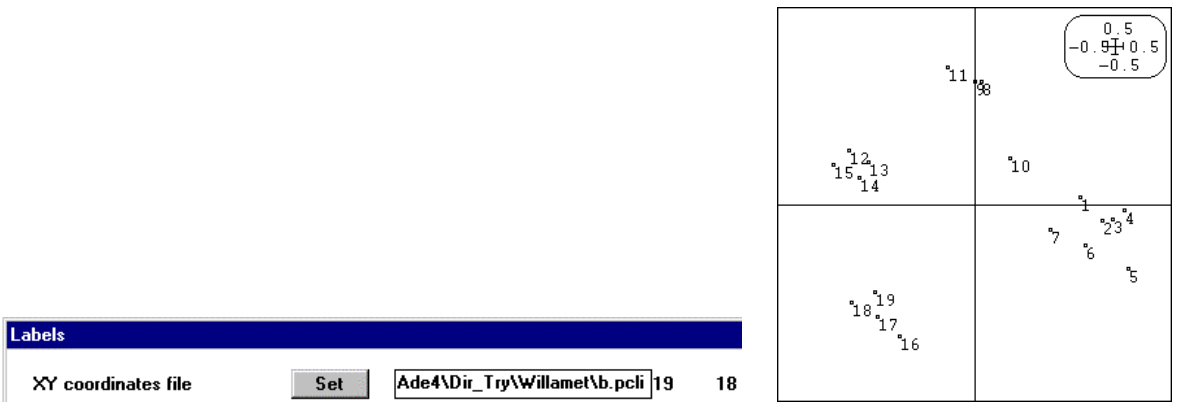


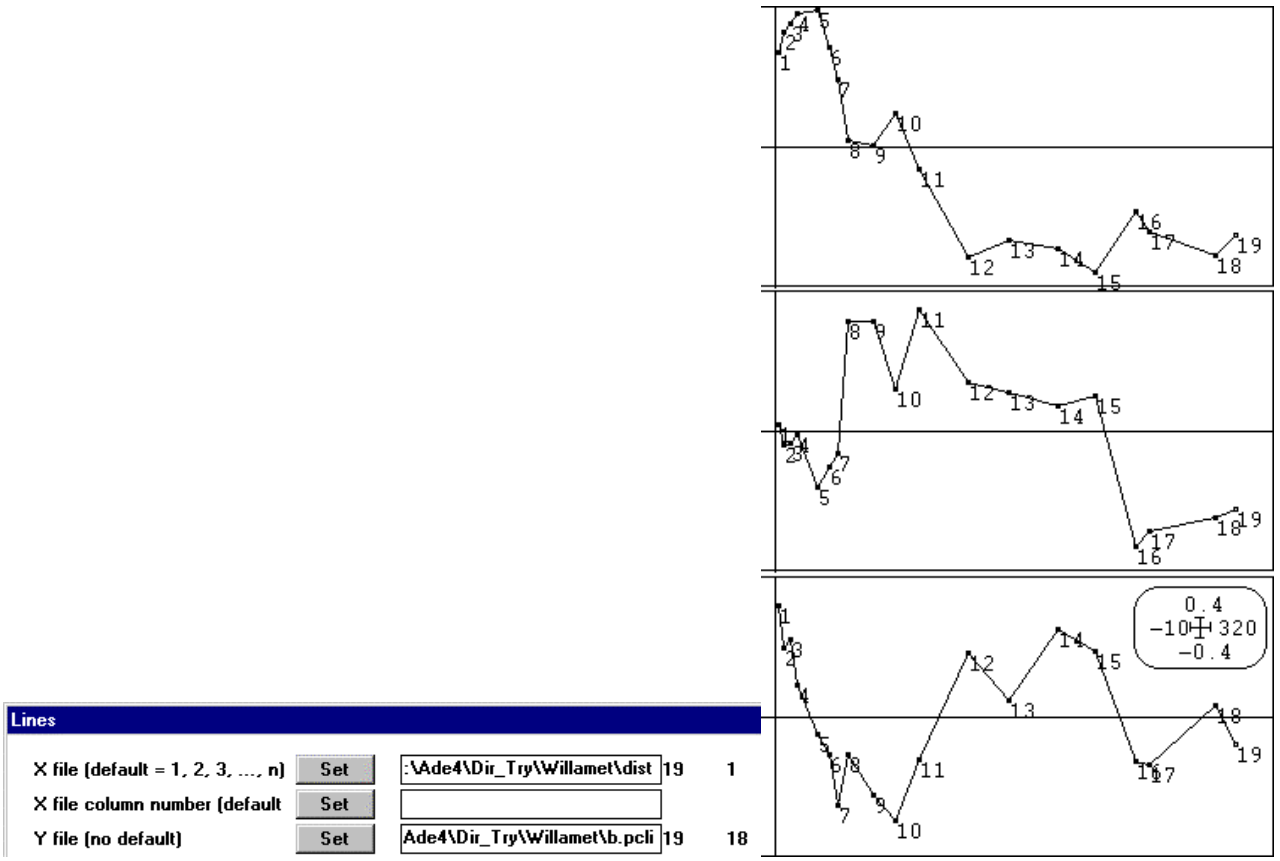

```
Initial partition - List of elements in each cluster :
1: 1,
2: 2, 3, 4, 5, 6,
3: 7, 10,
4: 8, 9, 11,
5: 12, 13, 14, 15,
6: 16, 17, 18, 19,

Final partition - List of elements in each cluster :
1: 1,
2: 2, 3, 4, 5, 6,
3: 7, 10,
4: 8, 9, 11,
5: 12, 13, 14, 15,
6: 16, 17, 18, 19
```

