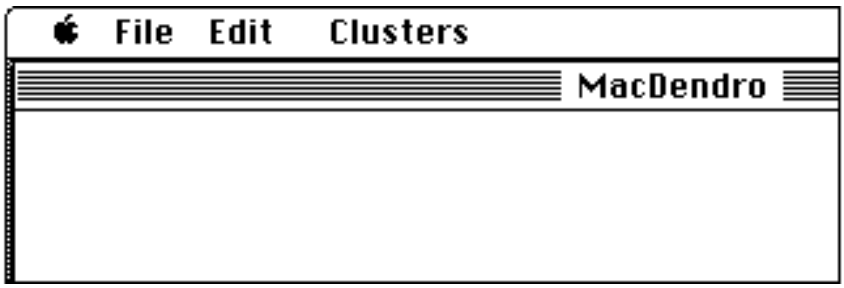
MacDendro

Cluster Analysis



Jean Thioulouse
U.R.A. CNRS 243 "Biométrie, Génétique et Biologie des populations"
Université Lyon 1 - I.A.S.B.S.E.- 69622 Villeurbanne Cedex - France
e-mail : jean@biomac.univ-lyon1.fr - THIOULOU@FRCISM51.BITNET

MacDendro

MacDendro is a Macintosh program for cluster analysis, compatible with MacMul and GraphMu: it uses the same type of binary files. It may be used to compute hierarchies and partitions according to several classical algorithms. GraphMu may be used to draw the graphical display (dendrograms) of hierarchies computed with MacDendro.

Most algorithms used in MacDendro come from the book of M. Roux (1985) *Algorithmes de classification*, Masson, Paris. See also: Roux M. (1991), Basic procedures in hierarchical cluster analysis, in: Devillers J. and Karcher W. (Eds.), *Applied Multivariate Analysis in SAR and Environmental Studies*, p. 115-136, Kluwer Academic Publishers, Dordrecht, The

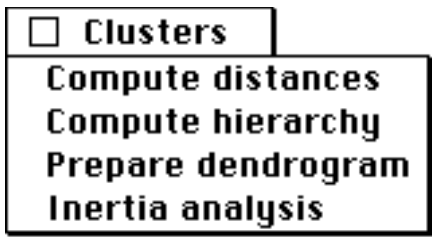
Netherlands.

A review on classical cluster analysis software can be found in Blashfield R.K., Aldenderfer M.S., and Morey L.C., (1982) Cluster Analysis Software, in: Krishnaiah P.R. and Kanal L.N. (Eds.), *Handbook of Statistics*, Vol. 2, p. 245-266, North Holland Publishing Company, Amsterdam.

File Menu

The "**File**" menu has the same commands than in MacMul and GraphMu (mainly transformation between TEXT and BIN file formats). The "**Edit**" menu is of no use: it may only be used with desk accessories (Apple menu).

Clusters Menu



The "**Clusters**" menu offers 4 commands, corresponding to the basic operations of the program. "**Compute distances**" is used to compute the distances between elements, starting from a data table with variables as columns ans elements as rows. "**Compute hierarchy**" is used to compute the hierarchical tree, or the partition. "**Prepare dendrogram**" is used to prepare a special file, containing the cluster number of each element for each level of the hierarchy. This file is to be used with GraphMu to draw the dendrogram or the convex hulls corresponding the clusters of elements. "**Inertia analysis**" is used to produce a text file containing a description of the cluster analysis.

Compute distances

The user can choose between three ways to compute the distances between elements: the classical Euclidean distance, the Chi-2 distance (for contingency tables), and the Jaccard index (for binary - or presence-abscence - data; d=1-c/(p+q-c), with c=number of species present in both samples, p and q =

number of species in each sample). The input file must be a binary file containing the elements as rows and the variables as columns. The name of the output file is built up by adding the ".dist" extension string to the name of the data file. This output file contains the matrix of distances between all elements, and it has a number of rows and columns equal to the number of rows of the data file.

Exemple of use:

In this example, the data file contains the number of death by 6 causes in 19 european countries (Table 1, from Roux, 1985). The death causes are: suicide (Sui), homicide (Hom), road accident (Rout), industry accident (Indu), cirrhosis (Cirh), and Other (Autr).

		Sui	Hom	Rout	Indu	Autr	Cirh
Austria	241	16	330	43	363	325	
France	156	9	225	10	535	328	
Portugal	85	19	349	7	281	345	
Germany		210	12	230	21	298	269
Belgium	156	10	260	13	367	144	
Finland	251	26	180	29	387	55	
Sweden	194	11	151	13	384	122	
Switzerland	225	9	195	26	276	128	
Italy		54	11	219	19	224	319
Ireland (North)	40	136	215	18	320	43	
Denmark		241	6	168	11	230	107
Iceland	101	5	179	23	380	9	
Scotland	82	15	155	18	342	59	
Spain		40	4	136	17	237	225
Norway	104	6	138	22	346	41	
Ireland (South)	38	7	182	32	314	37	
Netherlands	89	7	169	10	218	47	
England-Wales	79	10	130	14	203	36	
USA		121	102	220	26	273	158

Table 1: Number of death by 6 causes in 19 european countries

The correspondence analysis of this data table (19 rows, 6 columns) was first performed with MacMul, and the first 3 factor coordinates of rows (countries) are given in table 2.

