Fiche de Biostatistique

Problèmes d'analyse des données

D. Chessel

Résumé

La fiche contient 4 problèmes d'analyse des données.

Plan

1. Crânes

DEA AMSB - Février 96

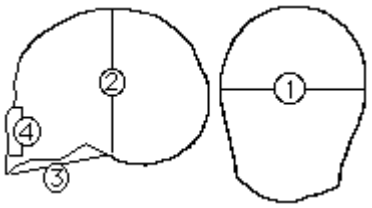
L'exercice est proposé dans l'ouvrage de Manly, B.F. (1994) *Multivariate Statistical Methods. A primer*. Second edition. Chapman & Hall, London. 1-215. L'exemple est traité pp. 6, 13, 51, 64, 72, 107, 112 et 117.

1 — Données traitées

Les mesures concernent 5 groupes de 30 crânes égyptiens. Les classes sont :

1 - période prédynastique ancienne (4000 avant JC)
2 - période prédynastique récente (3300 avant JC)
3 - 12 et 13 ème dynastie (1850 avant JC)
4 - période de Ptolémée (200 avant JC)
5 - période romaine (150 après JC)

Les variables sont définies par :



On utilise divers modules du logiciel ADE-4 pour obtenir les résultats suivants. On établit les données dans un fichier Skulls binaire à 150 lignes et 4 colonnes et un fichier Cla contenant une seule variable qualitative à 5 modalités représentées 30 fois.

2 — Analyses de variance

Le module Discrimin : Anova1-FF donne :

```
variable 1 from Skulls versus variable 1 from Cla
-----------------------------------------------------------
|Source  |        SS|d.f.|        MS|         F|     Proba|
|Between |     502.8|   4|     125.7|     5.955| 0.0002196|
|Within  |      3061| 145|     21.11|
|Total   |      3564| 149|
-----------------------------------------------------------

variable 2 from Skulls versus variable 1 from Cla
-----------------------------------------------------------
|Source  |        SS|d.f.|        MS|         F|     Proba|
|Between |     229.9|   4|     57.48|     2.447|   0.04839|
|Within  |      3405| 145|     23.48|
|Total   |      3635| 149|
-----------------------------------------------------------
```

```
variable 3 from Skulls versus variable 1 from Cla
------------------------------------------------------------
|Source  |        SS|d.f.|        MS|         F|     Proba|
|Between |     803.3|   4|     200.8|     8.306| 7.349E-06|
|Within  |      3506| 145|     24.18|
|Total   |      4309| 149|
------------------------------------------------------------

variable 4 from Skulls versus variable 1 from Cla
------------------------------------------------------------
|Source  |        SS|d.f.|        MS|         F|     Proba|
|Between |      61.2|   4|      15.3|     1.507|    0.2019|
|Within  |      1472| 145|     10.15|
|Total   |      1533| 149|
------------------------------------------------------------
```

3 — Analyse en composantes principales

Le module PCA : Correlation matrix PCA donne :

```
Classical Principal Component Analysis (Hotteling 1933)
Input file: Skulls
---- Row weights:
File Skulls.cnpl contains the row weights
It has 150 rows and 1 column
Each row has 6.6667e-03 weight (Sum = 1)
---- Column weights:
File Skulls.cnpc contains the column weights
It has 150 rows and 1 column
Each column has unit weight (Sum = 4)
---- Table:
File Skulls.cnta contains the centred and normed table
Zero mean and unit variance for each column
It has 150 rows and 4 columns
File :Skulls.cnta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-3.072e+00| 2.878e+00|
|   2|-2.549e+00| 2.530e+00|
|   3|-2.884e+00| 3.272e+00|
|   4|-2.169e+00| 2.836e+00|
|----|----------|----------|
---- Info: means and variances
File Skulls.cnma contains the descriptive of the analysis
It contains successively:
     Number of rows: 150
     Number of columns: 4
     means and variances:
     Col.:   1 | Mean:  1.3397e+02 | Variance: 2.3759e+01
     Col.:   2 | Mean:  1.3255e+02 | Variance: 2.4234e+01
     Col.:   3 | Mean:  9.6460e+01 | Variance: 2.8728e+01
     Col.:   4 | Mean:  5.0933e+01 | Variance: 1.0222e+01
-----------------------------------------------------
File Skulls.cn+r contains the Correlation matrix
from statistical triplet Skulls.cnta
It has 4 rows and 4 columns
---------------------- Correlation matrix -------------------
[  1] 1000
[  2]  -62 1000
[  3] -157  264 1000
[  4]  183  147   -6 1000
```

```
------------------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: Skulls.cnta
--- Number of rows: 150, columns: 4
----------------------
Total inertia:        4
----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +1.3373E+00 +0.3343 +0.3343   |02   +1.2064E+00 +0.3016 +0.6359 |
03   +7.6240E-01 +0.1906 +0.8265   |04   +6.9391E-01 +0.1735 +1.0000 |
```

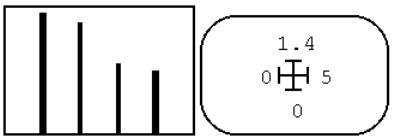


fig 1

```
File Skulls.cnvp contains the eigenvalues and relative inertia for each
axis
--- It has 4 rows and 2 columns

File Skulls.cnco contains the column scores
--- It has 4 rows and 2 columns
File :Skulls.cnco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-7.776e-01| 4.706e-01|
|   2|-8.211e-01| 1.401e-02|
|----|----------|----------|
```

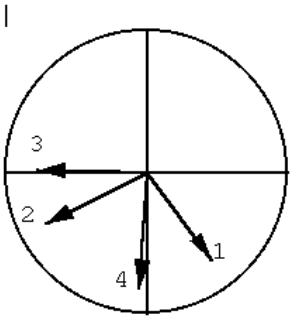


fig 2

```
File Skulls.cnli contains the row scores
--- It has 150 rows and 2 columns
File :Skulls.cnli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-2.916e+00| 3.231e+00|
|   2|-3.139e+00| 3.401e+00|
|----|----------|----------|
```

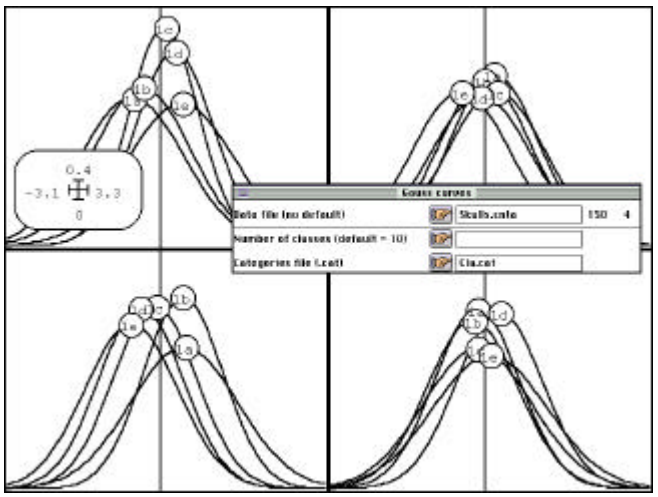


fig 3

4 — ACP inter-classes

L'analyse qui précède est associée à la partition par Discrimin : Initialize/LinkPrep :

L'option Discrimin : Between analysis/Test donne :

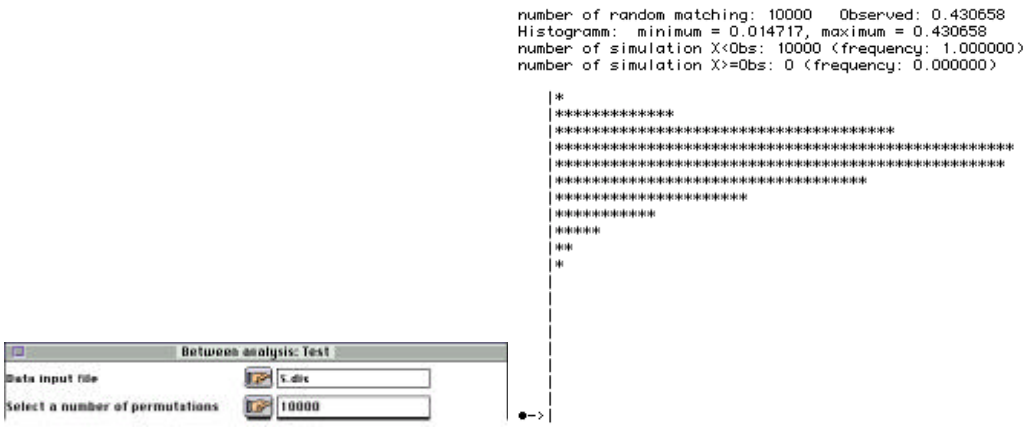


fig 4

L'option Discrimin : Between analysis/Run donne :

```
Between-class analysis
Categories defined by column 1 of file Cla
Input statistical triplet: table Skulls.cnta
Number of rows: 150, columns: 4
total inertia: 4.000000
------------------------------------------------------------
File S.beta contains multivariate means by classes
It has 5 rows and 4 columns
File S.bepl contains class weights
It has 5 rows and 1 column
File S.bepc contains column weights
It has 4 rows and 1 column
This statistical triplet is the gravity centers one
between-class inertia 0.430658 (ratio: 0.107665)
------------------------------------------------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
```

```
01   +3.8382E-01 +0.8913 +0.8913   |02   +3.1271E-02 +0.0726 +0.9639 |
03   +1.3758E-02 +0.0319 +0.9958   |04   +1.8041E-03 +0.0042 +1.0000 |

File S.bevp contains the eigenvalues and relative inertia for each axis
It has 4 rows and 2 columns

File S.bec1 contains column scores with unit norm
It has 4 rows and 1 columns
File :S.bec1
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -0.5986 Maxi =  0.6926

File S.beli contains standard gravity center scores with lambda norm
It has 5 rows and 1 columns

File :S.beli
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -0.82345   Maxi =  0.76771

File S.bels contains standard row scores with lambda norm
It has 150 rows and 1 columns

File :S.bels
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -3.028  Maxi =  2.5869
```