La partition n'a pas été modifiée. La méthode de Ward donne le dendrogramme :



qui pousse à une partition en 4 classes assez voisine de la précédente.





Tracer enfin les coordonnées des trois premiers axes de l'analyse en coordonnées principales en fonction de la distance à la source (fichier dist issu de la carte). La première partie 1-7 est en même temps originale et possède un ordre interne (axe 3). Les parties 12-15 et 16-19 sont clairement identifiées. 8-11 est une classe charnière qui explique les hésitations sur le groupe 7-10.

Retenir de ce qui précède que classification et ordination sont *nécessaires* pour rendre compte de l'ensemble de la structure de la matrice de similarité et qu'elles sont simultanément *possibles* et *complémentaires*.

La procédure des moyennes mobiles tolère un très grand nombre d'individus. Elle est particulièrement utile pour trouver les *formes fortes*. Ce sont, quand elles existent, les classes que l'on retrouve quand on fait successivement plusieurs partitions aleatoires initiales. On les appelle aussi *groupements stables*. Voir [1] p. 152-154.

Références

1 Lebart, L., Morineau, A. & Piron, M. (1995) Statistique exploratoire multidimensionnelle. Dunod, Paris. 1-439. Section 2.1 Agrégation autour des centres mobiles.

2 Verneaux, J. (1973) Cours d'eau de Franche-Comté (Massif du Jura). *Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie*. Thèse d'état, Besançon. 1-257.

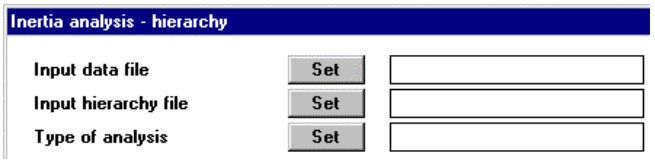
3 Van_Sickle, J. (1997) Using mean similarity dendrograms to evaluate classifications. *Journal of Agricultural, Biological, and Environmental Statistics* : 4, 370-388.

Clusters : Inertia analysis - hierarchy

Type Utilitaire de dépouillement d'une CAH.

Objet Pour interpréter en terme de variables une hiérarchie basée sur l'inertie inter-classe, on peut décomposer le niveau d'un noeud de la hiérarchie (l'inertie inter-classe atteinte par le dernier regroupement) entre les variables pour mesurer leur participation à ce niveau.