F1	F2	F3
.2203	.0064	-.1077
.2101	.0031	.1103
.3693	.2573	.0653
.2452	-.0170	-.1492
.0073	-.0954	.0373
-.2585	-.2700	-.1779
-.0537	-.2142	-.0582
.0150	-.2116	-.2109
.4840	.2873	.0896
-.7265	.6910	-.0476
.0214	-.2885	-.3343
-.3281	-.2828	.2414
-.2146	-.1087	.2025
.3924	.1777	.1833
-.2345	-.2501	.1761
-.2424	-.1003	.3793
-.1330	-.1425	.0682
-.2000	-.1408	.0652
-.2526	.4466	-.1946

Table 2: First 3 factor scores of the correspondence analysis for the 19 countries

The distances between countries were computed with MacDendro, starting from the first 3 factor coordinates, and using the Euclidean distance (Table 3). Using the correspondence

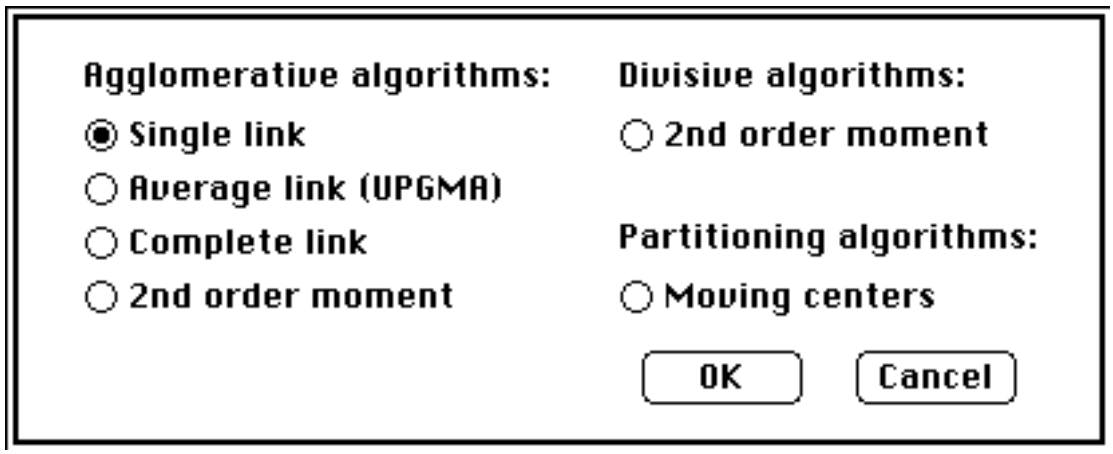
analysis factor scores instead of the real data provides a way to smooth the dataset by eliminating the variations which are not taken into account by the factor scores. To compute the distances between countries directly from the data, we should have used the Chi-2 distance on table 1, instead of the Euclidean distance on table 2.

.00	.17	.26	.04	.22	.43	.28	.25	.34	.91	.33	.55	.43	.30	.46	.53	.33	.37	.51
.17	.00	.24	.20	.18	.48	.30	.34	.31	.91	.44	.49	.35	.20	.40	.42	.29	.34	.55
.26	.24	.00	.29	.39	.67	.50	.51	.09	.92	.59	.70	.55	.11	.62	.60	.50	.54	.55
.04	.20	.29	.00	.24	.44	.29	.24	.35	.94	.31	.58	.46	.32	.49	.56	.35	.40	.53
.22	.18	.39	.24	.00	.30	.13	.21	.48	.84	.33	.34	.22	.39	.25	.33	.12	.17	.50
.43	.48	.67	.44	.30	.00	.19	.22	.75	.84	.25	.33	.32	.68	.28	.45	.24	.22	.56
.28	.30	.50	.29	.13	.19	.00	.13	.58	.88	.23	.32	.25	.50	.23	.38	.13	.16	.55
.25	.34	.51	.24	.21	.22	.13	.00	.58	.92	.11	.45	.38	.52	.36	.51	.25	.28	.55
.34	.31	.09	.35	.48	.75	.58	.58	.00	1.0	.66	.78	.63	.13	.70	.68	.59	.63	.63
.91	.91	.92	.94	.84	.84	.88	.92	1.0	.00	.99	.85	.77	.98	.85	.80	.80	.77	.43
.33	.44	.59	.31	.33	.25	.23	.11	.66	.99	.00	.52	.48	.61	.45	.61	.35	.37	.62
.55	.49	.70	.58	.34	.33	.32	.45	.78	.85	.52	.00	.16	.67	.09	.19	.23	.20	.66
.43	.35	.55	.46	.22	.32	.25	.38	.63	.77	.48	.16	.00	.52	.11	.14	.13	.11	.53
.30	.20	.11	.32	.39	.68	.50	.52	.13	.98	.61	.67	.52	.00	.59	.56	.49	.53	.62
.46	.40	.62	.49	.25	.28	.23	.36	.70	.85	.45	.09	.11	.59	.00	.20	.14	.12	.62
.53	.42	.60	.56	.33	.45	.38	.51	.68	.80	.61	.19	.14	.56	.20	.00	.26	.25	.62
.33	.29	.50	.35	.12	.24	.13	.25	.59	.80	.35	.23	.13	.49	.14	.26	.00	.05	.51
.37	.34	.54	.40	.17	.22	.16	.28	.63	.77	.37	.20	.11	.53	.12	.25	.05	.00	.50
.51	.55	.55	.53	.50	.56	.55	.55	.63	.43	.62	.66	.53	.62	.62	.62	.51	.50	.00

