Mise en œuvre de la classification (TP SAS)

Classification automatique sur un tableau présence absence

Il s'agit d'un tableau ayant comme individus des souches bactériennes, certaines étant étudiées au labo et inconnues, d'autres étant des souches de référence, issues de collection et parfaitement identifiées.

Les variables sont des critères biochimiques (croissance sur un substrat donné, à une température donnée, production de gaz, ...). Chaque souche peut répondre positivement (1) ou non (0) sur le test biochimique. L'objectif est de voir quelles sont les souches proches (dans le même groupe de la classification) des souches de collection (étiquetées T).

Les souches proches sont étudiées plus finement par une hybridation ADN/ADN pour valider leur appartenance à la même espèce bactérienne.

La classification est donc ici utilisée comme méthode de débroussaillage, l'hybridation étant lourde et coûteuse.

Acquisition et transformation des données

```
title1 "stat/jaccard6";

%include "stat/jaccard+"; /* inclusion du programme permettant */
                          /* de faire des beaux arbres         */

footnote;

data naza1a;
  infile "/nom du répertoire/deer301.prn" firstobs=2;
  input sou $ c4 c5 c6 c10 c12 c13 c18 c21-c28;

data naza1b;
  infile "/nom du répertoire/deer302.prn" firstobs=2;
  input c30-c37 c39 c40 e3 e4 e7 e8 e10 e11;

data naza1;
  merge naza1a naza1b;
run;

title2 "tableau d'origine";
proc print;
run;
```

tableau d'origine

11:45 Monday, October 20, 1997

OBS	OUT	C4	C5	C6	C10	C12	C13	C18	C21	C22	C23	C24	C25	C26	C27	C28	C29	C30	C31	C32	C33	C34	C35	C36	C37	C39	C40	C41	C42	C43	C44	C45
1	B2	0	0	0	1	1	1	0	1	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	0
2	B4	0	0	0	1	1	0	1	1	0	0	0	1	0	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	0	0
3	B5	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	0
4	B7	0	0	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	0
5	B8	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	0
6	B10	0	0	0	1	1	0	0	1	0	0	1	1	0	0	1	1	1	0	0	0	1	1	1	0	1	0	0	1	1	0	0
7	B11	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	0
8	B12	0	0	0	1	1	0	0	1	0	0	1	1	1	0	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	0	0
9	B14	0	0	0	1	1	0	0	1	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	0
10	B15	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	0
11	B16	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	0
12	B17	1	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	1	0	1	1	1	1	1	0	1
13	B18	0	0	0	1	1	0	0	1	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0
14	B21	0	0	0	1	1	0	0	1	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	0	0	0
15	B25	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	0
16	B26	0	1	0	1	1	0	0	1	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	0
17	B27	0	0	1	1	1	0	0	1	0	0	0	1	1	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	0
18	B29	1	1	1	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	1	1	0	0	1	1	1	1	0	0
19	T32	1	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1	1	0	0	0	1	0	0	0	0	1	1	1	1	0	0
20	T33	0	1	0	0	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0
21	T34	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0
22	T35	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	0	0	0	0
23	T39	0	1	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	0	0	0
24	B9	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	0
25	B19	1	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	1	0	1	1	1	1	1	0	0
26	T37	1	1	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	1	1	1	0	0	0
27	T42	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	1	1	1	0	0	0
28	T43	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0
29	T44	1	1	1	1	1	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	0
30	T45	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	1	1	1	1	1	0	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Étape de classification (CAH, centres mobiles, mixte, Jaccard)

La classification automatique sera abordée de plusieurs façons :

- À partir des composantes des individus obtenues après l'AFC sur le tableau des indicatrices,
  - classification type nuées dynamiques,
  - classification hiérarchique (critère de Ward).
  - classification mixte : CAH puis nuées dynamiques.
- Enfin, classification sur les données brutes, non plus sur une matrice de distance euclidienne, mais avec l'indice de similarité de Jaccard comme matrice de distance.

ACM et récupération des coordonnées sur les axes factoriels

```
title2 "AFC sur tableau d'origine";
proc corresp data=naza1 observed outc=coor dim=31;
  tables sou , c4 c5 c6 c10 c12 c13 c18 c21-c28 c30-c37 c39 c40
               e3 e4 e7 e8 e10 e11;
run;
```

```
stat/jaccard6                                  3
                          AFC sur tableau d'origine
                                               11:45 Monday, October 20, 1997
                      The Correspondence Analysis Procedure
stat/jaccard6                                  4
                          AFC sur tableau d'origine
                                               11:45 Monday, October 20, 1997

                              Contingency Table

              0         1         0         1         0         1         .

  B10         1         0         1         0         1         0         .
  B11         1         0         1         0         1         0         .
  B12         1         0         1         0         1         0         .
  B13         1         0         1         0         1         0         .
  B14         1         0         1         0         1         0         .
  B15         1         0         1         0         1         0         .
  B16         1         0         1         0         1         0         .
  B17         0         1         1         0         1         0         .
  B18         1         0         1         0         1         0         .
  B19         0         1         1         0         1         0         .
  B2          1         0         1         0         1         0         .
  B20         0         1         0         1         0         1         .
  B21         1         0         1         0         1         0         .
  B25         0         1         1         0         1         0         .
  B26         1         0         0         1         1         0         .
  B27         1         0         1         0         0         1         .
  B29         0         1         0         1         0         1         .
  B30         1         0         1         0         1         0         .
  B4          1         0         1         0         1         0         .
  B5          1         0         1         0         1         0         .
  .           .         .         .         .         .         .         .
```

AFC sur tableau d'origine

11:45 Monday, October 20, 1997

The Correspondence Analysis Procedure

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi-Squares	Percents	3 6 9 12 15
0.39665	0.15733	351.157	15.73%	**************************
0.34628	0.11991	267.634	11.99%	********************
0.27675	0.07659	170.948	7.66%	************
0.25987	0.06753	150.728	6.75%	**********
0.24759	0.06130	136.818	6.13%	**********
0.24245	0.05878	131.196	5.88%	**********
0.22407	0.05021	112.058	5.02%	********
0.20666	0.04271	95.327	4.27%	*******
0.19962	0.03985	88.943	3.98%	*******
0.18462	0.03408	76.073	3.41%	******
0.17979	0.03232	72.146	3.23%	*****
0.17043	0.02904	64.828	2.90%	*****
0.16354	0.02674	59.694	2.67%	****
0.15840	0.02509	56.004	2.51%	****
0.15131	0.02289	51.099	2.29%	****
0.14122	0.01994	44.515	1.99%	***
0.13692	0.01875	41.842	1.87%	***
0.12881	0.01659	37.031	1.66%	***
0.12192	0.01486	33.177	1.49%	**
0.11937	0.01425	31.806	1.42%	**
0.11591	0.01344	29.987	1.34%	**
0.11122	0.01237	27.612	1.24%	**
0.10446	0.01091	24.353	1.09%	**
0.09356	0.00875	19.537	0.88%	*
0.07736	0.00598	13.358	0.60%	*
0.07397	0.00547	12.211	0.55%	*
0.06919	0.00479	10.684	0.48%	*
0.05863	0.00344	7.671	0.34%	*
0.05568	0.00310	6.920	0.31%	*
0.04625	0.00214	4.774	0.21%	
0.02891	0.00084	1.866	0.08%	
	1.00000	2232	(Degrees of Freedom = 4331)	

```
title2 "Table issue de AFC sur tableau d'origine";
proc print data=coor(obs=10 drop=dim3-dim31
                                contr3-contr31
                                sqcos3-sqcos31
                                best3-best31);
run;
```


Table issue de AFC sur tableau d'origine

11:45 Monday, October 20, 1997

OBS	_TYPE_	_NAME_	QUALITY	MASS	INERTIA	DIM1	DIM2
1	INERTIA		.	.	1.00000	.	.
2	OBS	B10	1	0.013889	0.01360	-0.31919	0.03527
3	OBS	B11	1	0.013889	0.00332	-0.32436	0.07125
4	OBS	B12	1	0.013889	0.00940	-0.14547	-0.37360
5	OBS	B13	1	0.013889	0.00639	-0.33562	0.17287
6	OBS	B14	1	0.013889	0.00357	-0.34549	-0.05197
7	OBS	B15	1	0.013889	0.00463	-0.39893	0.06801
8	OBS	B16	1	0.013889	0.00463	-0.39893	0.06801
9	OBS	B17	1	0.013889	0.02436	-0.31645	0.13499
10	OBS	B18	1	0.013889	0.01960	-0.46900	-0.31600