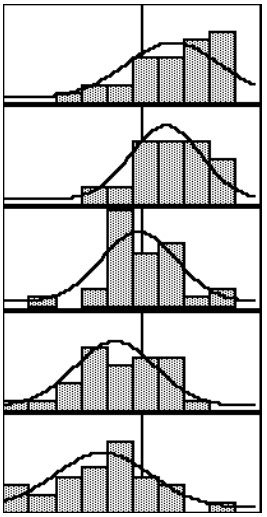


fig 5

```
File S.beco contains standard column scores with lambda norm
It has 4 rows and 1 columns
File :S.beco
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -0.37086   Maxi =  0.42909
------------------------------------------------------------
```

S.beli

	1
1	0.7677
2	0.6020
3	-0.0084
4	-0.5378
5	-0.8234

S.beco

	1
1	-0.3709
2	0.1866
3	0.4291
4	-0.1654

S.bec1

	1
1	-0.5986
2	0.3012
3	0.6926
4	-0.2669

5 — Distances temporelles et distances inter-classes

On décrit l'écart temporel entre les échantillons recueillis par une matrice de distance :

Dist_Tempo

	1	2	3	4	5
1	0	700	2150	3800	4150
2	700	0	1450	3100	3450
3	2150	1450	0	1650	2000
4	3800	3100	1650	0	350
5	4150	3450	2000	350	0

On décrit la distance entre groupes sur l'axe 1 de l'analyse inter-classes par :

Triplet To Distance Matrix

Input file	S.beta	5	4
Option: Output file			
Option: default = between rows			

S_MDbe

	1	2	3	4	5
1	0.0000	0.2910	0.8649	1.3180	1.6111
2	0.2910	0.0000	0.7528	1.1954	1.4294
3	0.8649	0.7528	0.0000	0.6386	0.9505
4	1.3180	1.1954	0.6386	0.0000	0.5007
5	1.6111	1.4294	0.9505	0.5007	0.0000

On compare les deux par un test de Mantel (Distances : Mantel Test) :

```
Correlation between two distance matrices
-----------------------------------------
First input file: Dist_Tempo
It has 5 rows and 5 columns
-----------------------------------------
Second input file: S_MDbe
It has 5 rows and 5 columns
-----------------------------------------
r index :  9.667e-01
Permutation signicant test (Manly 1994 p. 73)
Test on Z value (formula 5.9 p. 70)
number of random matching: 10000   Observed: 26616.509766
Histogramm:  minimum = 19723.402344, maximum = 26764.597656
number of simulation X<Obs: 9745 (frequency: 0.974500)
number of simulation X>=Obs: 255 (frequency: 0.025500)

    |************************************
    |*************************************************
    |*****************************************
    |************************************
    |**********************
    |
    |
    |
    |
    |*********
    |**********************
    |******************************
    |*********
    |**
    |****
    |
    |
    |****
    |*******
●->|*******
```

fig 6

6 — Analyse discriminante

On exécute l'analyse discriminante par Discrimin : Discriminant analysis/Ru. On obtient :

```
File S.dima contains the parameters:
   input file: Skulls.cnta
   categorical variable file: Cla
   n° of categorical variable used: 1

Discriminant analysis
Categories defined by column 1 of file Cla
Input statistical triplet: table Skulls.cnta
Number of rows: 150, columns: 4
total inertia (norm C- generalised inverse) = rank of the data matrix: 4.000000
```

```
between-class inertia (norm C-): 0.353306 (ratio: 0.088326)

Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.9829E-01 +0.8443 +0.8443   |02   +3.7535E-02 +0.1062 +0.9505 |
03   +1.5462E-02 +0.0438 +0.9943   |04   +2.0163E-03 +0.0057 +1.0000 |

File S.divp contains the eigenvalues and relative inertia for each axis
It has 4 rows and 2 columns

File S.difa contains coefficient of discriminant scores
It has 4 rows and 2 columns
File :S.difa
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -0.66273   Maxi =  0.52608
Col.: 2  Mini =  -0.36429   Maxi =  1.032

File S.dili contains canonical row scores with unit norm
It has 150 rows and 2 columns

File :S.dili
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -2.3822 Maxi =  2.5793
Col.: 2  Mini =  -2.4798 Maxi =  2.2028
```

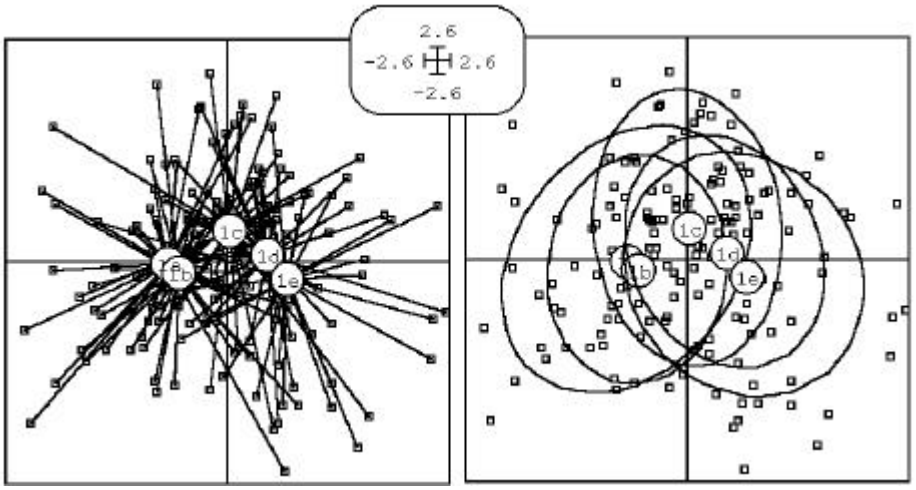


fig 7

```
File S.diax contains the principal axes
It has 4 rows and 2 columns
File :S.diax
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -0.78781   Maxi =  0.68092
Col.: 2  Mini =  -0.11947   Maxi =  0.88788

File S.dicp contains the correlations between PCA scores
and DA scores. It has 4 rows and 2 columns
File :S.dicp
-----------------------Minimum/Maximum:
Col.: 1  Mini =  -0.4636 Maxi =  0.86454
Col.: 2  Mini =  -0.87907   Maxi =  -0.069077
```

S.difa

	1	2
1	0.5261	0.1884
2	-0.1553	1.0320
3	-0.6627	-0.3643
4	0.2257	-0.2466

S.dicp

	1	2
1	0.8645	-0.3545
2	-0.4636	-0.3111
3	0.0471	-0.0691
4	-0.1882	-0.8791

Dans le module Curves : Lines, on observe :

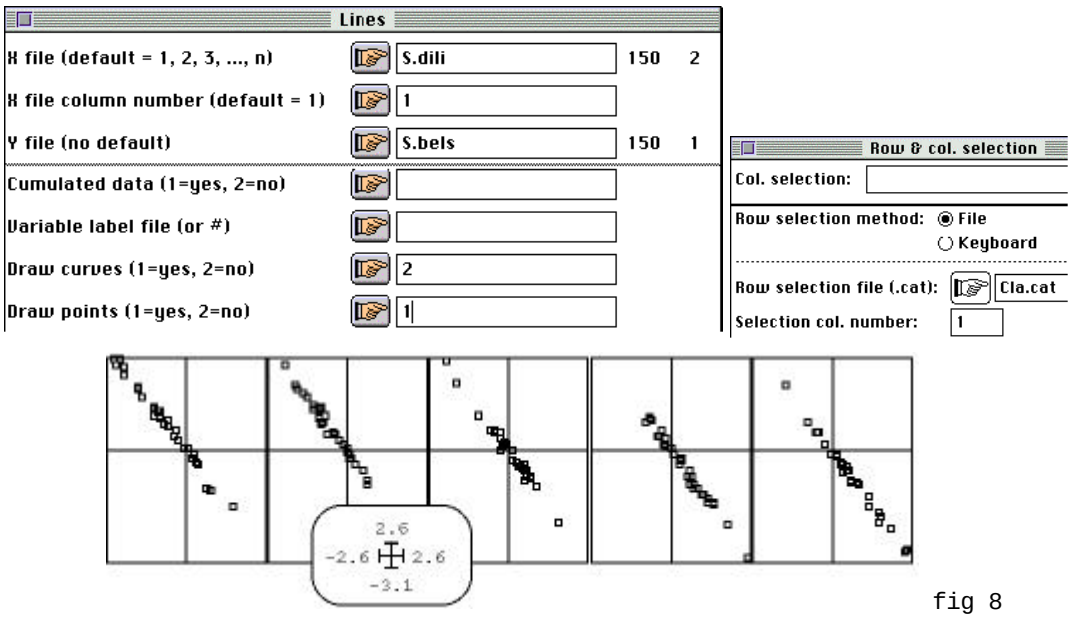
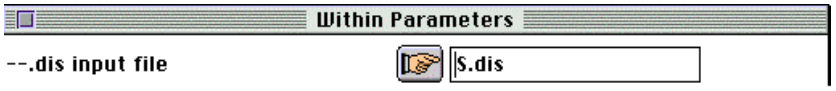