Table 3: Euclidean distances between the 19 countries, computed on the first 3 factor scores.

Compute hierarchy

Computation of hierarchies can be performed according to 4 agglomerative algorithms and 1 divisive algorithm. One partitionning algorithm is also available (it does not produce a hierarchy, but a partition). The 4 agglomerative methods are classical methods: single link, average link (or UPGMA method), complete link, plus the 2nd order moment criterion method, also known as "Ward's method".



For the first three agglomerative algorithms and the divisive algorithm, the input file must be the file containing the distances computed in the previous step.

For the 2nd order moment agglomerative method, the input file is the initial data file.

For the partitionning algorithm, the input file is also the initial data file, but the program also needs an initial partition to start the iterative partitionning algorithm. This partition can be contained in a separate file, or it can be randomly generated by MacDendro. The file containing the initial partition may have several columns, and must have a number of rows equal to the number of elements. One column is selected and it must contain the number of the cluster to which belongs each element. If no file is provided (i.e. if the user selects the "Cancel" button when asked to look for the cluster file), then the initial partition is set at random, and the user must only give the number of clusters for this random partition.

For hierarchical algorithms, the output file contains a description of the hierarchy under the form of a table with 5 columns and a number of rows equal to the number of nodes of the hierarchy (n-1 nodes for n elements). For each node, the 5 columns contain:

1- the node number (starting from n+1)
2- the number of the elder node (or the number of the elder element)
3- the number of the benjamin node (or the number of the benjamin element)
4- the node weight (number of elements attached to this node)
5- the node height.

The name of output files is built up starting from the distance file name, to which an extension string (corresponding to the algorithm) is added:

Algorithm:	**Extension string:**

Single link	.hasl
Average link	.haal
Complete link	.hacl
Agglomerative 2nd order moment	.ha2m
Divisive 2nd order moment	.hd2m

For the partitionning algorithm, there are two output file: one with a ".hctx" extension and one with a ".hcmc" extension. The ".hctx" file is a text file containing the description of the partition, and the ".hcmc" file is a BIN file with only one column, giving, for each element, the number of the cluster to which it belongs.

Exemples of use.

Hierarchy example:

On our example, the average link method produced the hierarchy given in table 4. The first column gives the node number of each node of the hierarchy (numbered starting from n+1). The second and third column give the number of the "elder" and "benjamin" nodes which are attached to the current node, or the number of the element if we are at the bottom of the tree (end nodes). The fourth column gives the weight of each node. This weight is simply the number of elements attached to this node (2 for end nodes). The fifth column contains the height of the node, i.e. the distance between the elements (or the nodes) which were grouped to form the current node.

Node	Eld.	Benj.	Weight	Height
20	1	4	2	.0419
21	17	18	2	.0523
22	12	15	2	.0926
23	3	9	2	.0943
24	8	11	2	.1134
25	21	13	3	.1179
26	23	14	3	.1228
27	5	7	2	.1279
28	22	25	5	.1631
29	20	2	3	.1874
30	28	16	6	.2069
31	27	24	4	.2253
32	31	6	5	.2397
33	29	26	6	.2900
34	30	32	11	.3249
35	10	19	2	.4310
36	33	34	17	.4829
37	36	35	19	.7226

Table 4: Hierarchy obtained with the average link agglomerative method.