OBS	CONTR1	CONTR2	SQCOS1	SQCOS2	BEST1	BEST2	BEST
1	0.15733	0.11991	.	.	.	.	.
2	0.00899	0.00014	0.10402	0.00127	0	0	21
3	0.00929	0.00059	0.44073	0.02127	0	0	24
4	0.00187	0.01617	0.03126	0.20620	0	0	28
5	0.00994	0.00346	0.24492	0.06498	0	0	8
6	0.01054	0.00031	0.46447	0.01051	1	0	1
7	0.01405	0.00054	0.47776	0.01389	17	0	17
8	0.01405	0.00054	0.47776	0.01389	17	0	17
9	0.00884	0.00211	0.05709	0.01039	0	0	18
10	0.01942	0.01157	0.15589	0.07077	10	0	10

Classification type nuées dynamiques sur l'ensemble des axes de l'AFC

```
title2 "nuées dynamiques issues des 31 axes factoriels de l'AFC";
run;
proc fastclus data=coor out=fast maxc=10 maxiter=7;
  where _type_ ='OBS';
  var dim1-dim31;
*  id sou;
  id _name_;
run;
```

```
stat/jaccard6                                  108
          nuées dynamiques issues des 31 axes factoriels de l'AFC
                                          11:45 Monday, October 20, 1997

   FASTCLUS Procedure:  Replace=FULL  Radius=0  Maxclusters=10 Maxiter=7
   Converge=0.02

                              Initial Seeds

Cluster    DIM1      DIM2      DIM3      DIM4      DIM5      DIM6      DIM7
--------------------------------------------------------------------------------
  1    0.337804  -.863009  0.097716  0.309637  -.092377  -.156062 0.029758
  2    -.319708  0.250199  0.104562  -.099779  -.354578  0.346817 -.311919
  3    0.624581  0.163032  -.289353  -.202054  -.051459  0.464818 0.125827
  4    0.526985  0.592662  0.159685  -.498219  0.165119  -.362227 -.276851
  5    -.106312  -.153132  0.387630  0.058134  -.283583  -.483853 0.983428
  6    -.518847  0.085990  0.003241  0.291296  0.513071  0.318467 -.057283
  7    0.977188  0.143284  0.333528  -.279175  -.503057  0.563297 -.190134
  8    0.029478  -.540695  -.421014  -.136465  0.644237  -.297051 -.046736
  9    0.489193  0.402006  0.216780  0.677414  0.222006  -.356101 -.106308
 10    0.467048  -.614003  0.151712  -.524233  0.569173  0.350179 0.308028
```

Minimum Distance Between Initial Seeds = 1.508092

Iteration	Criterion	Relative Change in Cluster Seeds 1	2	3	4	5
		6	7	8	9	10
1	0.1974	0.4848	0.6049	0.5312	0.4887	0.4572
		0.4821	0.4356	0.3281	0.5437	0
2	0.1347	0.1018	0.0486	0.1551	0	0.2838
		0.1247	0	0	0	0
3	0.1322	0	0.0122	0	0	0
		0.0456	0	0	0	0
4	0.1320	0	0.0255	0	0	0
		0.1328	0	0	0	0
5	0.1311	0	0.0274	0.0626	0	0
		0.1268	0	0	0	0
6	0.1305	0	0	0	0	0
		0	0	0	0	0

Convergence criterion is satisfied.

Criterion Based on Final Seeds = 0.13048

stat/jaccard6 110
nuees dynamiques issues des 31 axes factoriels de l'AFC
11:45 Monday, October 20, 1997

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=10 Maxiter=7
Converge=0.02

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	8	0.1664	1.0767	2	0.9979
2	35	0.1212	1.0240	3	0.8161
3	8	0.1524	0.8945	2	0.8161
4	5	0.1619	0.9208	9	0.9978
5	2	0.1712	0.6742	2	1.2405
6	6	0.1628	1.0732	2	0.8197
7	2	0.1669	0.6569	4	1.2289
8	2	0.1257	0.4948	2	1.1689
9	3	0.1725	0.8199	4	0.9978
10	1	.	0	1	1.4407

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
DIM1	0.399430	0.167213	0.846964	5.534406
DIM2	0.348707	0.195521	0.725464	2.642513
DIM3	0.278691	0.222673	0.442528	0.793811
DIM4	0.261690	0.184018	0.568204	1.315907
DIM5	0.249323	0.177657	0.556623	1.255417
DIM6	0.244147	0.164908	0.601605	1.510071
DIM7	0.225637	0.171138	0.497655	0.990663
DIM8	0.208113	0.152104	0.533535	1.143785
DIM9	0.201023	0.187609	0.239408	0.314765
DIM10	0.185912	0.154743	0.395021	0.652950
DIM11	0.181049	0.172987	0.202793	0.254379
DIM12	0.171622	0.165192	0.190967	0.236043
DIM13	0.164686	0.155145	0.225007	0.290335
DIM14	0.159514	0.159323	0.128854	0.147913
DIM15	0.152369	0.153053	0.118905	0.134951
DIM16	0.142215	0.146562	0.072562	0.078239
DIM17	0.137878	0.132513	0.193397	0.239768
DIM18	0.129710	0.124005	0.201895	0.252968
DIM19	0.122774	0.122490	0.130802	0.150486
DIM20	0.120211	0.122458	0.093816	0.103528
DIM21	0.116723	0.122401	0.039740	0.041385
DIM22	0.112005	0.115694	0.068297	0.073303
DIM23	0.105189	0.107718	0.084268	0.092022
DIM24	0.0942156	0.0991519	0.032860	0.033976
DIM25	0.0779056	0.0819636	0.033419	0.034574
DIM26	0.0744859	0.0781559	0.038589	0.040137
DIM27	0.0696720	0.0728353	0.045664	0.047849
DIM28	0.0590368	0.0618250	0.042331	0.044202
DIM29	0.0560713	0.0583931	0.052942	0.055902
DIM30	0.0465732	0.0478222	0.079296	0.086126
DIM31	0.0291172	0.0307008	0.029191	0.030069
OVER-ALL	0.180866	0.140605	0.472259	0.894870

nuees dynamiques issues des 31 axes factoriels de l'AFC

11:45 Monday, October 20, 1997

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=10 Maxiter=7 Converge=0.02

Statistics for Variables

Pseudo F Statistic = 6.16

Approximate Expected Over-All R-Squared = 0.46655

Cubic Clustering Criterion = 0.455

WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	DIM1	DIM2	DIM3	DIM4	DIM5	DIM6	DIM7
1	0.477809	-.585246	0.109903	0.189902	-.053914	-.036889	-.054324
2	-.267119	-.013165	0.028424	-.056893	-.081255	-.010711	-.054383
3	0.231743	0.051420	-.477447	-.001335	0.032041	0.234669	0.181257
4	0.561251	0.565100	0.178846	-.247038	0.137272	-.174213	-.137852
5	-.211382	-.009071	0.031916	-.094835	-.238779	-.561331	0.707652
6	-.325346	0.262394	0.194529	0.372981	0.261226	0.248331	0.092542
7	0.862391	0.209921	0.066677	-.338026	-.490274	0.319847	-.253780
8	-.219759	-.428349	-.183979	-.231613	0.590340	-.232107	-.133327
9	0.496345	0.466836	-.032280	0.444396	0.157813	-.409753	0.024276
10	0.467048	-.614003	0.151712	-.524233	0.569173	0.350179	0.308028