fig 8

7 — Variabilité intra-classes

Dans le module Discrimin : Within Parameters, on observe :




```
------------------------------------------------------------
Within-class analysis
Categories defined by column 1 of file Cla
Input statistical triplet: table Skulls.cnta
Number of rows: 150, columns: 4
total inertia: 4.000000
------------------------------------------------------------
File S.whmoy contains the array of means
It has 5 rows and 4 columns
```

S.whmoy

	1	2	3	4
1	-0.5348	0.2140	0.5050	-0.1251
2	-0.3296	0.0311	0.4863	-0.2189
3	0.1012	0.2545	-0.0795	-0.1147
4	0.3132	-0.0501	-0.3595	0.3232
5	0.4500	-0.4496	-0.5523	0.1355

```
File S.whvar contains the the array of variances
It has 5 rows and 4 columns

File S.whdiv contains within class inertia
It has 5 rows and 1 column
```

S.whvar

	1	2	3	4
1	1.0704	0.7967	1.1651	0.7222
2	0.9413	0.8614	0.6357	0.8262
3	0.4931	0.9887	0.6973	1.1914
4	0.6250	1.0513	0.7095	0.7532
5	1.1647	0.9857	0.8604	1.3075

S.whdiv

	1
1	3.7544
2	3.2647
3	3.3705
4	3.1389
5	4.3183

8 — Distances de Mahalanobis

La matrice de distances entre groupes est calculée en reprenant l'ACP de départ :

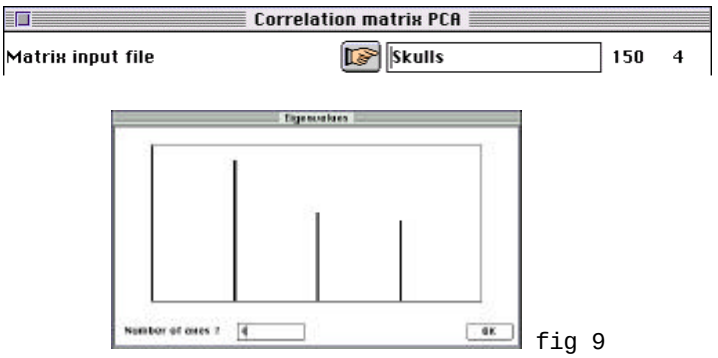
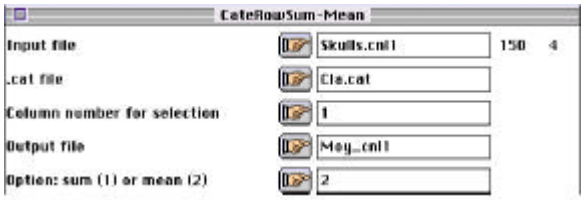


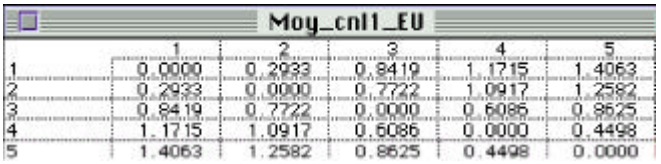
fig 9

On récupère les coordonnées normalisées (composantes principales normées) par DDUtil : Add normed scores :

Le fichier Skulls.cnl1 est utilisé pour calculer les moyennes par classe :



On calcule les distances euclidiennes classiques entre groupes :

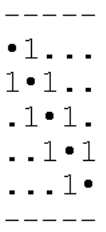


Moy_cnl1_EU

	1	2	3	4	5
1	0.0000	0.2933	0.8419	1.1715	1.4063
2	0.2933	0.0000	0.7722	1.0917	1.2582
3	0.8419	0.7722	0.0000	0.6086	0.8625
4	1.1715	1.0917	0.6086	0.0000	0.4498
5	1.4063	1.2582	0.8625	0.4498	0.0000

L'option édite Distances : Minimal Spanning Tree donne :


```
Neighboorhood relation from Minimal Spanning Tree
Input file (distances matrix): Moy_cnl1_EU

Neighboorhood graph in binary file: Moy_cnli_EU$G
It contains graph matrix (LEBART's M) with 5 rows and columns
Neighboorhood weights in binary file: Moy_cnli_EU$Gpl
It contains  5 rows and 1 column
Symetric neighboorhood matrix (. for 0):
```



01 D'accord - Pas d'accord ? Les techniques d'élimination de l'effet taille sont, dans le cas étudié, sans objet.

02 D'accord - Pas d'accord ? L'étude morphométrique rapportée est caractéristique de la mentalité "raciologique" du début du siècle. Elle se trouve invalidée par l'analyse statistique.

03 D'accord - Pas d'accord ? L'évolution morphométrique mise en évidence semble relativement continue.

04 D'accord - Pas d'accord ? L'analyse discriminante permettra facilement de dater un crâne d'origine inconnue.

05 D'accord - Pas d'accord ? Les schémas de dualité utilisés en ACP interclasses et en analyse discriminante ne diffèrent que par un seul paramètre.

06 D'accord - Pas d'accord ? S'il y avait un plus grand nombre de groupes d'individus, on devrait s'en tenir à l'ACP interclasses.

07 D'accord - Pas d'accord ? La figure 8 est la conséquence graphique des valeurs éditées dans les fenêtres S.bec1 et S.difa .

08 D'accord - Pas d'accord ? Le pourcentage de variance interclasse de la première colonne du fichier S.bels ne peut dépasser 29.8% et sa variance elle-même ne peut dépasser 1.34.

09 D'accord - Pas d'accord ? L'ACP interclasses et l'analyse discriminante donne des résultats voisins comme c'est très souvent le cas.

10 D'accord - Pas d'accord ? Le choix d'une bonne méthode statistique ne dépend que de la revue dans laquelle on veut publier.

2. Municipales

DEUG MASS — 1997

Les résultats à l'élection législative du 25 mai 1997 dans la quatrième circonscription du Rhône ont été publié par *Le Progrès* et rendus disponibles sur le serveur http://www.leprogres.fr :

Il y a p = 12 candidats désignés par leur appartenance politique :

1	LO	Lutte Ouvrière
2	UDF-RPR	Union pour la démocratie française - Rassemblement pour la République
3	FN	Front national
4	MEI	Mouvement écologiste indépendant
5	PS	Parti socialiste
6	GE	Génération écologie
7	GAE	Gauche alternative écologie
8	LDI	La droite indépendante
9	Verts	les Verts
10	IR	Initiative républicaine
11	MDC	Mouvements des citoyens
12	SE	Sans étiquette

Les 71 bureaux de votes sont regroupés en n = 24 groupes par leur adresse. La carte de répartition géographique des 24 unités spatiales est consignée dans la figure 1. Le tableau 1 donne les résultats (nombre de voix obtenues par chacun des 12 candidat dans chacun des 24 groupes de bureaux de votes). Le tableau 2 donne les mêmes résultats en pourcentages. Ce tableau est conservé dans un fichier portant le nom PC.

Une analyse en composantes principales centrée est exécutée sur le tableau PC et le listing est consigné dans l'annexe 1. La carte factorielle des variables sur le plan des deux premières composantes principales est tracée dans la figure 2. La première coordonnée factorielle des lignes du tableau est cartographiée dans la figure 3. Les contributions à la trace et les deux premières coordonnées des variables sont éditées dans l'annexe 2. Les statistiques d'inertie sur les variables sont reportées dans l'annexe 3. La matrice des covariances est éditée dans l'annexe 4.

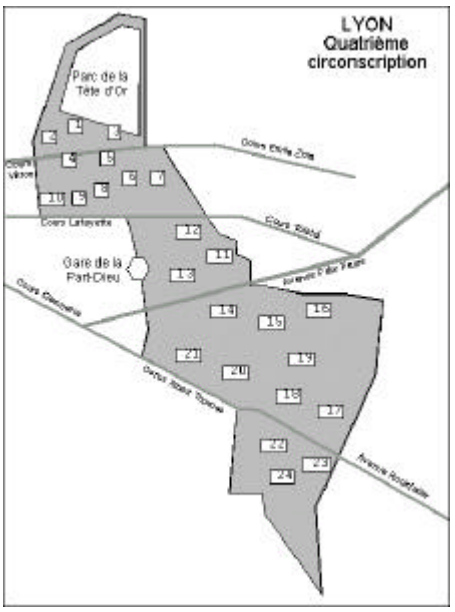


Figure 1 : 24 unités spatiales de regroupement des bureaux de la circonscription Lyon-4.