Dialogue L'option utilise une seule fenêtre de dialogue :



1) Nom du fichier binaire des données.

2) Nom du fichier de la hiérarchie.

3) Type de variables dans le tableau de données : taper 1 si le tableau de données contient des variables quantitatives et 2 si ce sont des variables qualitatives.

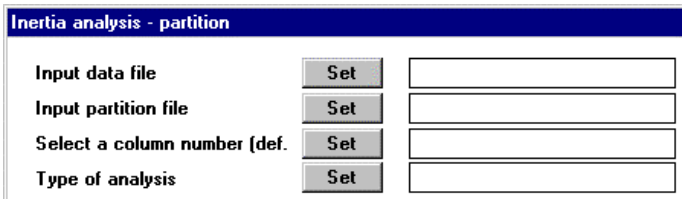
Exemple Voir Clusters: Compute hierarchy : divisive algorithm.

Clusters : Inertia analysis - partition

Type Utilitaire de statistique descriptive.

Objet Pour interpréter en terme de variables une partition, on peut décomposer l'inertie inter-classe entre les variables et entre les classes. L'option s'apparente à DDUtil : Columns/Inertia analysis pour l'ordination.

Dialogue L'option utilise une seule fenêtre de dialogue :



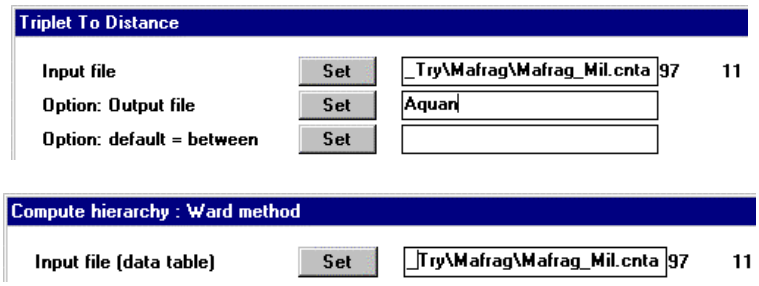
1) Nom du fichier binaire du tableau de données.

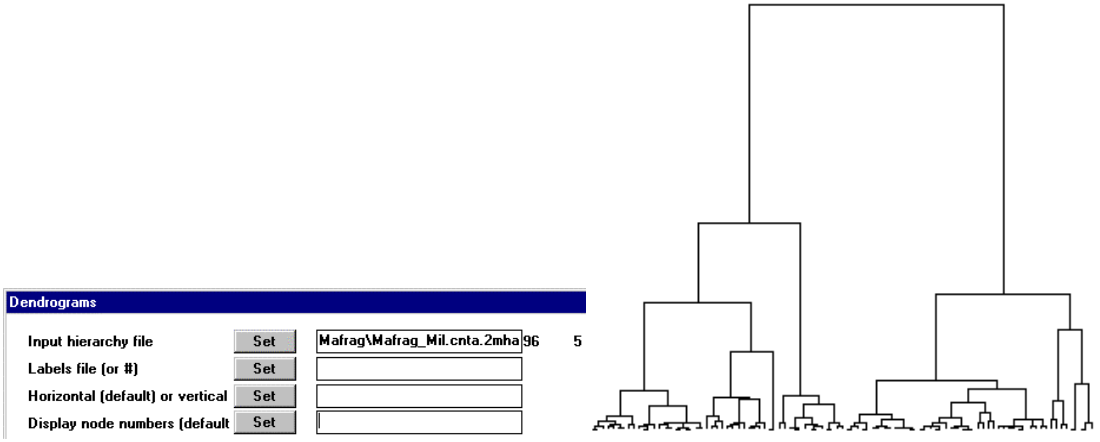
2) Nom du fichier binaire contenant la partition. Ce fichier a le même nombre de lignes que le précédent et contient dans une de ses colonnes une variable qualitative indiquant pour chaque individu le numéro de la classe à laquelle il appartient.