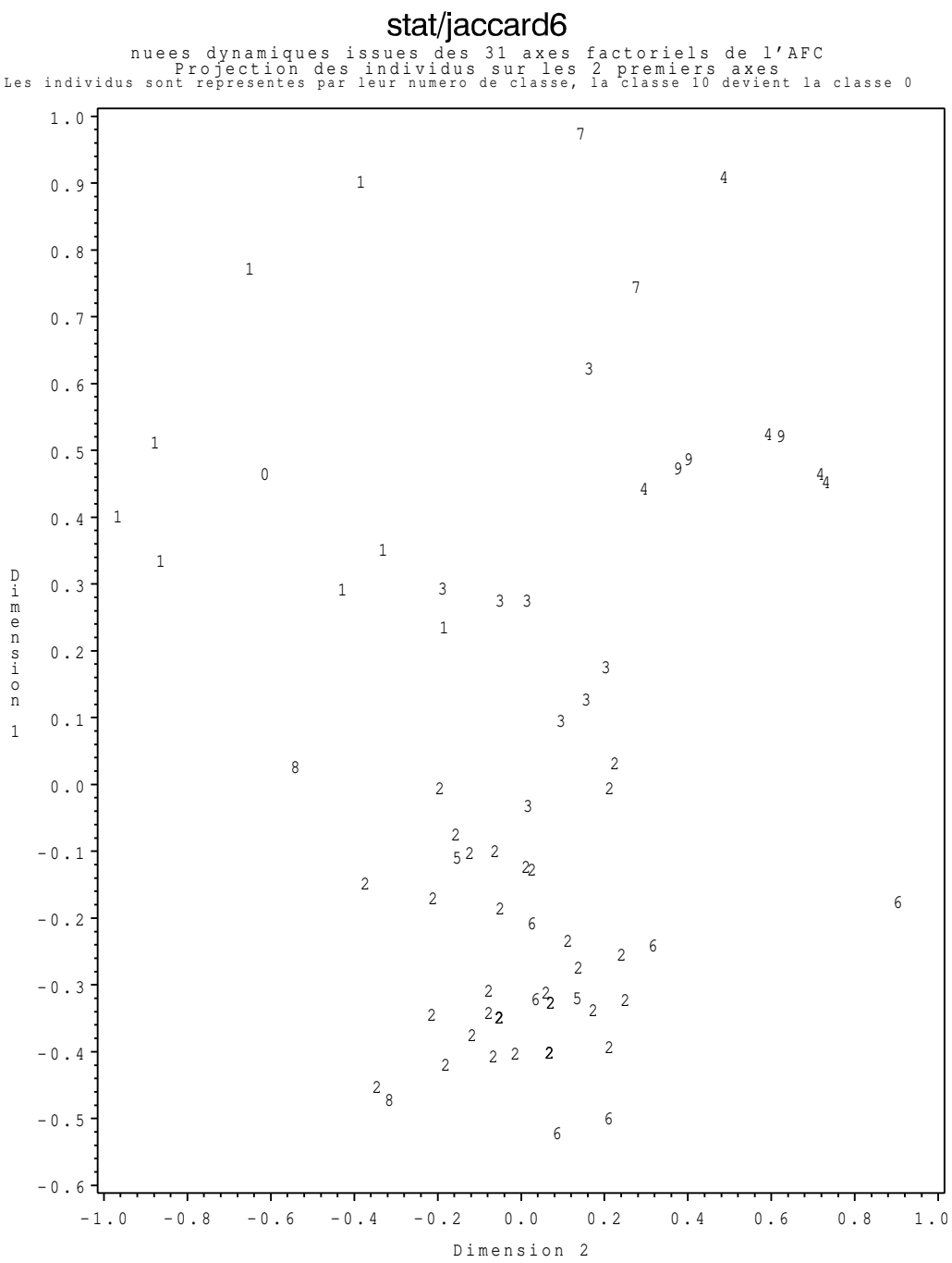
Cluster Standard Deviations

Cluster	DIM1	DIM2	DIM3	DIM4	DIM5	DIM6	DIM7
1	0.240207	0.293729	0.178111	0.307051	0.180260	0.117650	0.151047
2	0.135415	0.158182	0.202739	0.124794	0.139174	0.132442	0.160521
3	0.193530	0.130137	0.170800	0.241935	0.254970	0.279967	0.153972
4	0.198506	0.180857	0.164886	0.181085	0.176988	0.241738	0.114429
5	0.148592	0.203734	0.503055	0.216331	0.063362	0.109570	0.390006
6	0.149953	0.332798	0.288702	0.190883	0.303727	0.070813	0.212731
7	0.162348	0.094238	0.377384	0.083228	0.018078	0.344290	0.090010
8	0.352474	0.158881	0.335218	0.134559	0.076222	0.091845	0.122457
9	0.025277	0.135465	0.329418	0.239896	0.097604	0.152169	0.283084
10	.	.	.	.	.	.	.

Projection des individus

```
/* cette etape permet de projeter les individus sur les axes */
data fast2;
  set fast;
clus=put(cluster,2.);   * text ne peut prendre que des chaines de;
text = substr(clus,2);  * caracteres. On traduit donc cluster en clus;

*text=_name_;         * Cette ligne aurait fait imprimer le numero des ;
                      * individus au lieu de celui des classes.;
y = dim1;
x = dim2;
xsys = '2';
ysys = '2';
size = 1;
label y = "Dimension 1";
label x = "Dimension 2";
keep x y text xsys ysys size;
run;
*proc print;
*run;
filename figure "plot1.eps";
title3 "Projection des individus sur les 2 premiers axes";
title4 "Les individus sont representes par leur numero de classe, la "
       "classe 10 devient la classe 0";
proc gplot data=fast2;
  symbol1 i=none;
  plot y*x=1 / annotate=fast2 frame;
run;
quit;
```


stat/jaccard6
nuees dynamiques issues des 31 axes factoriels de l'AFC
Projection des individus sur les 2 premiers axes
Les individus sont representes par leur numero de classe, la classe 10 devient la classe 0
Dimension 1
Dimension 2
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0

CAH sur les composantes principales

```
title2 "CAH (critere de Ward) issue des 31 axes factoriels de l'AFC";
proc cluster data=coor
             outtree=treeward
             method=ward;
  where _type_ ='OBS';
  var dim1-dim31;
*  id sou;
  id _name_;
run;
*proc print data=treeward;
*title3 "treeward";
*run;
```

CAH (critere de Ward) issue des 31 axes factoriels de l'AFC

11:45 Monday, October 20, 1997

Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	0.159544	0.037948	0.157328	0.15733
2	0.121596	0.043928	0.119908	0.27724
3	0.077668	0.009187	0.076590	0.35383
4	0.068482	0.006320	0.067530	0.42136
5	0.062162	0.002554	0.061298	0.48265
6	0.059608	0.008695	0.058780	0.54143
7	0.050912	0.007601	0.050205	0.59164
8	0.043311	0.002901	0.042709	0.63435
9	0.040410	0.005847	0.039849	0.67420
10	0.034563	0.001784	0.034083	0.70828
11	0.032779	0.003325	0.032323	0.74060
12	0.029454	0.002333	0.029045	0.76965
13	0.027121	0.001677	0.026745	0.79639
14	0.025445	0.002228	0.025091	0.82149
15	0.023216	0.002991	0.022894	0.84438
16	0.020225	0.001215	0.019944	0.86432
17	0.019010	0.002186	0.018746	0.88307
18	0.016825	0.001751	0.016591	0.89966
19	0.015073	0.000623	0.014864	0.91452
20	0.014451	0.000826	0.014250	0.92877
21	0.013624	0.001079	0.013435	0.94221
22	0.012545	0.001480	0.012371	0.95458
23	0.011065	0.002188	0.010911	0.96549
24	0.008877	0.002807	0.008753	0.97424
25	0.006069	0.000521	0.005985	0.98023
26	0.005548	0.000694	0.005471	0.98570
27	0.004854	0.001369	0.004787	0.99049
28	0.003485	0.000341	0.003437	0.99392
29	0.003144	0.000975	0.003100	0.99703
30	0.002169	0.001321	0.002139	0.99916
31	0.000848	.	0.000836	1.00000