45	2403	528	45	464	59	48	260	86	14	75	3
64	2238	478	30	494	74	69	194	61	14	55	5
29	836	233	21	273	35	26	77	27	8	50	3
46	676	245	20	295	32	29	75	32	13	49	2
54	613	218	22	324	30	37	37	42	16	60	5
34	674	180	9	263	42	18	62	29	6	52	3
42	633	210	24	344	27	33	66	54	23	73	1
25	680	137	9	136	17	14	77	27	4	20	0
18	290	82	4	93	15	6	37	18	7	9	1
31	870	186	24	258	39	28	106	39	8	45	2
105	1098	510	46	767	79	59	118	106	16	171	5
35	366	137	23	290	34	24	36	33	6	38	0
56	565	272	21	454	34	33	46	42	10	83	3
51	431	206	18	257	23	18	40	29	6	51	0
31	440	224	20	266	29	33	38	21	10	37	2
26	420	226	22	263	25	30	42	25	12	74	0
18	435	214	16	202	22	20	29	16	7	57	0
33	473	186	15	271	27	17	44	44	4	50	2
38	525	217	28	305	24	22	51	35	9	43	0
30	488	176	25	275	18	27	52	23	10	39	0
83	737	353	34	612	48	49	69	78	15	109	3
90	630	458	47	695	71	42	73	40	11	208	4
38	283	256	21	292	22	21	42	13	10	78	2
58	284	281	14	343	37	30	32	16	21	105	2

Tableau 1 : Résultats bruts.

0.011	0.596	0.131	0.011	0.116	0.015	0.012	0.066	0.021	0.003	0.019	0.001
0.017	0.593	0.127	0.008	0.131	0.020	0.018	0.051	0.016	0.004	0.015	0.001
0.018	0.517	0.144	0.013	0.169	0.022	0.016	0.048	0.017	0.005	0.031	0.002
0.030	0.447	0.162	0.013	0.196	0.021	0.019	0.050	0.021	0.009	0.032	0.001
0.037	0.420	0.190	0.015	0.222	0.021	0.025	0.025	0.029	0.011	0.041	0.003
0.025	0.491	0.131	0.007	0.192	0.031	0.013	0.045	0.021	0.004	0.039	0.002
0.027	0.414	0.137	0.016	0.225	0.018	0.022	0.043	0.035	0.015	0.049	0.001
0.022	0.593	0.120	0.008	0.119	0.015	0.012	0.067	0.024	0.003	0.017	0.000
0.031	0.500	0.141	0.007	0.160	0.026	0.010	0.064	0.031	0.012	0.016	0.002
0.019	0.532	0.114	0.015	0.158	0.024	0.017	0.065	0.024	0.005	0.028	0.001
0.034	0.396	0.186	0.015	0.249	0.026	0.019	0.038	0.034	0.005	0.056	0.002
0.034	0.358	0.134	0.023	0.284	0.033	0.023	0.035	0.032	0.006	0.037	0.000
0.035	0.349	0.168	0.013	0.280	0.021	0.020	0.028	0.026	0.006	0.051	0.002
0.045	0.381	0.182	0.016	0.227	0.020	0.016	0.035	0.026	0.005	0.045	0.000
0.027	0.382	0.195	0.017	0.231	0.025	0.029	0.033	0.018	0.009	0.032	0.002
0.022	0.361	0.194	0.019	0.226	0.021	0.026	0.036	0.021	0.010	0.064	0.000
0.017	0.420	0.207	0.015	0.196	0.021	0.019	0.028	0.015	0.007	0.055	0.000
0.028	0.406	0.160	0.013	0.232	0.023	0.015	0.036	0.036	0.003	0.043	0.002
0.029	0.405	0.167	0.022	0.236	0.019	0.017	0.039	0.027	0.007	0.033	0.000
0.026	0.420	0.151	0.021	0.236	0.015	0.023	0.045	0.020	0.009	0.034	0.000
0.038	0.337	0.161	0.016	0.279	0.022	0.022	0.032	0.036	0.007	0.050	0.001
0.038	0.266	0.193	0.020	0.293	0.030	0.018	0.031	0.017	0.005	0.088	0.002
0.035	0.263	0.237	0.019	0.271	0.020	0.019	0.039	0.012	0.009	0.072	0.002
0.047	0.232	0.290	0.011	0.260	0.030	0.025	0.026	0.013	0.017	0.086	0.002

Tableau 2 : Résultats en pourcentages.

Annexe 1

```
Centered Principal Component Analysis (Pearson 1901)
Input file: PC
---- Row weight:
File PC.cppl contains the row weight
It has 24 rows and 1 column
Each row has 4.1667e-02 weight (Sum = 1)
---- Column weights:
File PC.cppc contains the column weights
It has 12 rows and 1 column
Each column has unit weight (Sum = 12)
---- Table:
File PC.cpta contains the (column) centred table
It has 24 rows and 12 columns
File :PC.cpta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.776e-02| 1.850e-02|
|   2|-1.860e-01| 1.780e-01|
|   3|-4.885e-02| 7.493e-02|
|   4|-8.121e-03| 7.824e-03|
|   5|-1.018e-01| 7.647e-02|
|   6|-7.786e-03| 1.084e-02|
|   7|-8.684e-03| 9.642e-03|
|   8|-1.654e-02| 2.527e-02|
|   9|-1.188e-02| 1.379e-02|
|  10|-3.934e-03| 9.806e-03|
|  11|-2.830e-02| 4.493e-02|
|  12|-1.164e-03| 2.265e-03|
|----|----------|----------|

---- Info: means and variances
File PC.cpma contains the descriptive of the analysis
It contains successively:
     Number of rows: 24
     Number of columns: 12
     means and variances:
     Col.:   1 | Mean: 2.8922e-02 | Variance: 7.9489e-05
     Col.:   2 | Mean: 4.1824e-01 | Variance: 9.8125e-03
     Col.:   3 | Mean: 1.6254e-01 | Variance: 1.0643e-03
     Col.:   4 | Mean: 1.4681e-02 | Variance: 2.0278e-05
```

```
     Col.:   5 | Mean: 2.1690e-01 | Variance: 2.7387e-03
     Col.:   6 | Mean: 2.2426e-02 | Variance: 2.3610e-05
     Col.:   7 | Mean: 1.9029e-02 | Variance: 2.1735e-05
     Col.:   8 | Mean: 4.1922e-02 | Variance: 1.5531e-04
     Col.:   9 | Mean: 2.3941e-02 | Variance: 5.4086e-05
     Col.:  10 | Mean: 7.3645e-03 | Variance: 1.2828e-05
     Col.:  11 | Mean: 4.2868e-02 | Variance: 3.8528e-04
     Col.:  12 | Mean: 1.1641e-03 | Variance: 8.0326e-07

DiagoRC: General program for two diagonal inner product analysis
Input file: PC.cpta
--- Number of rows: 24, columns: 12
-----------------------
Total inertia: 0.0143689
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +1.3455E-02 +0.9364 +0.9364   |02   +6.3097E-04 +0.0439 +0.9803 |
03   +9.7819E-05 +0.0068 +0.9871   |04   +7.0392E-05 +0.0049 +0.9920 |
05   +3.9477E-05 +0.0027 +0.9947   |06   +2.6881E-05 +0.0019 +0.9966 |
07   +2.2252E-05 +0.0015 +0.9982   |08   +1.5692E-05 +0.0011 +0.9993 |
09   +7.1496E-06 +0.0005 +0.9998   |10   +3.0347E-06 +0.0002 +1.0000 |
11   +5.5450E-07 +0.0000 +1.0000   |12   +0.0000E+00 +0.0000 +1.0000 |

File PC.cpvp contains the eigenvalues and relative inertia for each axis
--- It has 12 rows and 2 columns

File PC.cpco contains the column scores
--- It has 12 rows and 2 columns
File :PC.cpco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-5.012e-02| 9.899e-02|
|   2|-1.932e-02| 1.427e-02|
|----|----------|----------|

File PC.cpli contains the row scores
--- It has 24 rows and 2 columns
File :PC.cpli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-2.105e-01| 2.101e-01|
|   2|-5.058e-02| 6.005e-02|
|----|----------|----------|
```

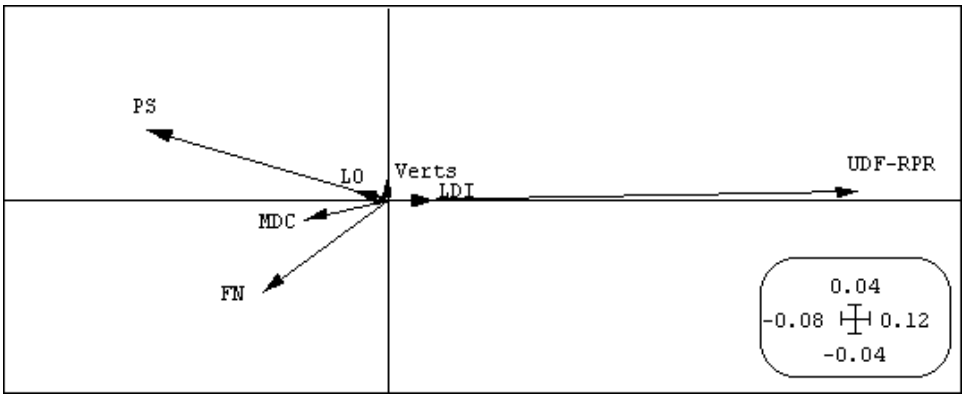


Figure 2: Carte factorielle des variables. Les points non représentés sont proches de l'origine.

