Université Panthéon-Assas

SESSION : Mai 2019.
ANNEE D'ETUDE : Master 1 Economie Managériale et Industrielle
MATIERE : ANALYSE DE DONNEES Enseignant : Mr FAKHFAKH

Exercice N°1

Une enquête a été effectuée auprès de 56 marques d'eau. Nous disposons ainsi des principales caractéristiques de chaque marque en terme de teneur (en mg/litre) des composants suivants :

CA : calcium. MG : magnésium. NA : sodium. K : potassium.
SUL : sulfates. NO3 : nitrates. HCO3 : carbonates. CL : chlorures.

Un premier résumé de ces caractéristiques est donné dans le tableau suivant :

	Moyenne	Ecart Type	Matrice de corrélation							
			CA	MG	NA	K	SUL	NO3	HCO3	CL
CA	102,449	120,656	1,000	0,717	0,092	0,037	0,918	-0,015	0,109	0,264
MG	25,052	27,040	0,717	1,000	0,551	0,070	0,652	-0,108	0,561	0,420
NA	83,896	184,108	0,092	0,551	1,000	-0,009	0,079	-0,110	0,831	0,533
K	48,156	292,690	0,037	0,070	-0,009	1,000	0,099	0,754	-0,038	-0,028
SUL	143,525	333,325	0,918	0,652	0,079	0,099	1,000	-0,029	-0,068	0,342
NO3	4,992	10,356	-0,015	-0,108	-0,110	0,754	-0,029	1,000	-0,095	-0,101
HCO3	407,956	567,256	0,109	0,561	0,831	-0,038	-0,068	-0,095	1,000	0,080
CL	46,507	137,193	0,264	0,420	0,533	-0,028	0,342	-0,101	0,080	1,000

Les deux premiers vecteurs propres sont les suivants :

	CA	MG	NA	K	SUL	NO3	HCO3	CL
Prin1	0,431	0,527	0,384	0,002	0,409	-0,085	0,316	0,337
Prin2	0,332	0,036	-0,362	0,485	0,381	0,479	-0,380	-0,058

Matrice de corrélation des variables avec les 4 premières composantes principales

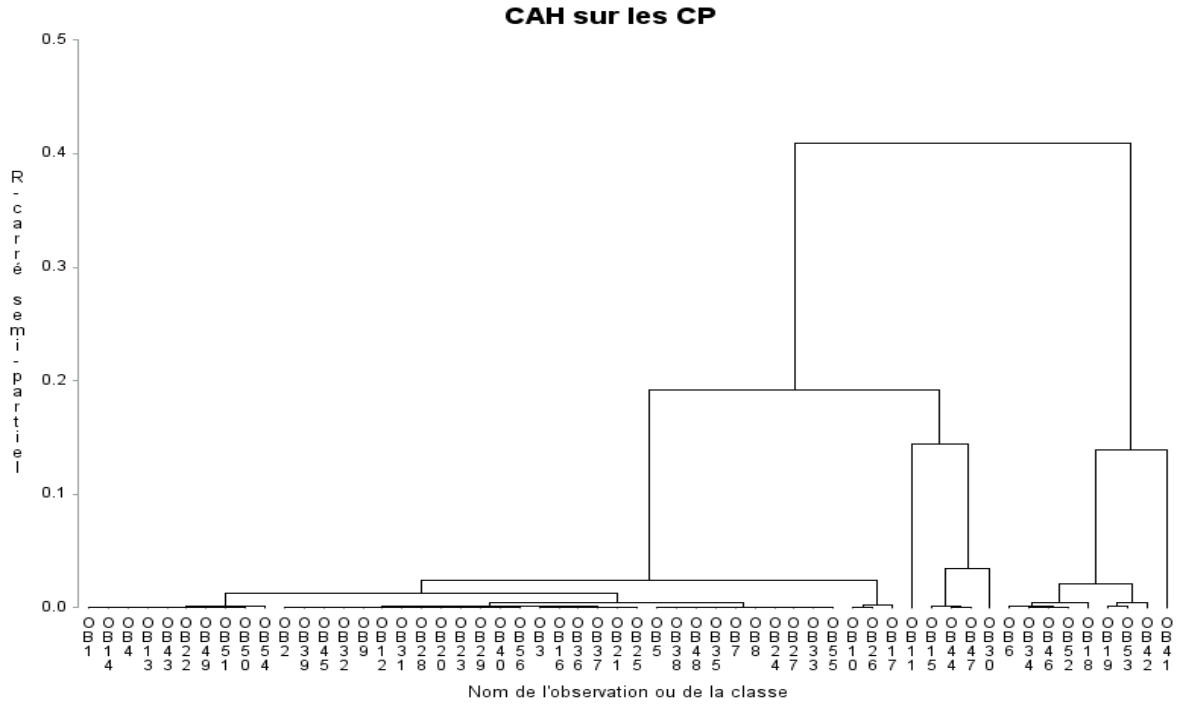
	Prin1	Prin2	Prin3	Prin4
CA	0,762	0,457	-0,350	-0,209
MG	0,931	0,050	0,034	-0,176
NA	0,679	-0,497	0,482	0,134
K	0,004	0,668	0,661	0,041
SUL	0,722	0,524	-0,393	-0,030
NO3	-0,150	0,659	0,650	0,035
HCO3	0,558	-0,522	0,510	-0,381
CL	0,595	-0,080	0,002	0,793

1) Examiner les moyennes et les écrat-types. Quels sont les critères qui paraissent les plus importants.
2) a. A partir de la matrice de corrélation, représenter le dendogramme des variables. Interpréter.
b. Est-il indispensable d'effectuer l'analyse sur les variables centrées et réduites.
3) Les quatre premières valeurs propres obtenues dans cet exemple, classées par ordre décroissant, sont les suivantes : (3.117), (1.893), (1.630) et (0.870).
Déterminer la valeur de l'inertie totale et donner la contribution de chaque axe à l'inertie totale. Interpréter.
4) a. Quelles sont les variables qui contribuent le plus à la définition de la première composante principale ? Que traduit cette première composante principale.
b. Même question que (a.) pour les axes 2 et 3.
c. Quel est le nombre d'axes que l'on peut retenir? Justifier votre réponse.
5) Représenter le cercle de corrélation (pour les deux premiers axes). L'interpréter.