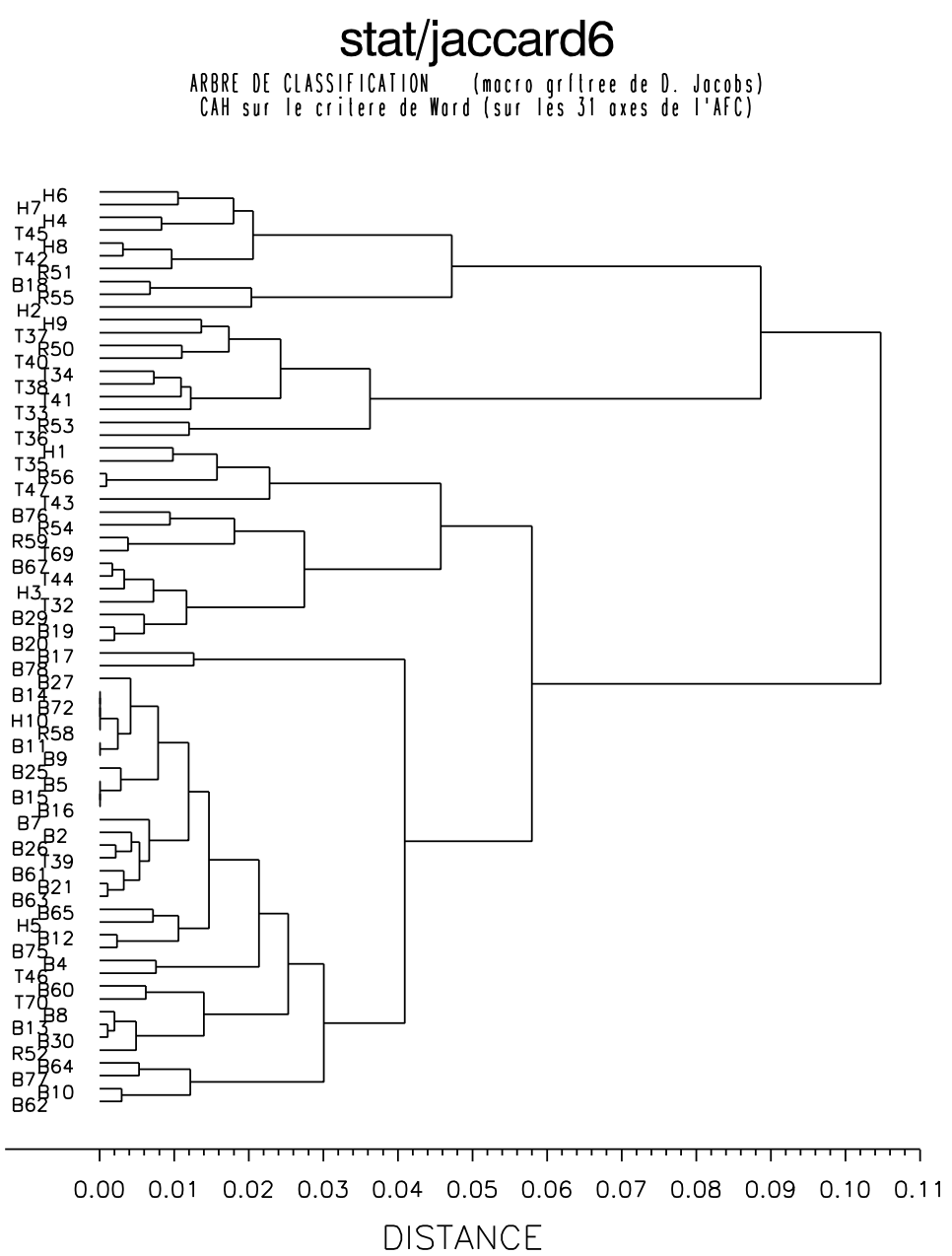
Root-Mean-Square Total-Sample Standard Deviation = 0.180866

Root-Mean-Square Distance Between Observations = 1.424138

Ward's Minimum Variance Cluster Analysis

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
71	B15	B16	2	0.000000	1.000000	T
70	CL71	B5	3	0.000000	1.000000	T
69	B14	B72	2	0.000000	1.000000	T
68	B11	B9	2	0.000000	1.000000	T
67	CL69	H10	3	0.000000	1.000000	T
66	CL67	R58	4	0.000000	1.000000	
65	R56	T47	2	0.000914	0.999086	
64	B13	B30	2	0.001084	0.998002	
63	B21	B63	2	0.001117	0.996885	
62	B67	T44	2	0.001731	0.995155	
61	B19	B20	2	0.002006	0.993149	
60	CL64	B8	3	0.002017	0.991131	
59	B26	T39	2	0.002156	0.988976	
58	B12	B75	2	0.002359	0.986617	
57	CL68	CL66	6	0.002436	0.984180	
56	CL70	B25	4	0.002827	0.981353	
55	B10	B62	2	0.002962	0.978392	
54	H8	T42	2	0.003115	0.975277	
53	CL63	B61	3	0.003246	0.972031	
52	CL62	H3	3	0.003308	0.968723	
51	R59	T69	2	0.003846	0.964877	
50	CL57	B27	7	0.004162	0.960715	
49	B2	CL59	3	0.004255	0.956460	
48	CL60	R52	4	0.004908	0.951551	
47	B64	B77	2	0.005320	0.946231	
46	CL49	CL53	6	0.005365	0.940866	
45	CL61	B29	3	0.005976	0.934890	
44	B60	T70	2	0.006210	0.928680	
43	CL46	B7	7	0.006677	0.922003	
42	B18	R55	2	0.006800	0.915203	
41	B65	H5	2	0.007192	0.908010	
40	CL52	T32	4	0.007238	0.900772	
..	....	....	.	....	....	

```
title2 "ARBRE DE CLASSIFICATION    (macro grftree de D. Jacobs)";
title3 "CAH sur le critere de Ward (sur les 31 axes de l'AFC)";
filename figure "cah.eps";
%grftree(clusdsn=treeward,items=72,axis=d);*!!!!!!!!!!!!!!!!!!!;
run;
```



```
proc tree data=treeward n=10 noprint out=trees;
  id _name_;
  copy dim1--dim31;
run;

title2 "sortie de la procedure tree (10 1eres obs, 5 1ers axes)";
proc print data=trees(obs=10);
  var _name_ dim1-dim5 cluster clusname;
* on n imprime pas les axes 6 a 31;
```

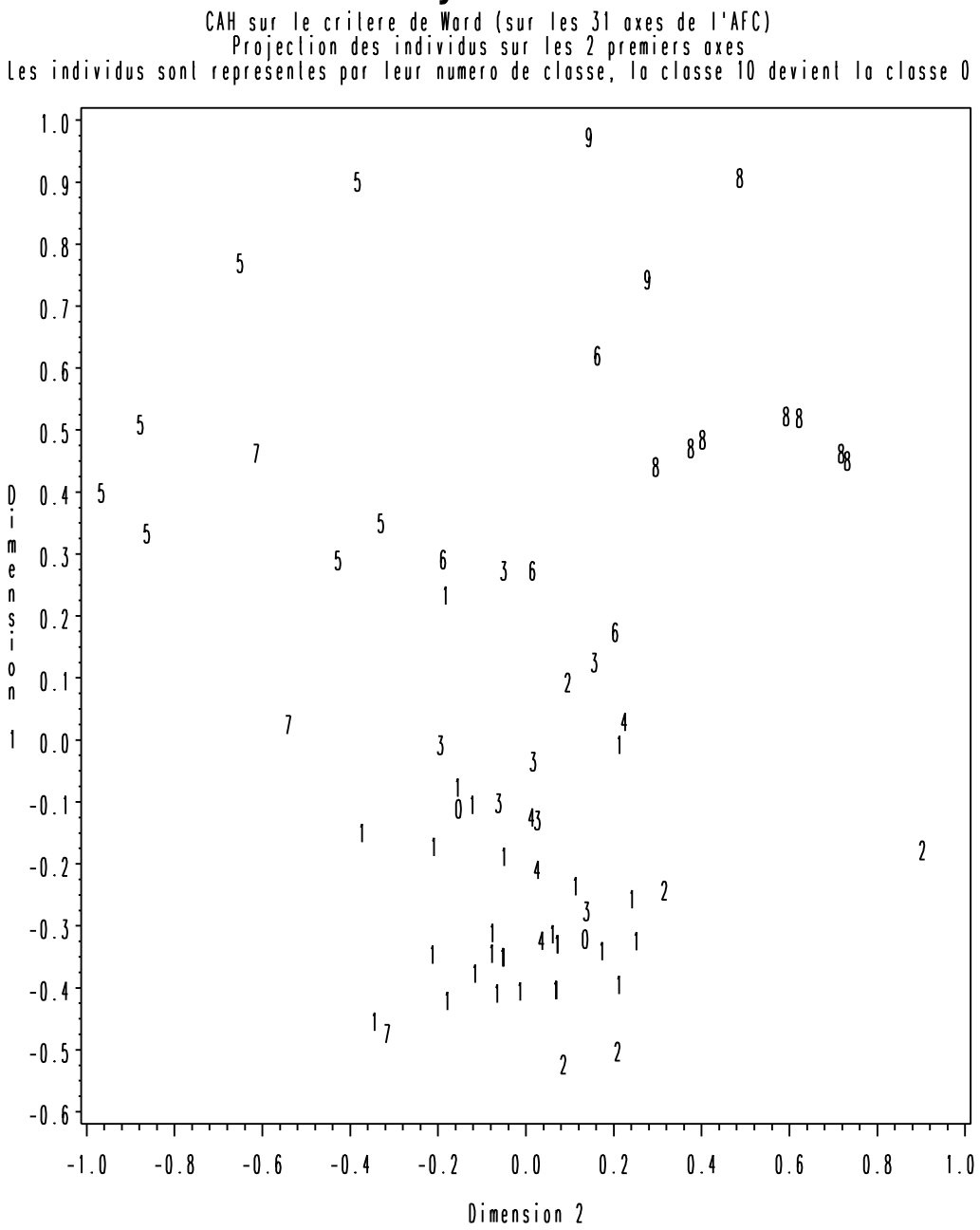
```
stat/jaccard6                                  117
          sortie de la procedure tree (10 1eres obs, 5 1ers axes)
                                               11:45 Monday, October 20, 1997
```