Partition example:

Table 5 lists the contents of the file with extension ".hctx" for our example. The initial partition was randomly generated. First, the initial partition is recalled under the form of a series of couples (number of element : number of cluster to which it belongs). Then the successive steps of the iterative partitionning algorithm are printed, with for each step, the corresponding moment and the inertie ratio R = (between clusters moment) / (total moment). This ratio provides a way to judge the quality of the partition (the higher it is, the better is the partition). After the last step, a summary of the process is given (number of steps, total moment, between-clusters moment, and between / total moment ratio of last partition). The final partition is presented in the form of a series of couples (number of element : number of cluster to which it belongs). Then, the final partition between-cluster moment and relative contribution of each cluster are listed.

```
Description of partition:

Cluster number for each element (initial partition):
  1: 1   2: 3   3: 3   4: 1   5: 1   6: 3   7: 3   8: 3   9: 1  10: 2
 11: 2  12: 1  13: 1  14: 1  15: 2  16: 3  17: 2  18: 2  19: 1

Step#:  1   Moment:  .4160   R:  .1170
Step#:  2   Moment:  1.715   R:  .4820
Step#:  3   Moment:  1.821   R:  .5120

Number of iterations       :  3
Total moment               :  3.557
Between-clusters moment    :  1.821
Between/Total moment ratio :  .5120

Final partition :

Cluster number for each element (final partition):
  1: 1   2: 1   3: 1   4: 1   5: 3   6: 3   7: 3   8: 3   9: 1  10: 2
 11: 3  12: 2  13: 2  14: 1  15: 2  16: 2  17: 3  18: 2  19: 2

Cluster#:   Between-clusters moment:   Relative contribution:
1          .8666                    .4758
2          .6293                    .3455
3          .3254                    .1787
```

Table 5: Output file of the iterative partitionning algorithm.

Prepare dendrogram

This command builds the file which will be used by GraphMu to draw the dendrogram of a hierarchy (using the "**Dendrograms**" command), or the convex hulls corresponding to a partition (in this case, the file must first be transposed, and then used as a row selection file for the "**Convex hulls**" command).

The name of output files is built up starting from the distance file name, to which an extension string (corresponding to the algorithm) is added:

Algorithm:	**Extension string:**
Single link	.hssl
Average link	.hsal
Complete link	.hscl
Agglomerative 2nd order moment	.hs2a
Divisive 2nd order moment	.hs2d

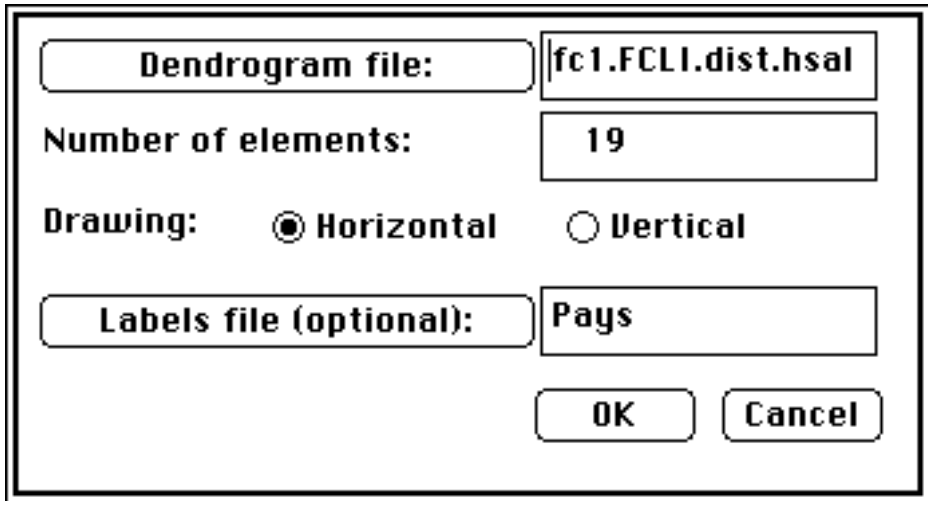
Exemples of use.

Table 6 shows the contents of the output file produced for our example, in the case of the average link method. Each row of this table corresponds to a hierarchical level, the first one corresponding to one cluster containing all elements, and the penultimate corresponding to a number of clusters equal to the number of elements, with one element in each cluster. The last row contains the height of each node (plus the value 1).