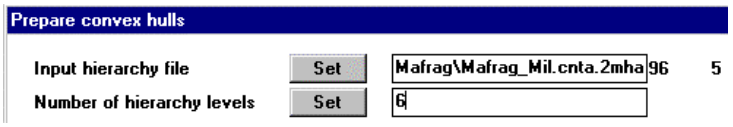
3) Numéro de colonne sélectionnée dans le précédent fichier (par défaut c'est la première).

4) Type de variables dans le tableau de données : taper 1 si le tableau de données contient des variables quantitatives et 2 si ce sont des variables qualitatives.

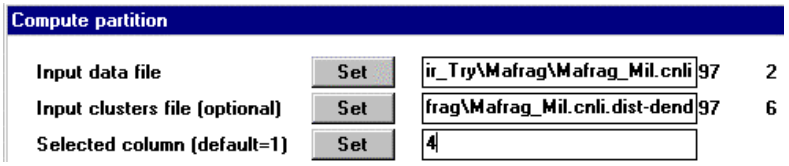
Exemple Utiliser la carte Mafragh+1. Faire l'ACP normée de Mafragh_Mil. Calculer la hiérarchie associée au tableau normalisé par la méthode de Ward, tracer le dendrogramme et conserver 6 niveaux de partition :



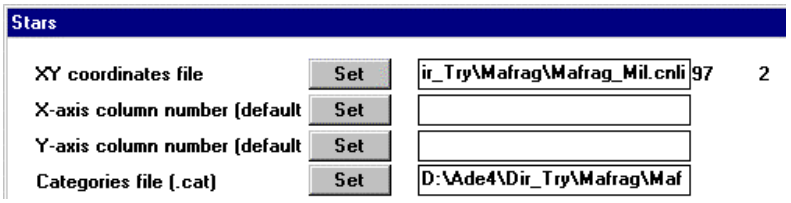


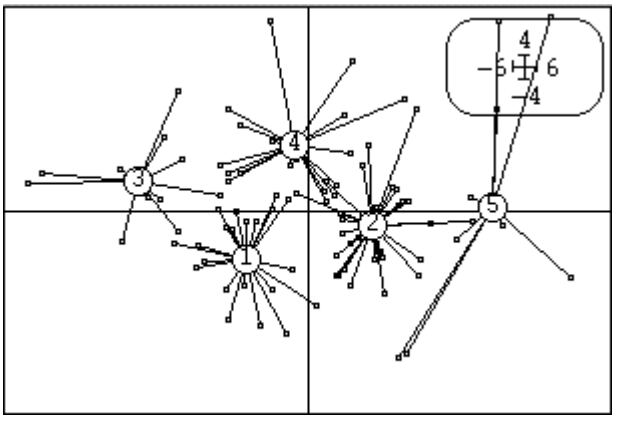


Faire une classification par moyennes mobiles à partir de la partition en 5 classes de la hiérarchie précédente :

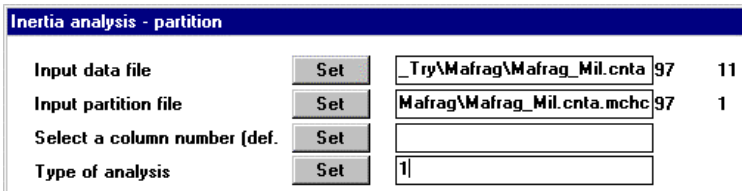



```
Qualitative variables file: D:\Ade4\Dir_Try\Mafrag\Mafrag_Mil.cnta.mchc
Number of rows: 97, variables: 1, categories: 5
```





Important On peut se dire que, sur la carte qui précède, la partition est bizarre. On aurait pu faire mieux. C'est oublier que, sur la carte on voit un nuage projeté et les distances entre points sont des approximations des distances dans l'espace. La classification utilise les distances totales, la projection en donne une approximation. L'écart entre les deux points de vue est un élément important de l'interprétation. La classe 5 est plus homogène dans l'espace que ne laisse croire la projection. Par contre le facteur 2 indique des différences importantes au centre du premier facteur. Si on prend la distance, non pas sur .cnta, mais sur .cnli, la distance représentée sur le plan est utilisée pour faire la classification et il y a redondance inutile.




```
Clusters: Inertia analysis
Interpretation of clusters - Continuous variables
Input file : D:\Ade4\Dir_Try\Mafrag\Mafrag_Mil.cnta
Clusters file : D:\Ade4\Dir_Try\Mafrag\Mafrag_Mil.cnta.mchc
```