OBS	_NAME_	DIM1	DIM2	DIM3	DIM4	DIM5	CLUSTER	CLUSNAME
1	B15	-0.39893	0.06801	-0.044307	-0.18210	-0.03402	1	CL10
2	B16	-0.39893	0.06801	-0.044307	-0.18210	-0.03402	1	CL10
3	B5	-0.39893	0.06801	-0.044307	-0.18210	-0.03402	1	CL10
4	B14	-0.34549	-0.05197	0.091357	-0.15301	-0.10554	1	CL10
5	B72	-0.34549	-0.05197	0.091357	-0.15301	-0.10554	1	CL10
6	B11	-0.32436	0.07125	0.028772	-0.11354	-0.15597	1	CL10
7	B9	-0.32436	0.07125	0.028772	-0.11354	-0.15597	1	CL10
8	H10	-0.34549	-0.05197	0.091357	-0.15301	-0.10554	1	CL10
9	R58	-0.34549	-0.05197	0.091357	-0.15301	-0.10554	1	CL10
10	R56	-0.49772	0.20921	-0.059343	0.33077	0.46264	2	CL12

```
*proc print data=trees;

data trees2;
  set trees;
clus=put(cluster,2.);   * text ne peut prendre que des chaines de;
text = substr(clus,2);  * caracteres. On traduit donc cluster en clus.;
*text=_name_;
y = dim1;
x = dim2;
xsys = '2';
ysys = '2';
size = 1;
label y = "Dimension 1";
label x = "Dimension 2";
keep x y text xsys ysys size;
run;
*proc print;
*run;
filename figure "plot2.eps";
title2 "CAH sur le critere de Ward (sur les 31 axes de l'AFC)";
title3 "Projection des individus sur les 2 premiers axes";
title4 "Les individus sont representes par leur numero de classe, la "
       "classe 10 devient la classe 0";
proc gplot data=trees2;
  symbol1 i=none;
  plot y*x=1 / annotate=trees2 frame;
run;
quit;
```

stat/jaccard6



Classification mixte
(Ward, puis nuées dynamiques)

```
proc summary nway data=trees;
  var dim1-dim31;
  class cluster;
  output out=summa mean= ;
run;

title2 "CLASSIFICATION MIXTE : CAH-Ward puis nuees dynamiques";
title3 "1) Centre de gravite des classes d'origine issues de la CAH";
proc print data=summa;
  var cluster dim1-dim5;
run;
```

```
stat/jaccard6                                  118
            CLASSIFICATION MIXTE : CAH-Ward puis nuees dynamiques
         1) Centre de gravite des classes d'origine issues de la CAH
                                          11:45 Monday, October 20, 1997
```

OBS	CLUSTER	DIM1	DIM2	DIM3	DIM4	DIM5
1	1	-0.28426	-0.02630	0.08113	-0.08707	-0.07106
2	2	-0.26604	0.32174	-0.09106	0.31966	0.37942
3	3	-0.01700	0.00420	-0.45366	0.06092	-0.11749
4	4	-0.15252	0.07467	0.36376	0.20679	-0.12437
5	5	0.51210	-0.64255	0.07405	0.25917	-0.04448
6	6	0.34442	0.04826	-0.38867	-0.07470	0.10357
7	7	0.00918	-0.49023	-0.07208	-0.32915	0.58328
8	8	0.53691	0.52825	0.09967	0.01225	0.14497
9	9	0.86239	0.20992	0.06668	-0.33803	-0.49027
10	10	-0.21138	-0.00907	0.03192	-0.09483	-0.23878

```
title2 "CLASSIFICATION MIXTE : CAH-Ward puis nuees dynamiques";
title3 "2) Nuees dyn. avec les centres de gravite des classes d'origine"
       "comme noyaux initiaux";
proc fastclus seed=summa data=trees mean=moy2 out=mixt maxc=10 maxiter=7;
id _name_;
var dim1-dim31;
run;
```


CLASSIFICATION MIXTE : CAH-Ward puis nuees dynamiques

2) Nuees dyn. avec les centres de gravite des classes d'origine comme noyaux i

11:45 Monday, October 20, 1997

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=10 Maxiter=7 Converge=0.02

Initial Seeds

Cluster	DIM1	DIM2	DIM3	DIM4	DIM5	DIM6	DIM7
1	-.284264	-.026300	0.081128	-.087072	-.071056	0.018886	-.075814
2	-.266041	0.321743	-.091063	0.319657	0.379421	0.322640	0.007877
3	-.017004	0.004200	-.453656	0.060923	-.117495	-.128833	-.000910
4	-.152519	0.074673	0.363764	0.206794	-.124366	0.049513	0.309153
5	0.512101	-.642554	0.074052	0.259174	-.044475	-.056695	-.071105
6	0.344420	0.048263	-.388671	-.074698	0.103573	0.404548	0.292461
7	0.009177	-.490233	-.072082	-.329153	0.583284	-.038012	0.013792
8	0.536911	0.528251	0.099674	0.012250	0.144975	-.262541	-.077054
9	0.862391	0.209921	0.066677	-.338026	-.490274	0.319847	-.253780
10	-.211382	-.009071	0.031916	-.094835	-.238779	-.561331	0.707652

Minimum Distance Between Initial Seeds = 0.684557

Iteration	Criterion	Relative Change in Cluster Seeds 1	2	3	4	5
		6	7	8	9	10
1	0.1295	0	0.3666	0	0	0
		0.2816	0	0	0	0
2	0.1287	0	0	0	0	0
		0	0	0	0	0

Convergence criterion is satisfied.

Criterion Based on Final Seeds = 0.12874

stat/jaccard6 121

CLASSIFICATION MIXTE : CAH-Ward puis nuees dynamiques

2) Nuees dyn. avec les centres de gravite des classes d'origine comme noyaux i

11:45 Monday, October 20, 1997

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=10 Maxiter=7 Converge=0.02

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	30	0.1165	1.0085	3	0.6846
2	4	0.1569	1.0237	1	0.9557
3	7	0.1112	0.7351	1	0.6846
4	4	0.1259	0.6495	1	0.7830
5	7	0.1648	1.0481	3	1.0511
6	5	0.1555	0.8804	3	0.8431
7	3	0.1775	0.9880	1	1.1279
8	8	0.1791	1.0645	3	1.0262
9	2	0.1669	0.6569	8	1.2774
10	2	0.1712	0.6742	1	1.2647