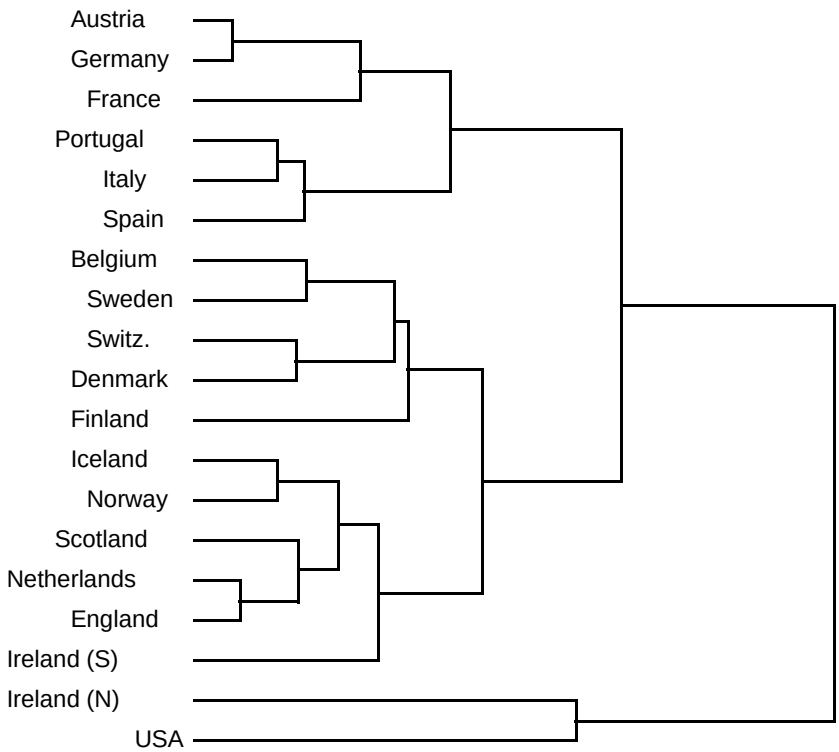
```
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 1  1  1  1  1  1  1  1  1  2  1  1  1  1  1  1  1  2
 1  1  1  1  3  3  3  3  1  2  3  3  3  1  3  3  3  2
 1  1  1  1  3  3  3  3  1  2  3  3  3  1  3  3  3  4
 1  1  1  1  3  3  3  3  1  2  3  5  5  1  5  5  5  4
 1  1  6  1  3  3  3  3  6  2  3  5  5  6  5  5  5  4
 1  1  6  1  3  7  3  3  6  2  3  5  5  6  5  5  5  4
 1  1  6  1  3  7  3  8  6  2  8  5  5  6  5  5  5  4
 1  1  6  1  3  7  3  8  6  2  8  5  5  6  5  9  5  4
 1 10  6  1  3  7  3  8  6  2  8  5  5  6  5  9  5  4
 1 10  6  1  3  7  3  8  6  2  8  5 11  6  5  9 11 11  4
 1 10  6  1  3  7 12  8  6  2  8  5 11  6  5  9 11 11  4
 1 10  6  1  3  7 12  8  6  2  8  5 11 13  5  9 11 11  4
 1 10  6  1  3  7 12  8  6  2  8  5 11 13  5  9 14 14  4
 1 10  6  1  3  7 12  8  6  2 15  5 11 13  5  9 14 14  4
 1 10  6  1  3  7 12  8 16  2 15  5 11 13  5  9 14 14  4
 1 10  6  1  3  7 12  8 16  2 15  5 11 13 17  9 14 14  4
 1 10  6  1  3  7 12  8 16  2 15  5 11 13 17  9 14 18  4
 1 10  6 19  3  7 12  8 16  2 15  5 11 13 17  9 14 18  4

.04 .05 .09 .09 .11 .12 .12 .13 .16 .19 .21 .23 .24 .29 .32 .43 .48 .72 1.0
```

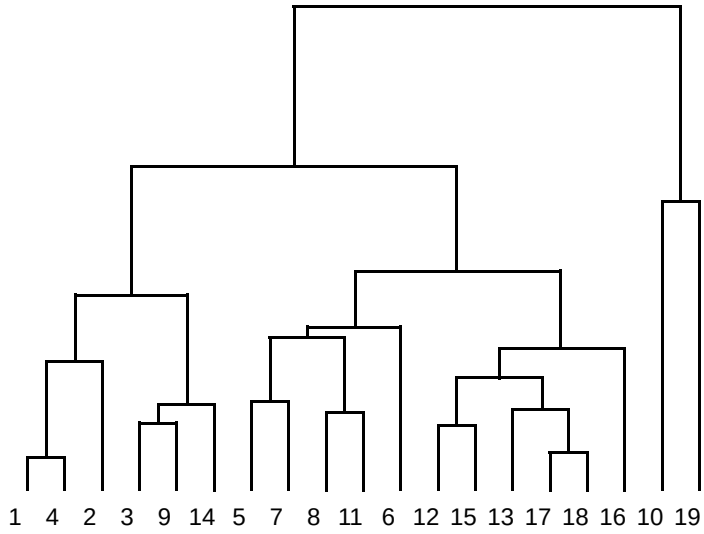
GraphMu must then be used to draw the dendrogram. The "**Dendrograms**" command of GraphMu allows the user to select the dendrogram file (coming from MacDendro), to choose between horizontal or vertical drawing, and (for horizontal dendrograms only) to select an optional label file, containing the names of the elements (the country names here).



The resulting graphic can be saved into a PICT file, which may then be opened with MacDraw, or any graphical software compatible with the PICT format. The dendrogram may then be modified (font and size of text, thickness and aspect of lines, etc.) and printed. The graphic can also be copied to the clipboard, and then pasted into another software (word processor, etc.).