La classe de chacun des points est indiquée :

```
Cluster number for each element (elt nb : clust. nb) :
1:1, 2:2, 3:2, 4:3, 5:1, 6:2, 7:4, 8:4, 9:2, 10:4,
11:1, 12:1, 13:5, 14:2, 15:2, 16:1, 17:1, 18:4, 19:4, 20:2,
21:2, 22:2, 23:2, 24:4, 25:4, 26:3, 27:4, 28:3, 29:4, 30:4,
31:2, 32:3, 33:1, 34:3, 35:3, 36:4, 37:4, 38:1, 39:1, 40:2,
41:2, 42:2, 43:2, 44:5, 45:5, 46:5, 47:5, 48:5, 49:2, 50:2,
51:2, 52:2, 53:2, 54:2, 55:2, 56:2, 57:5, 58:2, 59:2, 60:2,
61:3, 62:3, 63:1, 64:3, 65:3, 66:2, 67:5, 68:2, 69:4, 70:4,
71:5, 72:2, 73:2, 74:5, 75:4, 76:4, 77:4, 78:1, 79:4, 80:1,
81:1, 82:1, 83:1, 84:1, 85:3, 86:2, 87:3, 88:1, 89:1, 90:1,
91:1, 92:1, 93:1, 94:1, 95:3, 96:4, 97:1,
```

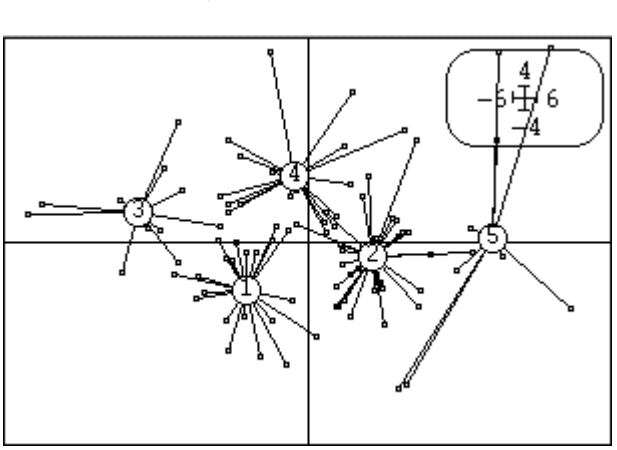
Le contenu de chaque classe est édité :

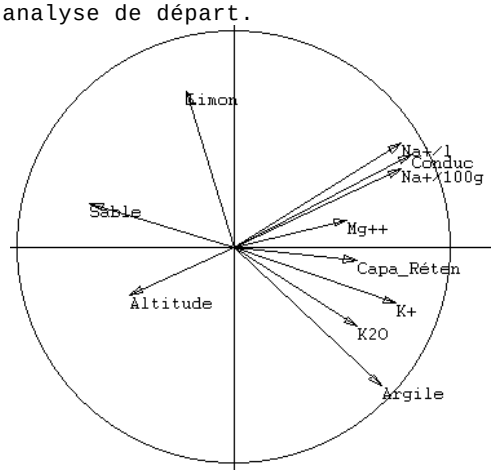
```
List of elements in each cluster :
1: 1, 5, 11, 12, 16, 17, 33, 38, 39, 63, 78, 80, 81, 82, 83, 84, 88, 89, 90, 91, 92, 93, 94, 97,
2: 2, 3, 6, 9, 14, 15, 20, 21, 22, 23, 31, 40, 41, 42, 43, 49, 50, 51, 52, 53, 54, 55, 56, 58, 59,
60, 66, 68, 72, 73, 86,
3: 4, 26, 28, 32, 34, 35, 61, 62, 64, 65, 85, 87, 95,
4: 7, 8, 10, 18, 19, 24, 25, 27, 29, 30, 36, 37, 69, 70, 75, 76, 77, 79, 96,
5: 13, 44, 45, 46, 47, 48, 57, 67, 71, 74,

Total moment               : 1067.0004
Between-clusters moment    : 537.8963      R = 0.5041
```