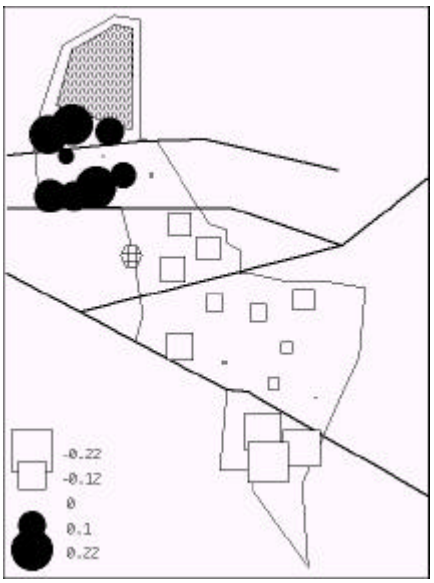


Figure 3: Cartographie de la première coordonnée factorielle de l'ACP.

Annexe 2

	Contributions	Coordonnées	
N°		F1	F2
1	0.006	-0.0069	0.0016
2	0.683	0.0990	0.0016
3	0.074	-0.0259	-0.0193
4	0.001	-0.0027	0.0008
5	0.191	-0.0501	0.0143
6	0.002	-0.0022	0.0010
7	0.002	-0.0027	0.0003
8	0.011	0.0102	-0.0005
9	0.004	0.0002	0.0052
10	0.001	-0.0015	-0.0007
11	0.027	-0.0173	-0.0044
12	0.000	-0.0001	0.0000

Annexe 3

```
Input file: PC.cpta
Number of rows: 24, columns: 12

Column inertia
All contributions are in 1/10000
----------Absolute contributions----------
|Num  |Fac 1|Fac 2|
|    1|   35|   39|
|    2| 7282|   40|
|    3|  498| 5915|
|    4|    5|   10|
|    5| 1866| 3226|
|    6|    3|   15|
|    7|    5|    1|
|    8|   78|    3|
|    9|    0|  433|
|   10|    1|    6|
|   11|  222|  307|
|   12|    0|    0|
----------Relative contributions----------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
|    1| 5996|  314|| 3688  |10000   |   55 |
|    2| 9985|    2||   11  |10000   | 6828 |
|    3| 6303| 3506||  190  |10000   |  740 |
```

```
|    4| 3505|  327|| 6167  |10000   |   14 |
|    5| 9171|  743||   84  |10000   | 1905 |
|    6| 2003|  410|| 7585  |10000   |   16 |
|    7| 3312|   48|| 6639  |10000   |   15 |
|    8| 6759|   13|| 3226  |10000   |  108 |
|    9|    4| 5053|| 4942  |10000   |   37 |
|   10| 1838|  331|| 7830  |10000   |    8 |
|   11| 7765|  504|| 1730  |10000   |  268 |
|   12|  135|   25|| 9839  |10000   |    0 |
```

Annexe 4

79e-06	-69e-05	14e-05	10e-06	36e-05	19e-06	13e-06	-67e-06	13e-06	13e-06	10e-05	19e-07
-69e-05	*********	-26e-04	-26e-05	-49e-04	-22e-05	-26e-05	10e-04	16e-06	-16e-05	-17e-04	-10e-06
14e-05	-26e-04	11e-04	57e-06	10e-04	32e-06	66e-06	-27e-05	-11e-05	48e-06	51e-05	13e-07
10e-06	-26e-05	57e-06	20e-06	15e-05	77e-08	12e-06	-27e-06	63e-08	18e-07	36e-06	-14e-07
36e-05	-49e-04	10e-04	15e-05	27e-04	12e-05	14e-05	-54e-05	56e-06	60e-06	79e-05	54e-07
19e-06	-22e-05	32e-06	77e-08	12e-05	24e-06	36e-07	-21e-06	59e-08	18e-07	38e-06	14e-07
13e-06	-26e-05	66e-06	12e-06	14e-05	36e-07	22e-06	-39e-06	-32e-07	83e-07	38e-06	23e-08
-67e-06	10e-04	-27e-05	-27e-06	-54e-05	-21e-06	-39e-06	16e-05	41e-08	-13e-06	-18e-05	-21e-07
13e-06	16e-06	-11e-05	63e-08	56e-06	59e-08	-32e-07	41e-08	54e-06	-11e-07	-31e-06	-24e-09
13e-06	-16e-05	48e-06	18e-07	60e-06	18e-07	83e-07	-13e-06	-11e-07	13e-06	26e-06	39e-08
10e-05	-17e-04	51e-05	36e-06	79e-05	38e-06	38e-06	-18e-05	-31e-06	26e-06	39e-05	24e-07
19e-07	-10e-06	13e-07	-14e-07	54e-07	14e-07	23e-08	-21e-07	-24e-09	39e-08	24e-07	80e-08

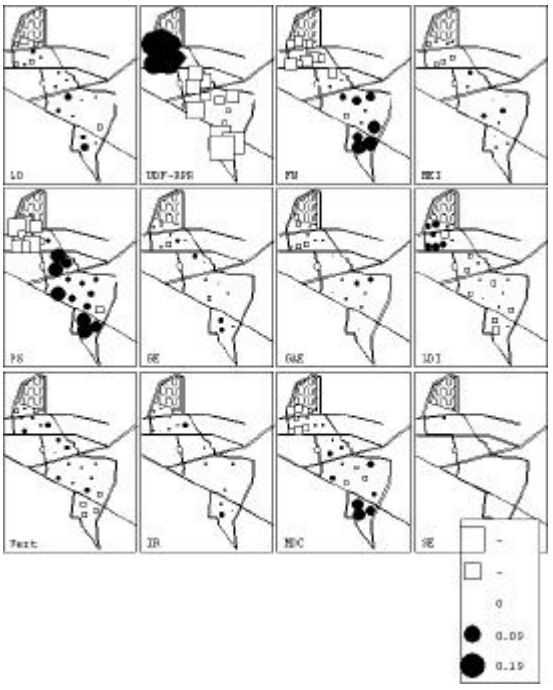


Figure 4 : Cartographie des données centrées.

01 Vrai ou Faux ? La moyenne des variables est le pourcentage des voix obtenu par chaque candidat.

02 Vrai ou Faux ? Le tableau de données après centrage est de rang 12.

03 Vrai ou Faux ? Une matrice de covariances sur *p* variables est toujours diagonalisable mais ses valeurs propres peuvent être négatives.

04 Vrai ou Faux ? L'annexe 2 donne les deux premiers vecteurs propres de la matrice de covariances.

05 Vrai ou Faux ? La valeur manquante de l'annexe 4 est 82e-05.

06 D'accord / Pas d'accord ? La circonscription est sociologiquement hétérogène.

07 D'accord / Pas d'accord ? Il est légitime de n'interpréter qu'un seul axe dans l'analyse.
08 D'accord / Pas d'accord ? La variabilité des scores des candidats est très organisée.
09 D'accord / Pas d'accord ? Les résultats permettent de distinguer trois parties dans cette circonscription.
10 D'accord / Pas d'accord ? Le découpage électoral vise à assurer la représentation de toutes les opinions.