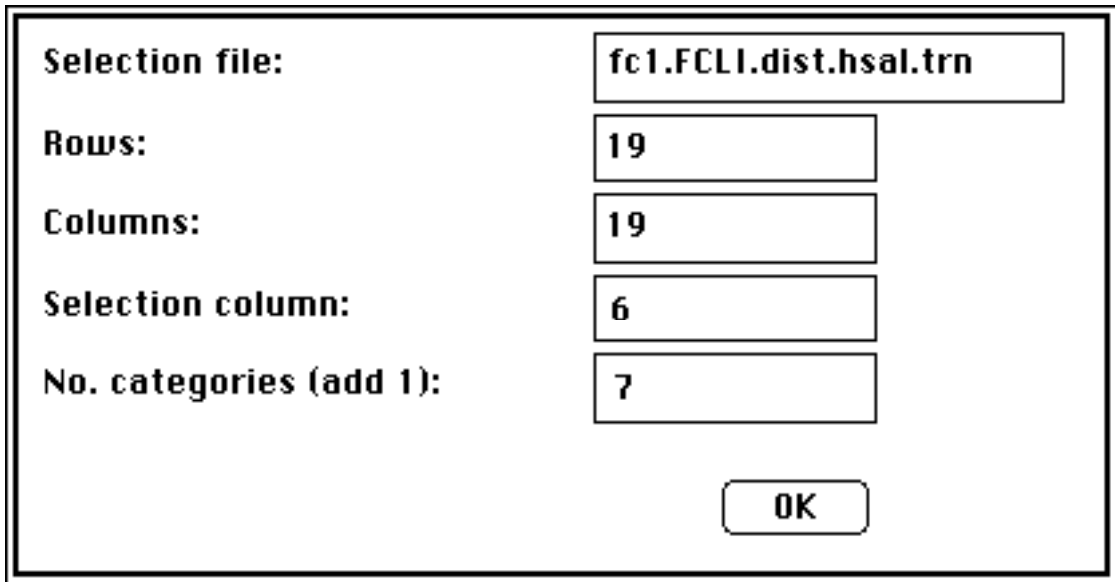


In the case of vertical dendrograms, only the element number is printed at the bottom of the tree:

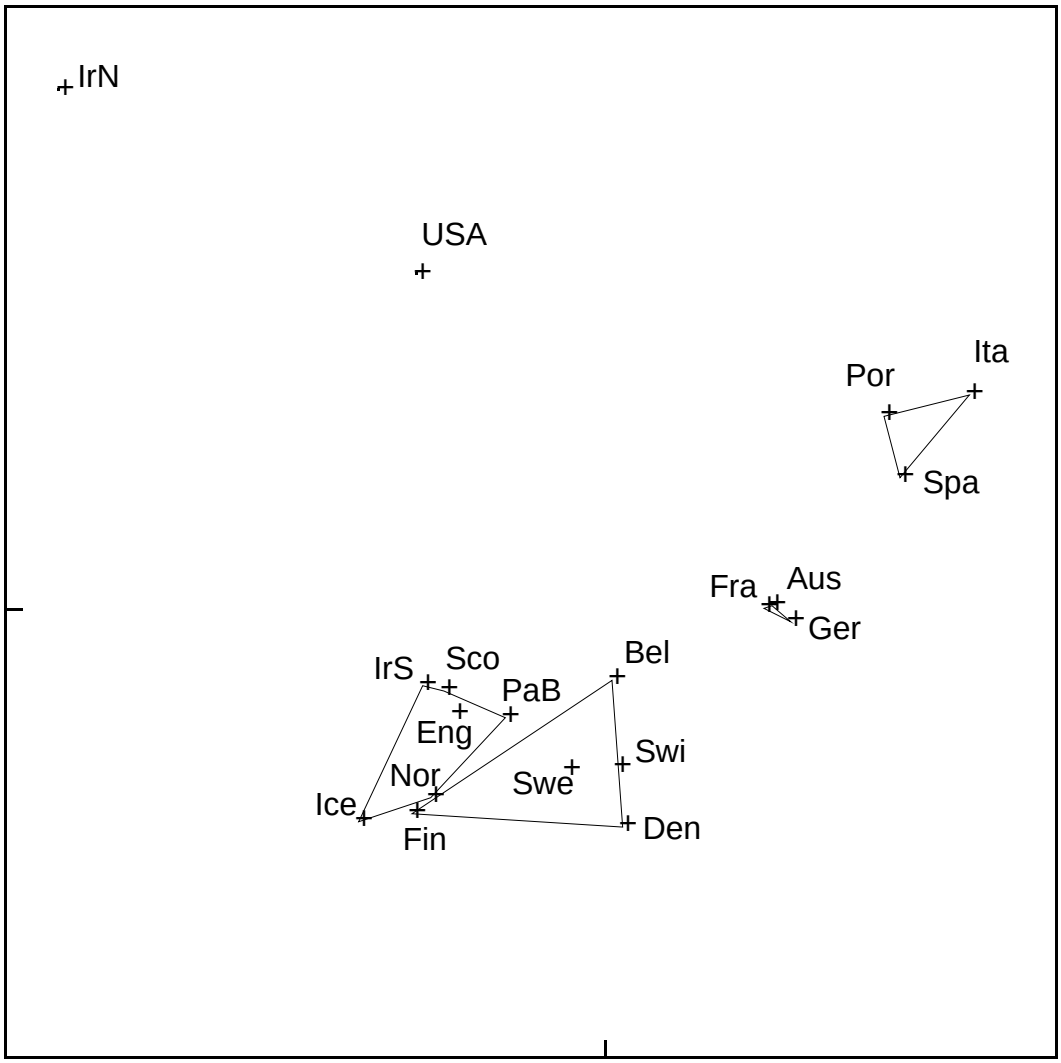


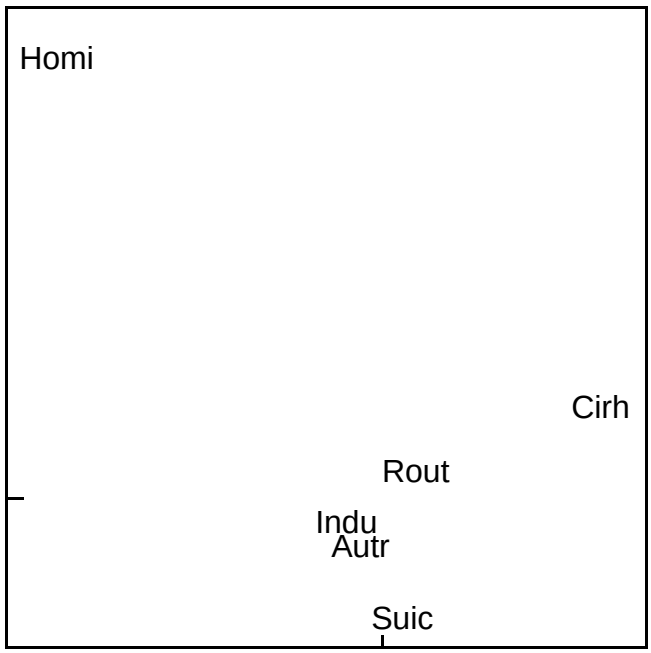
Remark: the commands of the "**Modif.**" menu in GraphMu cannot be used to modify the parameters of the dendrograms drawings.

GraphMu can also be used to draw the convex hulls corresponding to the clusters formed by each level of a hierarchy. These hulls can then be pasted over the factor map of a factor analysis to underline the discrimination between groups of elements. This is achieved with the "Convex hulls" command of GraphMu, by using the dendrogram file produced by MacDendro as a row selection file. The dendrogram file must first be transposed, and can then be used to select the rows corresponding to elements belonging to the same clusters. The user can directly select the hierarchy level (hence the number of clusters) by indicating this level in the "**Selection column**" field (6 clusters in this example):



The resulting graphic can then be pasted over the correspondance analysis factor map:


IrN
USA
Ita
Por
Spa
Fra
Aus
Ger
IrS
Sco
Bel
PaB
Eng
Swi
Nor
Swe
Ice
Den
Fin

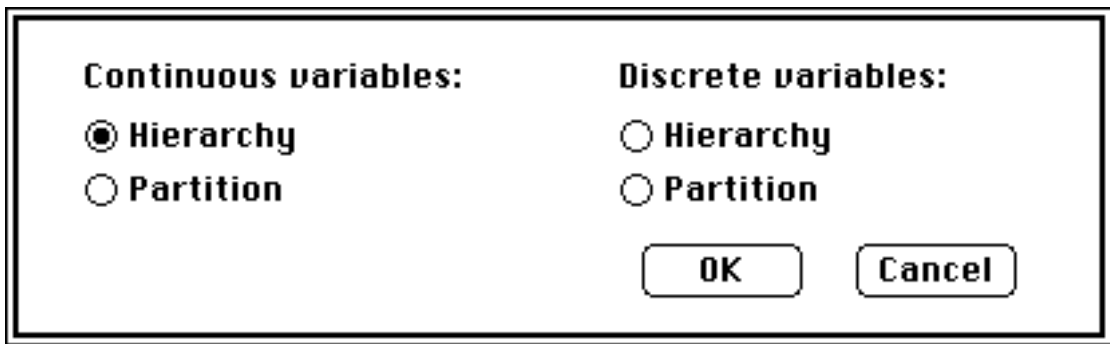

Homi
Cirh
Rout
Indu
Autr
Suic

Inertia analysis

The last step consists in interpretation aids, presented under the form of inertia analyses. Two types of variables in the initial data table are considered: continuous (or quantitative) variables, or discrete (qualitative) variables. Moreover, the inertia analysis may be performed for a hierarchy, or for a partition (a partition may result from any level of a hierarchy, by "cutting the tree" at this level and considering the resulting clusters).