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
DIM1	0.399430	0.185968	0.810709	4.282874
DIM2	0.348707	0.182697	0.760296	3.171807
DIM3	0.278691	0.184186	0.618583	1.621799
DIM4	0.261690	0.212406	0.424702	0.738229
DIM5	0.249323	0.158127	0.648744	1.846924
DIM6	0.244147	0.154506	0.650279	1.859426
DIM7	0.225637	0.154370	0.591267	1.446586
DIM8	0.208113	0.155585	0.511940	1.048928
DIM9	0.201023	0.182558	0.279816	0.388535
DIM10	0.185912	0.148528	0.442639	0.794170
DIM11	0.181049	0.173358	0.199371	0.249018
DIM12	0.171622	0.162791	0.214313	0.272771
DIM13	0.164686	0.167767	0.093778	0.103482
DIM14	0.159514	0.161893	0.100520	0.111754
DIM15	0.152369	0.153998	0.107991	0.121065
DIM16	0.142215	0.142531	0.122867	0.140078
DIM17	0.137878	0.137006	0.137779	0.159795
DIM18	0.129710	0.127201	0.160216	0.190783
DIM19	0.122774	0.126123	0.078466	0.085147
DIM20	0.120211	0.119386	0.138701	0.161036
DIM21	0.116723	0.120450	0.070109	0.075394
DIM22	0.112005	0.118682	0.019546	0.019936
DIM23	0.105189	0.107198	0.093078	0.102631
DIM24	0.0942156	0.0963707	0.086355	0.094517
DIM25	0.0779056	0.0797048	0.085962	0.094046
DIM26	0.0744859	0.0758317	0.094919	0.104874
DIM27	0.0696720	0.0732036	0.035987	0.037331
DIM28	0.0590368	0.0618636	0.041135	0.042900
DIM29	0.0560713	0.0584869	0.049898	0.052518
DIM30	0.0465732	0.0477307	0.082817	0.090295
DIM31	0.0291172	0.0306241	0.034037	0.035236
OVER-ALL	0.180866	0.138733	0.486215	0.946340

Pseudo F Statistic = 6.52
Approximate Expected Over-All R-Squared = 0.46655
Cubic Clustering Criterion = 1.587
WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	DIM1	DIM2	DIM3	DIM4	DIM5	DIM6	DIM7
1	-.284264	-.026300	0.081128	-.087072	-.071056	0.018886	-.075814
2	-.356999	0.378226	0.077243	0.382998	0.406920	0.282554	-.034900
3	-.017004	0.004200	-.453656	0.060923	-.117495	-.128833	-.000910
4	-.152519	0.074673	0.363764	0.206794	-.124366	0.049513	0.309153
5	0.512101	-.642554	0.074052	0.259174	-.044475	-.056695	-.071105
6	0.295095	0.057774	-.463794	-.046500	0.136744	0.420235	0.269766
7	0.009177	-.490233	-.072082	-.329153	0.583284	-.038012	0.013792
8	0.536911	0.528251	0.099674	0.012250	0.144975	-.262541	-.077054
9	0.862391	0.209921	0.066677	-.338026	-.490274	0.319847	-.253780
10	-.211382	-.009071	0.031916	-.094835	-.238779	-.561331	0.707652

Cluster Standard Deviations

Cluster	DIM1	DIM2	DIM3	DIM4	DIM5	DIM6	DIM7
1	0.148395	0.161667	0.122344	0.109183	0.143916	0.113428	0.149325
2	0.176892	0.361813	0.289531	0.232684	0.255831	0.056605	0.096724
3	0.178904	0.121365	0.137892	0.176088	0.099856	0.161158	0.063935
4	0.148838	0.100374	0.077885	0.194960	0.123571	0.158956	0.080947
5	0.237361	0.264582	0.158159	0.255356	0.192556	0.111746	0.154884
6	0.200786	0.154835	0.189106	0.285562	0.259603	0.102086	0.109149
7	0.468352	0.155277	0.306184	0.193895	0.055265	0.342399	0.269127
8	0.154363	0.162851	0.241826	0.404028	0.143996	0.234241	0.193440
9	0.162348	0.094238	0.377384	0.083228	0.018078	0.344290	0.090010
10	0.148592	0.203734	0.503055	0.216331	0.063362	0.109570	0.390006

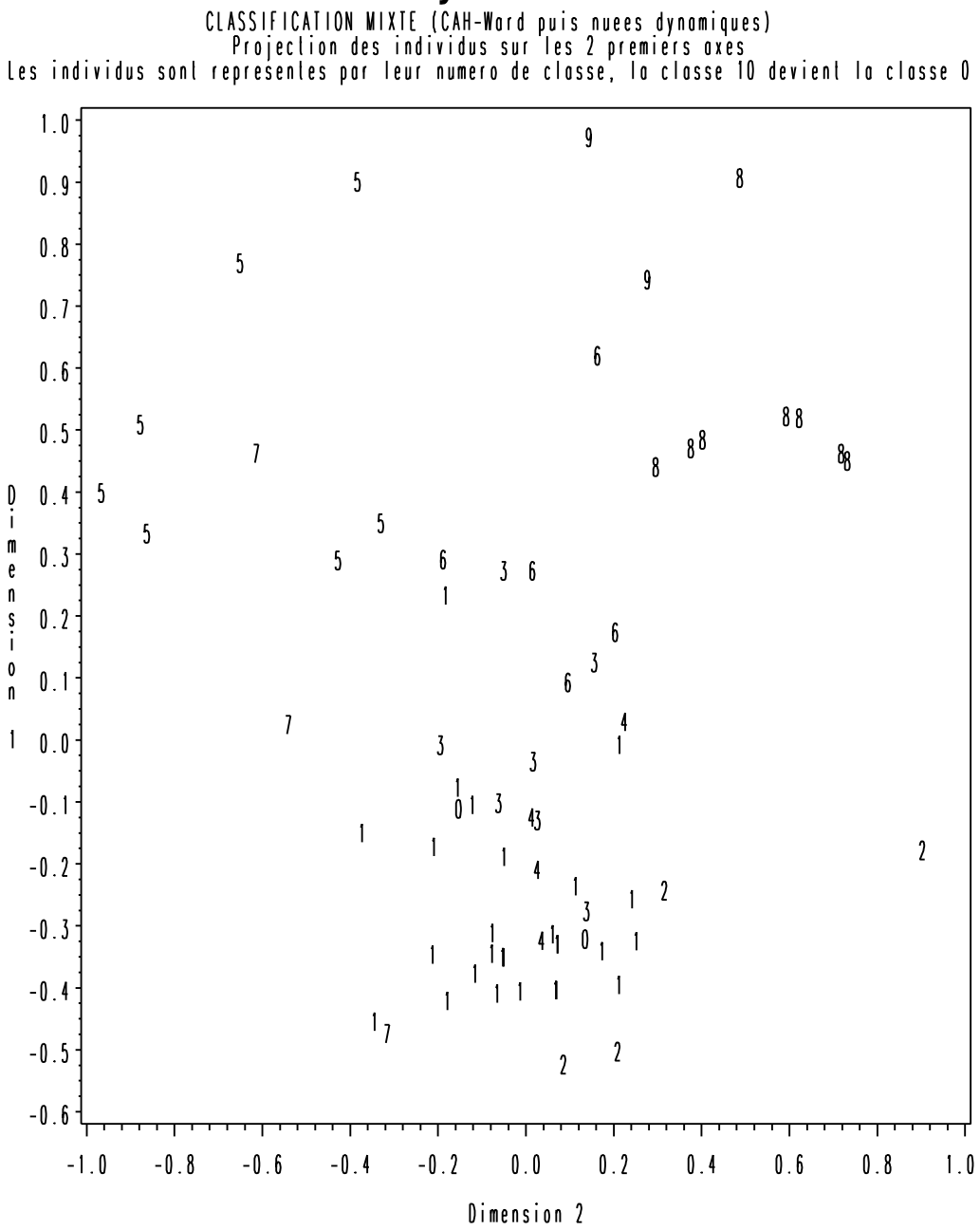
```
title2 "data=moy2 (table issue de la classif mixte)";
title3 "elle donne les moyennes des classes ainsi que d'autres
statistiques";
proc print data=moy2(drop=dim3-dim31);
* var dim1 dim2;
run;
```

stat/jaccard6 126
data=moy2 (table issue de la classif mixte)
elle donne les moyennes des classes ainsi que d'autres statistiques
11:45 Monday, October 20, 1997