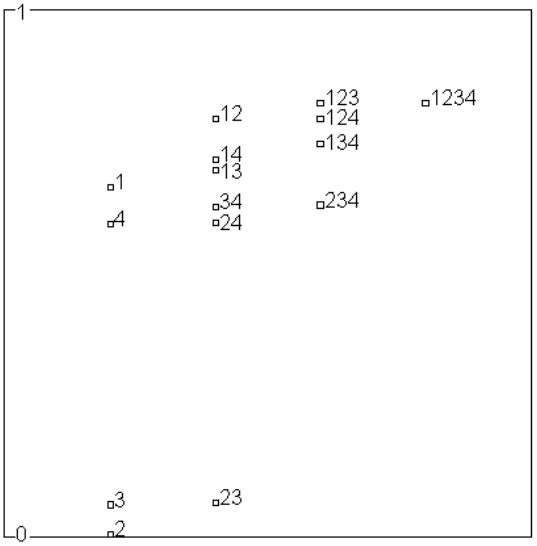
3. Européennes

Aux élections européennes de 1984, le candidat de l'extrême droite avait obtenu pour la première fois un score important. Il est calculé par région administrative (n = 21) dans le colonne A du tableau ci-dessous. Dans les mêmes régions, on connaît le taux de population immigrée de l'époque (B, en 1/100), le taux de chômage (C, en 1/100), l'évolution du taux de chômage (D, en 1/100) et le taux d'urbanisation (E, en 1/100). Le tableau normalisé associé est calculé :

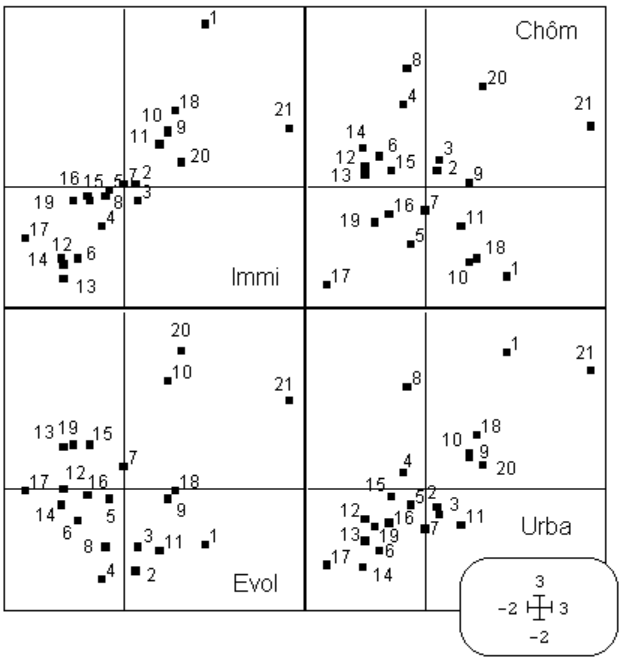
%voix	Pop. immi.	Chomage	Evol. Chom.	Urbani.					
A	B	C	D	E	A*	B*	C*	D*	E*
14.5	13.3	7.1	0.23	93.6	1.372	2.709	-1.498	-0.936	2.295
10.7	5.4	9.5	0.07	62.4	0.190	0.032	0.274	-1.366	-0.305
10.8	4.6	9.7	0.22	60.7	0.221	-0.239	0.422	-0.963	-0.446
8.9	3.3	11	0.01	69.1	-0.371	-0.679	1.382	-1.527	0.254
9.3	5.1	7.8	0.51	62.9	-0.246	-0.069	-0.981	-0.184	-0.263
7.6	1.7	9.8	0.38	53.4	-0.775	-1.221	0.496	-0.533	-1.055
10.1	5.4	8.6	0.72	57.9	0.003	0.032	-0.390	0.380	-0.680
9.1	4.8	11.8	0.21	86.4	-0.308	-0.171	1.973	-0.990	1.695
12.4	8	9.2	0.51	72.4	0.719	0.913	0.053	-0.184	0.529
12.5	8.1	7.4	1.25	73.2	0.750	0.947	-1.277	1.803	0.595
12	7.4	8.2	0.19	58.8	0.594	0.710	-0.686	-1.044	-0.605
6.8	1.4	9.6	0.58	60.1	-1.024	-1.323	0.348	0.004	-0.496
6.8	0.7	9.4	0.84	55.6	-1.024	-1.560	0.200	0.702	-0.871
6.7	1.7	10	0.48	50.5	-1.055	-1.221	0.644	-0.265	-1.296
8.3	4.6	9.5	0.85	64.6	-0.557	-0.239	0.274	0.729	-0.121
8.1	4.8	8.5	0.54	59.3	-0.619	-0.171	*-0.464*	-0.104	-0.563
4.8	2.7	6.9	0.57	50.9	-1.647	-0.883	-1.646	-0.023	-1.263
12.9	9.1	7.5	0.57	76.9	0.874	1.286	-1.203	-0.023	0.904
7.4	4.6	8.3	0.85	58.2	-0.837	-0.239	-0.612	0.729	-0.655
13.2	6.5	11.4	1.44	70.7	0.968	0.405	1.678	2.313	0.387
19	8.2	10.5	1.13	89.6	*2.773*	0.981	1.013	1.481	1.962

```
     Col.:   1 | Mean:  1.0090e+01 | Variance: 1.0325e+01
     Col.:   2 | Mean:  5.3048e+00 | Variance: 8.7100e+00
     Col.:   3 | Mean:  9.1286e+00 | Variance: 1.8335e+00
     Col.:   4 | Mean:  5.7857e-01 | Variance: 1.3865e-01
     Col.:   5 | Mean:  6.6057e+01 | Variance: 1.4400e+02

---------------------- Correlation matrix -------------------
[  1] 1000
[  2]  814 1000
[  3]   47 -355 1000
[  4]  245   76   -2 1000
[  5]  769  762  131   77 1000
------------------------------------------------------------
```

Le pourcentage de variance expliquée par la régression multiple du nombre de voix sur chacune des combinaisons d'explicatives est représentée dans la figure :



Les nuages normalisés (explicatives - expliquée) sont :



Commenter.

4. Paris

ISFA 2° année

Les données analysées sont publiées dans les deux articles du *Monde* reproduits ci-dessous. Le tableau figurant dans l'article du 18 mars comporte 20 lignes (Arrondissements) et 7 colonnes (catégories des candidats). Il est utilisé dans le format ci-dessous (fichier MuniPC) :

0.4132	0.3474	0.1169	0.0318	0.0395	0.0395	0.0117
0.3247	0.4366	0.1037	0.0433	0.0329	0.0395	0.0193
0.3106	0.4628	0.0820	0.0476	0.0426	0.0339	0.0205
0.3629	0.4104	0.0912	0.0386	0.0359	0.0398	0.0212
0.3982	0.3871	0.0819	0.0377	0.0415	0.0342	0.0194
0.4580	0.3313	0.0882	0.0230	0.0390	0.0438	0.0167
0.5857	0.2005	0.1114	0.0146	0.0290	0.0449	0.0139
0.5918	0.1782	0.1318	0.0165	0.0265	0.0430	0.0122
0.3960	0.3570	0.1114	0.0403	0.0349	0.0395	0.0209
0.2988	0.4214	0.1261	0.0583	0.0424	0.0281	0.0249
0.2853	0.4531	0.1082	0.0612	0.0406	0.0291	0.0225
0.3737	0.3727	0.1181	0.0426	0.0396	0.0367	0.0166
0.3111	0.4355	0.1100	0.0518	0.0414	0.0297	0.0205
0.3652	0.3934	0.1041	0.0423	0.0425	0.0343	0.0182
0.4884	0.2918	0.1031	0.0278	0.0372	0.0372	0.0145
0.6447	0.1464	0.1208	0.0119	0.0225	0.0423	0.0114
0.4971	0.2530	0.1285	0.0290	0.0372	0.0398	0.0154
0.2987	0.4022	0.1485	0.0602	0.0411	0.0279	0.0214
0.2779	0.4436	0.1412	0.0523	0.0341	0.0275	0.0234
0.2702	0.4385	0.1388	0.0634	0.0393	0.0264	0.0234

Ce tableau est soumis à une Analyse en Composantes Principales (*PCA: Covariance matrix PCA*) centrée.

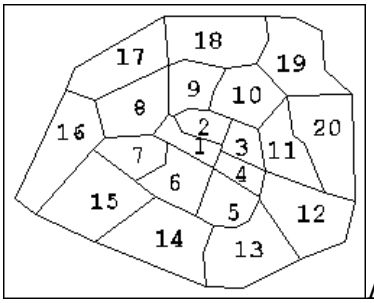
```
Centered Principal Component Analysis (Pearson 1901)
Input file: E:\Ade4\ELEPARIS\MuniPC
---- Row weight:
File E:\Ade4\ELEPARIS\MuniPC.cppl contains the row weight
It has 20 rows and 1 column
Each row has 5.0000e-02 weight (Sum = 1)
---- Column weights:
File E:\Ade4\ELEPARIS\MuniPC.cppc contains the column weights
It has 7 rows and 1 column
Each column has unit weight (Sum = 7)
---- Table:
File E:\Ade4\ELEPARIS\MuniPC.cpta contains the (column) centred table
It has 20 rows and 7 columns
---- Info: means and variances
File E:\Ade4\ELEPARIS\MuniPC.cpma contains the descriptive of the analysis
It contains successively:
     Number of rows: 20
     Number of columns: 7
     means and variances:
     Col.:   1 | Mean: 3.9761e-01 | Variance: 1.2105e-02
     Col.:   2 | Mean: 3.5814e-01 | Variance: 8.7598e-03
     Col.:   3 | Mean: 1.1329e-01 | Variance: 3.4531e-04
     Col.:   4 | Mean: 3.9710e-02 | Variance: 2.3650e-04
     Col.:   5 | Mean: 3.6985e-02 | Variance: 2.9893e-05
     Col.:   6 | Mean: 3.5855e-02 | Variance: 3.4307e-05
     Col.:   7 | Mean: 1.8400e-02 | Variance: 1.6287e-05
-------------------------------------------------
```

```
DiagoRC: General program for two diagonal inner product analysis
Input file: E:\Ade4\ELEPARIS\MuniPC.cpta
--- Number of rows: 20, columns: 7
-----------------------
Total inertia: 0.0215269
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.0914E-02 +0.9715 +0.9715   |02   +5.7037E-04 +0.0265 +0.9980 |
03   +2.3276E-05 +0.0011 +0.9991   |04   +1.1334E-05 +0.0005 +0.9996 |
05   +5.8951E-06 +0.0003 +0.9999   |06   +2.3903E-06 +0.0001 +1.0000 |
07   +0.0000E+00 +0.0000 +1.0000

File E:\Ade4\ELEPARIS\MuniPC.cpvp contains the eigenvalues and relative inertia
for each axis
--- It has 7 rows and 2 columns

File E:\Ade4\ELEPARIS\MuniPC.cpco contains the column scores
--- It has 7 rows and 2 columns

File E:\Ade4\ELEPARIS\MuniPC.cpli contains the row scores
--- It has 20 rows and 2 columns
```