L'inertie totale présente 50 % d'inertie inter-classe (on retrouve se résultat en faisant l'ACP inter-classe). Le nuage des centres de gravité des classes a une inertie totale de 1067. On peut la décomposer en somme pondérée de carrés de distances pour toutes les variables et toutes les classes et l'exprimer en pourcentage soit par variable soit par classe :

```
Contribution of variables to clusters:
Var. :          1      2      3      4      5      6      7      8      9     10     11
-------------------------------------------------------------------------------------
Clust 1         3     -1     -2     -1     -7    -17    -10    -18     -2    -16     22
Clust 2        22    -12     -8      0      0      7     12     12      4      8    -14
Clust 3       -19      0     38     -5     -2     -6     -6     -7    -12     -5      1
Clust 4       -18     65     -3     -2      1      1     -4      0      7      0      0
Clust 5         4     -2     -1     19     17     11     16     12      2     13     -3

Explanation of clusters by variables:
Var. :          1      2      3      4      5      6      7      8      9     10     11
-------------------------------------------------------------------------------------
Clust 1         3     -2     -2     -2    -14    -27    -13    -24     -5    -24     50
Clust 2        22    -15     -7      0      0     11     16     16      8     12    -32
Clust 3       -51     -1     86    -22     -9    -23    -21    -24    -66    -20      4
Clust 4       -16     77     -2     -2      1      1     -4      0     13      0      0
Clust 5         8     -5     -3     74     75     38     46     36      7     44    -14

Code des variables (1_Argile, 2_Limon, 3_Sable, 4_K2O, 5_Mg++, 6_Na+/100g, 7_K+, 8_Conduc,
9_Capa_Réten, 10_Na+/l, 11_Altitude).
Clust 1     Points hauts à salinité faible
Clust 2     Points bas argileux
Clust 3     Sols sableux  ne retenant pas l'eau et lessivés
Clust 4     Sols limoneux
Clust 5     Sols riches et fortement salés
```

Pour mémoire, cercle des corrélations de l'analyse de départ.



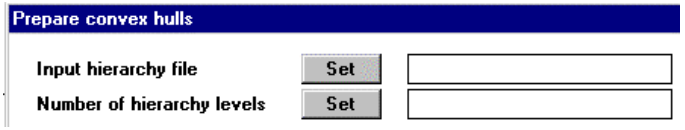


Clusters : Prepare convex hulls

Type Utilitaire d'emploi d'une hiérarchie de classes.

Objet Les options Clusters: Compute hierarchy : distance methods-Compute hierarchy, Ward method et Compute hierarchy : divisive algorithm calcule une hiérarchie de classes indicées. Le dernier niveau de la hiérarchie est la classe unique contenant tous les éléments (rang 1 : partition triviale à une classe). Cette classe unique est le regroupement des 2 classes présentes à l'avant dernier niveau (rang 2 : partition en deux classes). Au rang *k*, on a une partition en *k* classes. L'option édite le nombre de partition désirées, chacune d'entre elles étant une variable qualitative. Le fichier de sortie est accompagné de ses utilitaires et l'option CategVar : Read Categ File n'est pas nécessaire. Le nom de l'option rappelle qu'on peut tracer sur une carte les polygones de contour des classes, mais le fichier créé est un fichier de variables qualitatives ordinaires utile en diverses circonstances.

Dialogue L'option utilise une seule fenêtre de dialogue :



1) Nom du fichier binaire d'entrée (fichiers créés par Clusters : Compute hierarchy).

2) Nombre de niveaux désirés (*k* indique qu'on obtiendra les partitions à 1, 2, ..., *k* classes).

Exemple Voir Clusters: Compute hierarchy : divisive algorithm.

Voir Clusters: Compute partition.