OBS	_CRIT_	CLUSTER	_FREQ_	_RMSSTD_	_RADIUS_	_NEAR_	_GAP_	DIM1	DIM2
1	0.12874	1	30	0.11645	1.00853	3	0.68456	-0.28426	-0.02630
2	0.12874	2	4	0.15686	1.02369	1	0.95574	-0.35700	0.37823
3	0.12874	3	7	0.11117	0.73510	1	0.68456	-0.01700	0.00420
4	0.12874	4	4	0.12585	0.64953	1	0.78302	-0.15252	0.07467
5	0.12874	5	7	0.16482	1.04807	3	1.05107	0.51210	-0.64255
6	0.12874	6	5	0.15551	0.88039	3	0.84306	0.29509	0.05777
7	0.12874	7	3	0.17753	0.98804	1	1.12793	0.00918	-0.49023
8	0.12874	8	8	0.17914	1.06450	3	1.02619	0.53691	0.52825
9	0.12874	9	2	0.16686	0.65692	8	1.27736	0.86239	0.20992
10	0.12874	10	2	0.17124	0.67418	1	1.26473	-0.21138	-0.00907

```
data mixt2;
set mixt;
clus=put(cluster,2.); * text ne peut prendre que des chaines de;
text = substr(clus,2); * caracteres. On traduit donc cluster en clus;
*text=_name_;
y = dim1;
x = dim2;
xsys = '2';
ysys = '2';
size = 1;
label y = "Dimension 1";
label x = "Dimension 2";
keep x y text xsys ysys size;
run;
*proc print;
*run;
filename figure "plot3.eps";
title2 "CLASSIFICATION MIXTE (CAH-Ward puis nuees dynamiques)";
title3 "Projection des individus sur les 2 premiers axes";
title4 "Les individus sont representes par leur numero de classe, la "
       "classe 10 devient la classe 0";
proc gplot data=mixt2;
symbol1 i=none;
plot y*x=1 / annotate=mixt2 frame;
run;
quit;
```

stat/jaccard6



Méthode des bactériologistes
Distance Jaccard

```
data distjacc(type=distance);  *la table creee est une table de dist.;
  array dj(*) dj1-dj72;        * il y a 72 individus;
  retain dj1-dj72 .;           *initialisation des 72 ind. (a val.manq.);
do row=1 to 72;
  set naza1 point=row;
  array caract(*) c4--e11;             * liste des variables numeriques;
  array save(*) save1-save31;          * il y en a 31;
  do g=1 to 31;
    save(g)=caract(g);
  end;

    do col=1 to row;
      set naza1(drop=sou) point=col;
      cont00=0;
      cont01=0;
      cont10=0;
      cont11=0;
      do g=1 to 31;
        if caract(g)=0 and save(g)=1 then cont01=cont01+1;
        if caract(g)=1 and save(g)=0 then cont10=cont10+1;
        if caract(g)=0 and save(g)=0 then cont00=cont00+1;
        if caract(g)=1 and save(g)=1 then cont11=cont11+1;
      end;
      dj(col)=1-(cont11/(cont01+cont10+cont11));
    end;
    output;
  end;
  stop;
*  keep dj1-dj72 sou;*!!!!!!!!!!!!!!!!!!!!!!!!!;
   keep sou dj1-dj72;*!!!!!!!!!!!!!!!!!!!!!!!!!;
 run;

title2 "tableau de distance : 1 - indice de Jaccard";
proc print data=distjacc(obs=10);
  id sou;
  var dj1-dj10;
run;
```

Distance Jaccard

```
stat/jaccard6                                  127
                tableau de distance : 1 - indice de Jaccard
                                       11:45 Monday, October 20, 1997
```

SOU	DJ1	DJ2	DJ3	DJ4	DJ5	DJ6	DJ7	DJ8	DJ9	DJ10
B2	0.00000	.	.	.	.	.	.	.	.	.
B4	0.16667	0.00000	.	.	.	.	.	.	.	.
B5	0.17647	0.22222	0.00000	.	.	.	.	.	.	.
B7	0.11765	0.16667	0.17647	0.00000	.	.	.	.	.	.
B8	0.33333	0.36842	0.18750	0.33333	0.00000	.	.	.	.	.
B10	0.33333	0.36842	0.29412	0.33333	0.35294	0.00000	.	.	.	.
B11	0.23529	0.27778	0.06667	0.23529	0.13333	0.25000	0.00000	.	.	.
B12	0.30000	0.25000	0.26316	0.30000	0.31579	0.22222	0.22222	0.00000	.	.
B14	0.17647	0.22222	0.12500	0.17647	0.18750	0.18750	0.06667	0.16667	0.000	.
B15	0.17647	0.22222	0.00000	0.17647	0.18750	0.29412	0.06667	0.26316	0.125	0

CAH sur le tableau de distance de Jaccard

```
title2 "CAH sur le tableau de distance (de Jaccard)";
proc cluster data=distjacc rsquare rmsstd
             outtree=treejacc
             method=average;
  var dj1-dj72;
  id sou;
  run;
```

CAH sur le tableau de distance (de Jaccard)