Par ailleurs, on calcule la matrice des corrélations du tableau MuniPC :

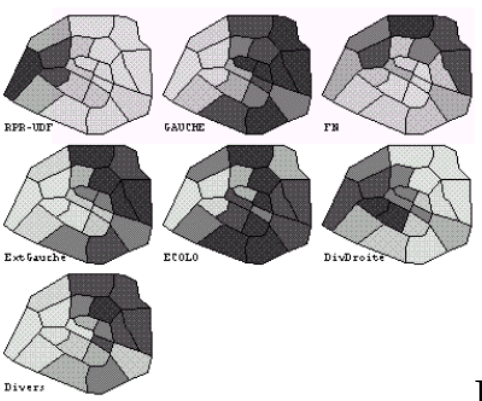
```
---------------------- Correlation matrix -------------------
[  1] 1000
[  2] -980 1000
[  3]  -47 -149 1000
[  4] -954  890  231 1000
[  5] -747  764 -225  693 1000
[  6]  820 -736 -355 -916 -609 1000
[  7] -873  841   69  886  562 -788 1000
-------------------------------------------------------------
```

On donne la carte des arrondissements (*Areas: Area Edit*) :

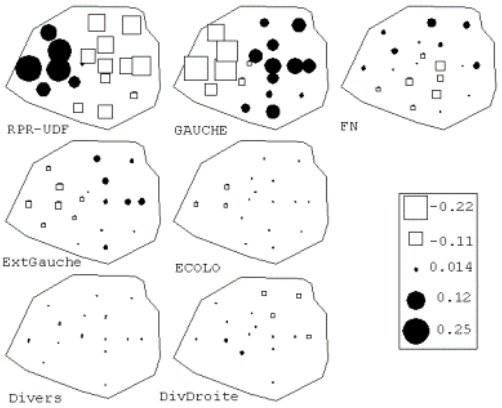


A

les cartes par niveau de gris des valeurs observées (*Areas: Gray levels areas*, échelles séparées) et les cartes par symboles des valeurs centrées (*Maps: Values*, échelle commune) :



B

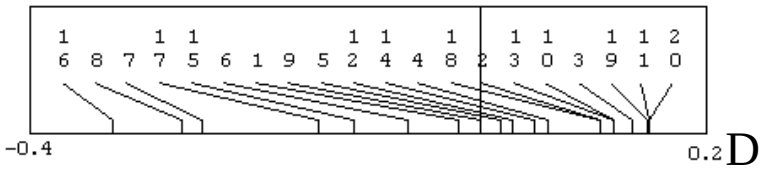


C

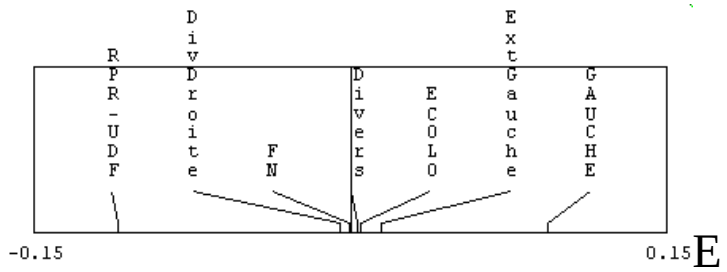
Les valeurs des coordonnées factorielles sur le premier facteur sont :

```
E:\Ade4\ELEPARIS\MuniPC.cpli
   1 | -0.0197     6 | -0.0648    11 |  0.1487    16 | -0.3270
   2 |  0.1059     7 | -0.2470    12 |  0.0278    17 | -0.1443
   3 |  0.1344     8 | -0.2659    13 |  0.1169    18 |  0.1056
   4 |  0.0598     9 |  0.0004    14 |  0.0477    19 |  0.1471
   5 |  0.0183    10 |  0.1179    15 | -0.1127    20 |  0.1510

E:\Ade4\ELEPARIS\MuniPC.cpco
   1 | -0.1097   3 | -0.0006     5 |  0.0041     7 |  0.0035
   2 |  0.0929   4 |  0.0144     6 | -0.0046
```

On obtient les figures qui suivent avec *Graph1D: Labels* :

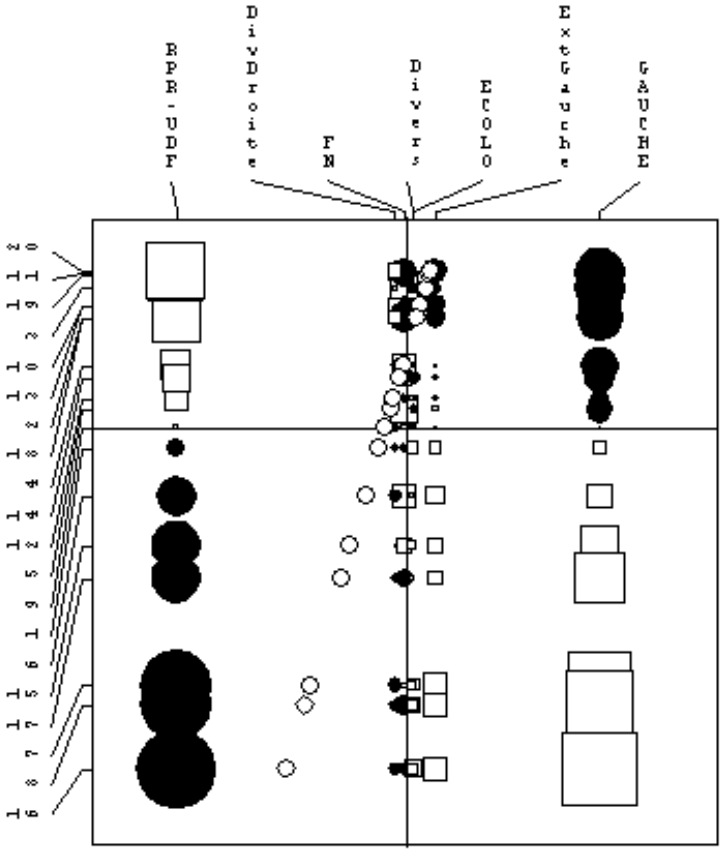


D



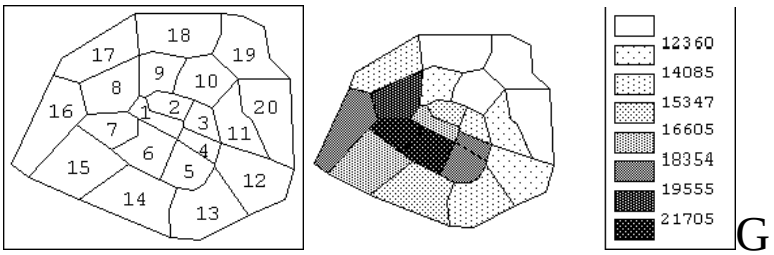
E

avec *Tables: Values* et *Tables: TabMeanVar* :



F

Les données de l'article du 4 mai se résume à un vecteur de 20 valeurs appelé Prix. Les valeurs sont cartographiables :



G

On calcule simplement la corrélation existant entre le score de chacune des catégories de candidat et la variable prix :

```
[  1]  7.5803e-01    [  5] -4.8356e-01
[  2] -6.4404e-01    [  6]  8.4580e-01
[  3] -5.0549e-01    [  7] -6.9320e-01
[  4] -8.6103e-01
```

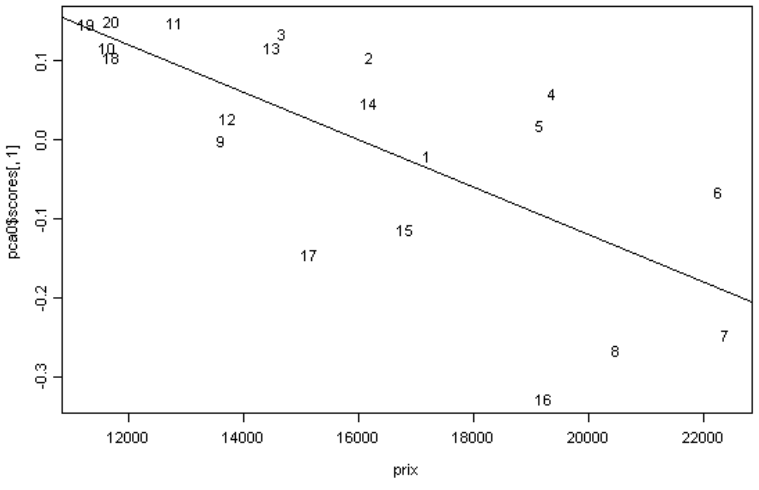
On calcule de même la corrélation existant entre la première composante principale et la variable Prix :

```
[  1] -7.1597e-01
```