For a hierarchy, the inertia analysis gives the contribution of each variable to the formation of the nodes of the hierarchy.

For partitions, the inertia analysis gives the contribution of the variables to the formation of each cluster, and, reciprocally, the explanation of each cluster by the variables.



Exemples of use.

The following listing shows the results of the inertia analysis of the hierarchy (average link method) computed on the initial data set (19 rows, 6 columns) with the Chi-2 distance (the results are very similar to the results of the cluster analysis performed on the correspondence analysis factor scores with the Euclidean distance, but in this case, we wanted a decomposition of the hierarchy nodes over the 6 death causes).

```
Interpretation of a hierarchy - Continuous variables
Input file : psysoc
Hierarchy file : psysoc.dist.haal

Contribution of variables to the nodes of the hierarchy:

Node#     Suic   Homi   Rout   Indu   Autr   Cirh
-------------------------------------------------------
N 20       5      0     53      3     22     17
```

N 21	5	0	76	1	11	6
N 22	40	7	24	1	1	27
N 23	7	0	19	6	56	12
N 24	2	1	26	0	31	41
N 25	10	0	23	0	47	20
N 26	1	0	43	0	1	55
N 27	36	2	9	3	0	49
N 28	4	0	75	0	7	14
N 29	57	0	10	2	30	0
N 30	1	1	1	0	93	4

N 31	7	0	5	1	70	16
N 32	9	0	6	1	82	2
N 33	46	0	2	0	52	0
N 34	88	0	0	0	0	12
N 35	28	5	0	0	10	57
N 36	0	0	9	0	0	90
N 37	18	63	2	0	2	15

One can easily see, for example, that homicides play a prominent part in the discrimination between USA plus North Ireland and the rest of the world. Indeed, death rate by homicide is much higher in these two countries (see data table).

For a partition, an example of inertia analysis listing is given hereunder. The corresponding partition is:

Cluster 1: Austria, France, Germany, Italy, Portugal, Spain
Cluster 2: Belgium, Sweden, Scotland, The Netherlands, England, Iceland, Norway, Ireland (South), Finland, Switzerland, Denmark
Cluster 3: Ireland (North), USA

This partition may come from the partitionning algorithm, or from the selection of a particular level in a hierarchy, or may be arbitrarily chosen by the user.

Interpretation of clusters - Continuous variables
Input file : psysoc
Clusters file : Cla

Cluster number for each element:
1: 1 2: 1 3: 1 4: 1 5: 2 6: 2 7: 2 8: 2 9: 1 10: 3
11: 2 12: 2 13: 2 14: 1 15: 2 16: 2 17: 2 18: 2 19: 3

Total moment : 5.5603E+05
Between-clusters moment : 2.6183E+05 R : .4709

Contribution of variables to clusters:

Clust.#	Suic	Homi	Rout	Indu	Autr	Cirh
1	0	0	8	0	0	91
2	1	-2	-12	0	0	-85
3	-18	63	2	0	-2	-15

Explanation of clusters by variables:

Clust.#	Suic	Homi	Rout	Indu	Autr	Cirh
1	0	-3	58	0	38	68
2	17	-8	-39	-13	-2	-30
3	-83	89	2	87	-60	-2

First, the partition is recalled under the form of a series of couples (number of element : number of cluster to which it belongs). Then the total moment, the between clusters moment, and the ratio R = (between clusters moment) / (total moment) are printed.

The first table (contribution of variables to clusters) must be read line by line (the sum of absolute values on one line is equal to 100). For each cluster, it gives the contribution of all variables in the formation of this cluster. For our example, the contribution of variables to clusters shows that homicides are characteristic of cluster 3, while cirrhosis explains the oposition between clusters 1 and 2. Industry accidents do not contribute to any cluster.

The second table (explanation of clusters by variables) must be read column by column (the sum of absolute values on one column is equal to 100). For each variable, it gives its explanation by the different clusters. For our example, we see that cluster two is characterized mainly by a high suicide rate and a low road accident rate.

Limits

The maximum number of elements is 100 for the 800Kb versions of MacDendro and GraphMu. For data sets up to 600 elements, a set of "plain vanilla" programs (i.e. with simple line-mode interface) is available. Computation time are negligible for about 20 elements, but grows rapidly with the number of elements. The following results have been obtained on a Macintosh SE/30 for 400 elements:

Distances computation: 20 minutes
Hierarchy computation: 3 minutes
Dendrogram preparation: 35 minutes.

and for 600 elements, on a Macintosh IIfx with 8Mb RAM:

Distances computation: 12 minutes
Hierarchy computation: 3 minutes
Dendrogram preparation: 5 minutes.