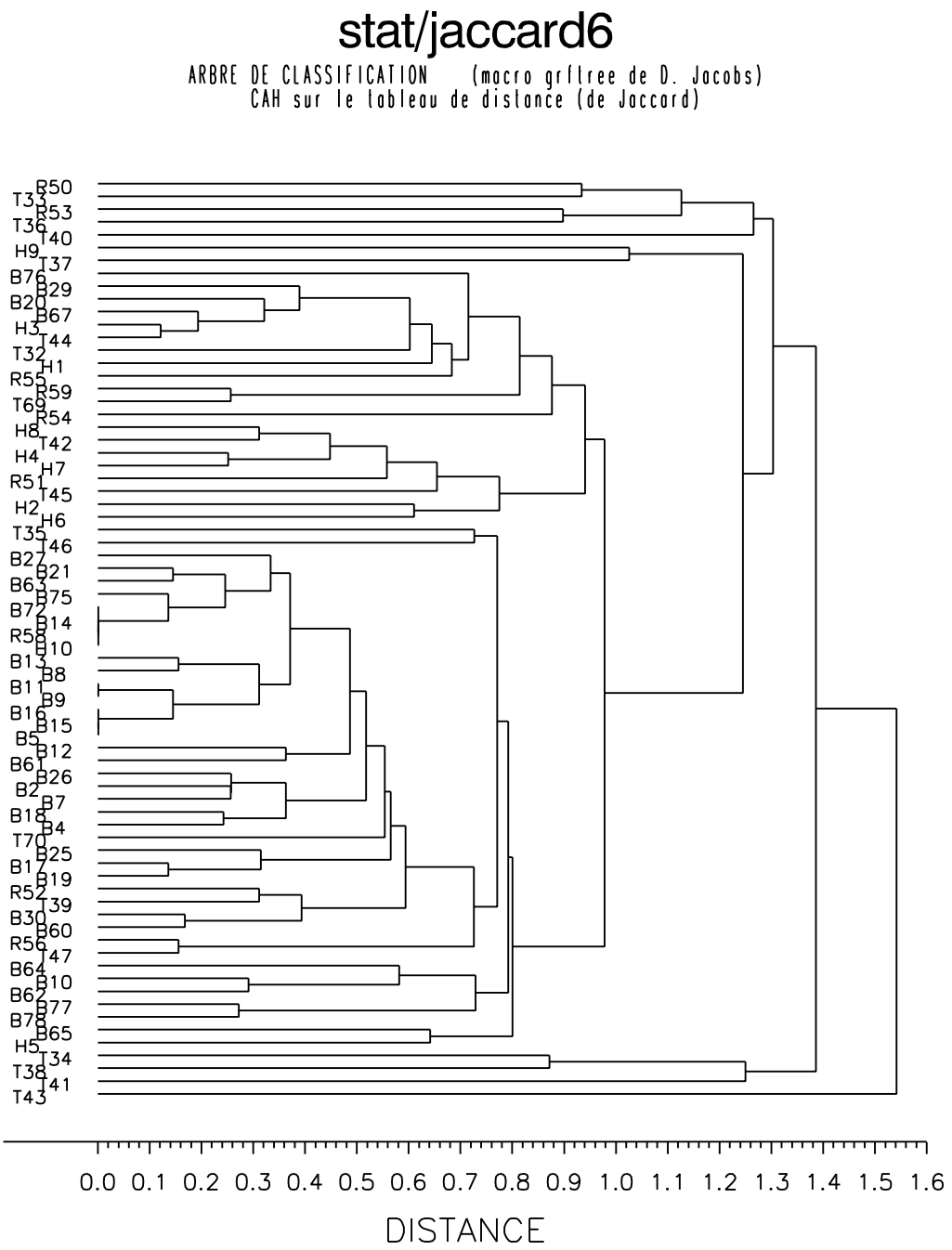
11:45 Monday, October 20, 1997

Average Linkage Cluster Analysis

Root-Mean-Square Distance Between Observations = 0.458716

NCL	-Clusters Joined-		FREQ	RMS STD	SPRSQ	RSQ	Norm RMS Dist	Tie
71	B5	B15	2	0.00000	0.000000	1.000000	0.000000	T
70	CL71	B16	3	0.00000	0.000000	1.000000	0.000000	T
69	B11	B9	2	0.00000	0.000000	1.000000	0.000000	T
68	B14	R58	2	0.00000	0.000000	1.000000	0.000000	T
67	CL68	H10	3	0.00000	0.000000	1.000000	0.000000	T
66	CL67	B72	4	0.00000	0.000000	1.000000	0.000000	
65	T44	H3	2	0.03928	0.000207	0.999793	0.121111	
64	B17	B19	2	0.04419	0.000261	0.999532	0.136250	T
63	CL66	B75	5	0.02795	0.000418	0.999114	0.136250	
62	CL70	CL69	5	0.03651	0.000714	0.998400	0.145333	T
61	B21	B63	2	0.04714	0.000297	0.998102	0.145333	
60	T47	R56	2	0.05051	0.000342	0.997761	0.155714	T
59	B8	B13	2	0.05051	0.000342	0.997419	0.155714	T
58	B30	B60	2	0.05439	0.000396	0.997023	0.167692	
57	CL65	B67	3	0.05611	0.000636	0.996387	0.193799	
56	B4	B18	2	0.07857	0.000826	0.995560	0.242222	
55	CL63	CL61	7	0.05918	0.002098	0.993463	0.245717	
54	H4	H7	2	0.08159	0.000891	0.992571	0.251538	
53	B2	B7	2	0.08319	0.000926	0.991645	0.256471	T
52	CL53	B26	3	0.08319	0.000926	0.990719	0.256471	T
51	T69	R59	2	0.08319	0.000926	0.989792	0.256471	
50	B77	B78	2	0.08839	0.001046	0.988746	0.272500	
49	B10	B62	2	0.09428	0.001190	0.987556	0.290667	
48	T39	R52	2	0.10102	0.001366	0.986190	0.311429	T
47	T42	H8	2	0.10102	0.001366	0.984824	0.311429	
46	CL62	CL59	7	0.07501	0.003464	0.981360	0.311795	
45	CL64	B25	3	0.08711	0.001770	0.979590	0.314474	T
44	CL57	B20	4	0.08364	0.001966	0.977624	0.321021	
43	CL55	B27	8	0.07450	0.002387	0.975237	0.333339	
42	CL52	CL56	5	0.10501	0.003225	0.972011	0.363333	T
..	....	....	.	.....	.....	.....	.....	

```
title2 "ARBRE DE CLASSIFICATION    (macro grftree de D. Jacobs)";
title3 "CAH sur le tableau de distance (de Jaccard)";
filename figure "jacc.eps";
%grftree(clusdsn=treejacc,items=72,axis=d);*!!!!!!!!!!!!!!!!!!!;
run;
```

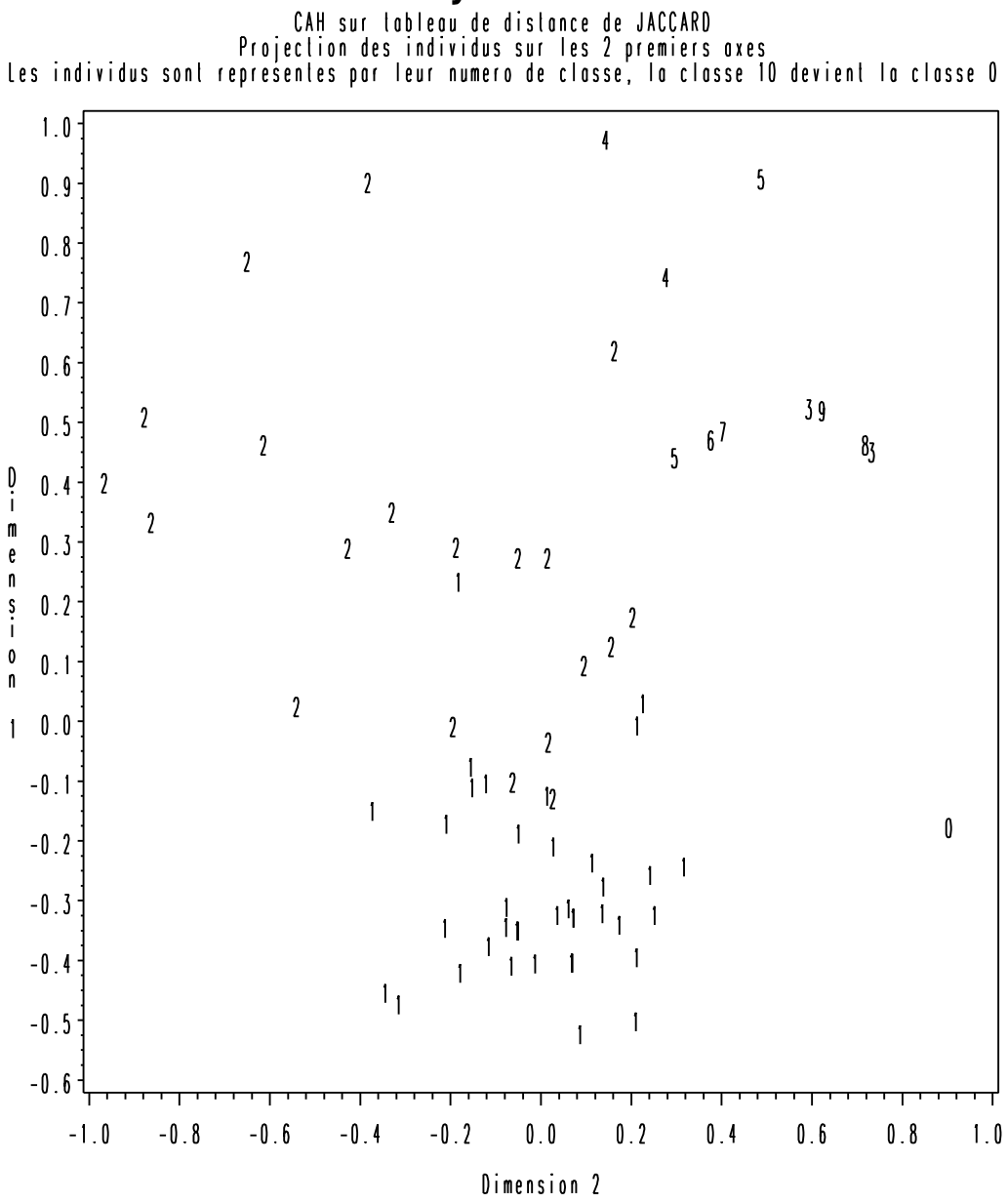


```
proc tree data=treejacc noprint n=10 out=treej;
  id sou;
run;

proc sort data=trees; by _name_;
proc sort data=treej; by sou;

data treej2;
  merge treej trees(keep=_name_ dim1 dim2 rename=(_name_=sou));
  by sou;
clus=put(cluster,2.);   * text ne peut prendre que des chaines de;
text = substr(clus,2);  * caracteres. On traduit donc cluster en clus;
*text=_name_;
y = dim1;
x = dim2;
xsys = '2';
ysys = '2';
size = 1;
label y = "Dimension 1";
label x = "Dimension 2";
keep x y text xsys ysys size;
run;
*proc print;
*run;
filename figure "plot4.eps";
title2 "CAH sur tableau de distance de JACCARD";
title3 "Projection des individus sur les 2 premiers axes";
title4 "Les individus sont representes par leur numero de classe, la "
       "classe 10 devient la classe 0";
footnote1 "Cette projection est discutable puisque les distances "
          "utilisees ";
footnote2 "dans la CAH-Jaccard et dans l'AFC ne sont pas les meme"
          "s         ";
proc gplot data=treej2;
  symbol1 i=none;
  plot y*x=1 / annotate=treej2 frame;
run;
*footnote;
run;
quit;
```

stat/jaccard6



Celle projection est discutable puisque les distances utilisees dans la CAH-Jaccard et dans l'AFC ne sont pas les memes