Les données sont importées dans un autre logiciel de statistique :

```
> munipc
   RPR-UDF GAUCHE     FN ExtGauche  ECOLO DivDroite Divers
 1  0.4132 0.3474 0.1169    0.0318 0.0395    0.0395 0.0117
 2  0.3247 0.4366 0.1037    0.0433 0.0329    0.0395 0.0193
```

```
 3  0.3106 0.4628 0.0820    0.0476 0.0426    0.0339 0.0205
...
18  0.2987 0.4022 0.1485    0.0602 0.0411    0.0279 0.0214
19  0.2779 0.4436 0.1412    0.0523 0.0341    0.0275 0.0234
20  0.2702 0.4385 0.1388    0.0634 0.0393    0.0264 0.0234

> cor(munipc,prix)
             [,1]
  RPR-UDF  0.7580      ECOLO -0.4836
   GAUCHE -0.6440  DivDroite  0.8458
       FN -0.5055     Divers -0.6932
ExtGauche -0.8610
> pca0_princomp(munipc, cor=F)
> cor(pca0$scores[,1],prix)
[1] -0.71597
> plot(prix,pca0$scores[,1],type="n")
> text(prix,pca0$scores[,1])
> abline(lm(pca0$scores[,1]~prix))
```



H

```
> pca0$sdev
 Comp. 1 Comp. 2  Comp. 3  Comp. 4  Comp. 5  Comp. 6 Comp. 7
  0.1446 0.02388 0.004825 0.003367 0.002428 0.001546       0
```

Partie 1

Q1 Vrai ou Faux ? Dans une ACP normée normée l'inertie totale est égale au nombre de lignes du tableau traité.

Q2 Vrai ou Faux ? Dans une ACP centrée sur un tableau de pourcentage par ligne, la somme des colonnes du tableau centré est nulle.

Q3 Vrai ou Faux ? Une matrice de covariance est carrée, symétrique et diagonalisable. Toutes ses valeurs propres sont strictement positives.

Q4 Vrai ou Faux ? Dans une ACP centré sur un tableau de pourcentage par ligne, la somme des coordonnées des colonnes sur une composante principale est nulle.

Q5 Sur le graphe F on a ajouté par un cercle blanc le position moyenne des position des colonnes en utilisant la distribution de fréquences de chaque ligne. Démontrer que ces ces points sont sur une droite.

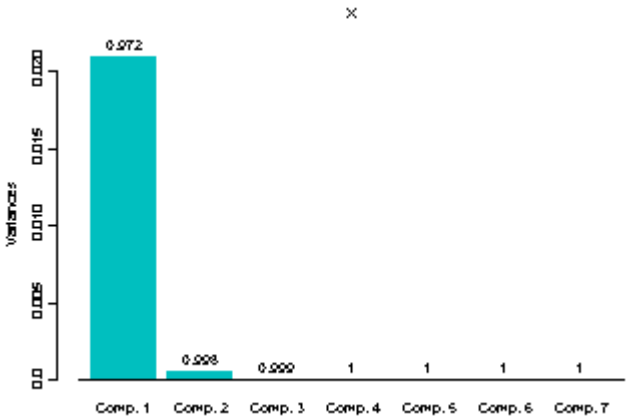
Partie 2

On s'intéresse à l'ACP de MuniPC.

Q1 Qu'est-ce qui justifie qu'on dépouille de manière isolée le premier facteur de l'ACP ?

Q2 Vrai ou Faux ? Deux variables prennent en compte plus de 95 % de l'inertie totale.

Q3 A-ton dans ADE-4 l'équivalent de l'ordre `plot(pca0)` qui donne dans S-PLUS le dessin ci -dessous ?



Q4 Vrai ou Faux ? Sur le graphe F on a réécrit le tableau centré en positionnant les lignes et les colonnes par les coordonnées factorielles du premier facteur. Les cercles noirs sont les valeurs négatives et les carrés blancs sont les valeurs positives.

Q5 Que fournit l'édition suivante ?

```
> pca0$sdev
 Comp. 1 Comp. 2  Comp. 3  Comp. 4  Comp. 5  Comp. 6 Comp. 7
  0.1446 0.02388 0.004825 0.003367 0.002428 0.001546       0
```

Q6 Que fournit l'édition suivante ?

```
Column inertia
All contributions are in 1/10000

|Num  |Fac 1|Fac 2|
|    1| 5749| 1391|
|    2| 4127| 2186|
|    3|    0| 5968|
|    4|   98|  336|
|    5|    8|   13|
|    6|   10|   98|
|    7|    5|    4|
```

Partie 3

On s'intéresse maintenant à la signification des données.

Q1 En utilisant ce qui suit, indiquer la signification du facteur 2 de l'ACP centrée de MuniPC.

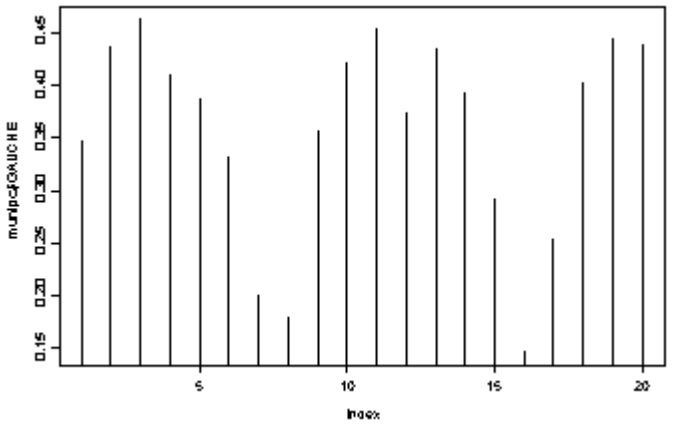
```
----------Relative contributions----------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
```

1	9933	65	1	9999	5623
2	9854	142	3	9999	4069
3	10	9858	130	9999	160
4	8718	811	470	9999	109
5	5746	261	3991	9999	13
6	6261	1633	2104	9999	15
7	7483	152	2363	9999	7

Q2 Commenter l'assertion « L'opposition Droite-Gauche dans l'opinion parisienne est fortement liée aux caractéristiques socio-économiques des quartiers ».

Q2 Commenter l'assertion «La catégorie Divers proposée par le Monde aurait pu s'appeler Divers Gauche ».

Q3 Quand on représente le pourcentage de voix de gauche en fonction du numéro du l'arrondissement on obtient une courbe périodique. Pourquoi ?



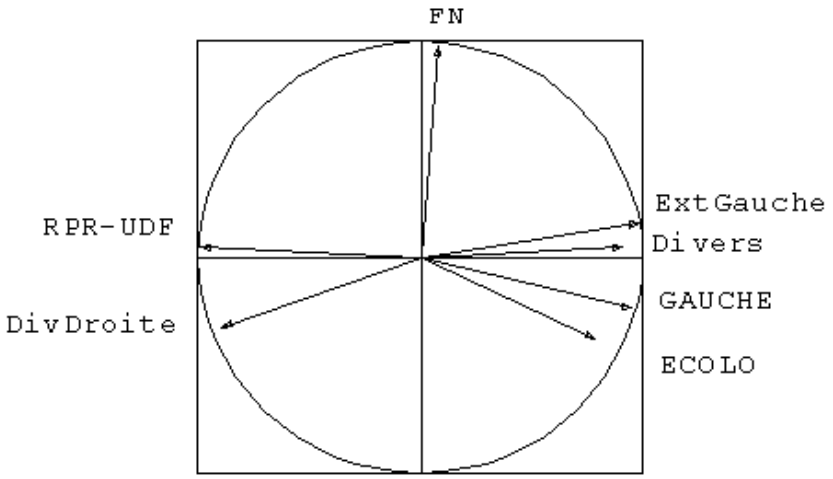
Q4 Commenter l'assertion « Les arrondissements 4 et 5 sont des quartiers "chics" qui votent plus à droite que prévu ».

Q5 Commenter l'assertion « Le commentaire de l'article du 18 mars s'appuie largement sur le tableau de données».

Partie 4

On utilise l'ACP normée de MuniCP.

Q1 En utilisant l'information ci-dessous représenter la variable prix sur la figure et justifier l'opération.




```
X input file: E:\Ade4\ELEPARIS\MuniPC.cnli
--- Number of rows: 20, columns: 2
Y input file: E:\Ade4\ELEPARIS\Prix
--- Number of rows: 20, columns: 1
Diagonal inner product: uniform weight
XtDY output file: screen
--- Number of rows: 2, columns: 1
Input file: screen
--- Number of rows: 2, columns: 1
-------------------
[  1] -7.9164e-01
[  2] -4.4079e-01
-------------------
```

Q2 Donner une légende pour la figure :