6) Nous avons décidé d'établir une classification ascendante hiérarchique sur les individus en fonction des mêmes variables.

a. Un résumé sur les derniers nœuds de la CAH est donné dans le tableau suivant, ainsi que l'arbre correspondant à cette classification. Donner le nombre de classes à retenir. Justifier votre réponse.

Classes	Ainé	Bénjamin	Perte Inertie
8	CL14	CL10	.964
7	CL9	CL11	.942
6	CL8	CL12	.918
5	CL17	OB30	.884
4	CL7	OB41	.745
3	OB11	CL5	.601
2	CL6	CL3	.409
1	CL2	CL4	.000



b. Nous avons décidé de retenir une typologie à 5 classes dont les caractéristiques sont données dans le tableau suivant :

	Moyennes des variables par cluster								
Cluster	Nb d'indiv	**CA**	**MG**	**NA**	**K**	**SUL**	**NO3**	**HCO3**	**CL**
1	6,0	108,9	55,9	351,8	38,9	64,2	3,3	1392,4	74,4
2	1,0	130,0	31,0	0,0	2195,0	387,0	63,0	10,2	1,4
3	4,0	471,3	78,0	168,0	7,6	1238,8	1,2	281,3	250,2
4	1,0	99,0	88,1	968,0	103,0	18,0	1,0	3380,5	88,0
5	44,0	67,5	14,5	21,5	3,1	52,1	4,3	226,7	24,3

Interpréter chacune des classes. Discuter de la pertinence de cette analyse à la lumière de ces résultats. Comment peut-on améliorer la qualité de cette analyse.

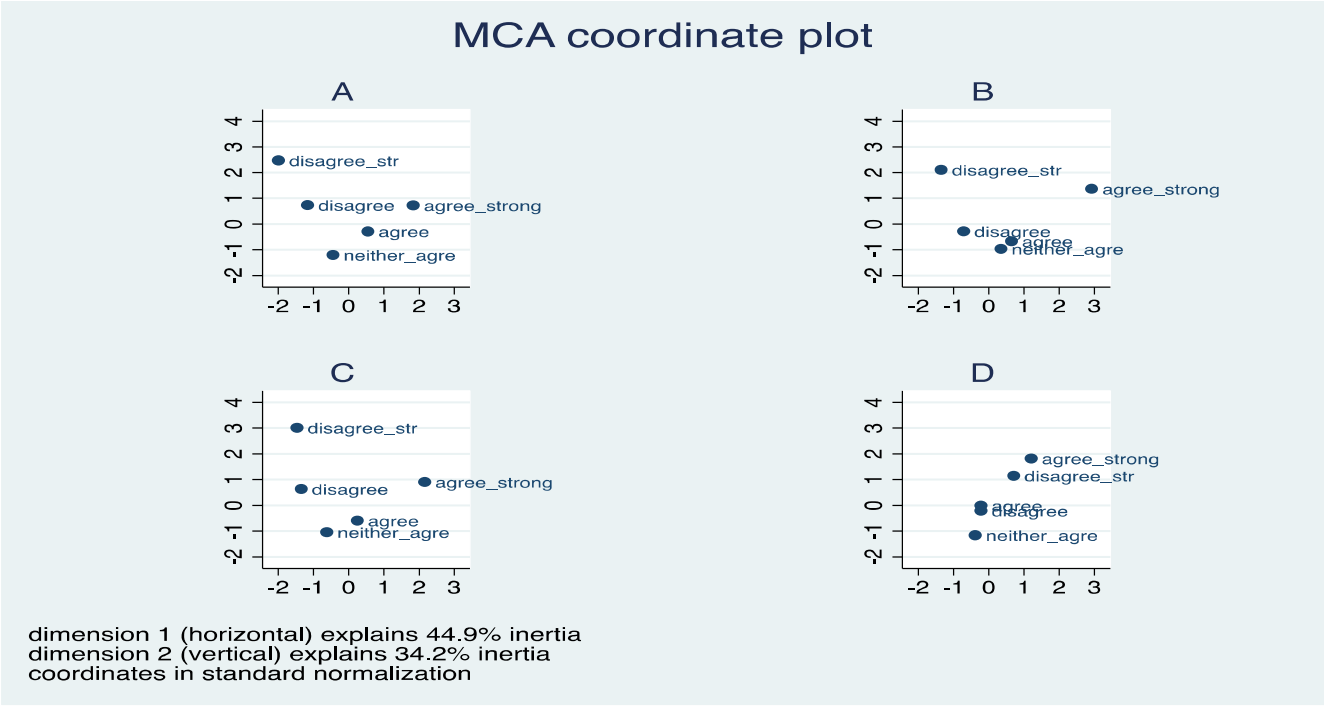
Exercice N°2

Nous cherchons dans cette application à cerner les attitudes des individus face aux effets de la science. Plus précisément, nous nous intéressons aux quatre effets suivants (les choix de réponse étant tout à fait d'accord : strongly agree ; d'accord : agree ; neutre : neither (agree or disagree): pas d'accord : disagree ; pas du tout d'accord : strongly disagree)

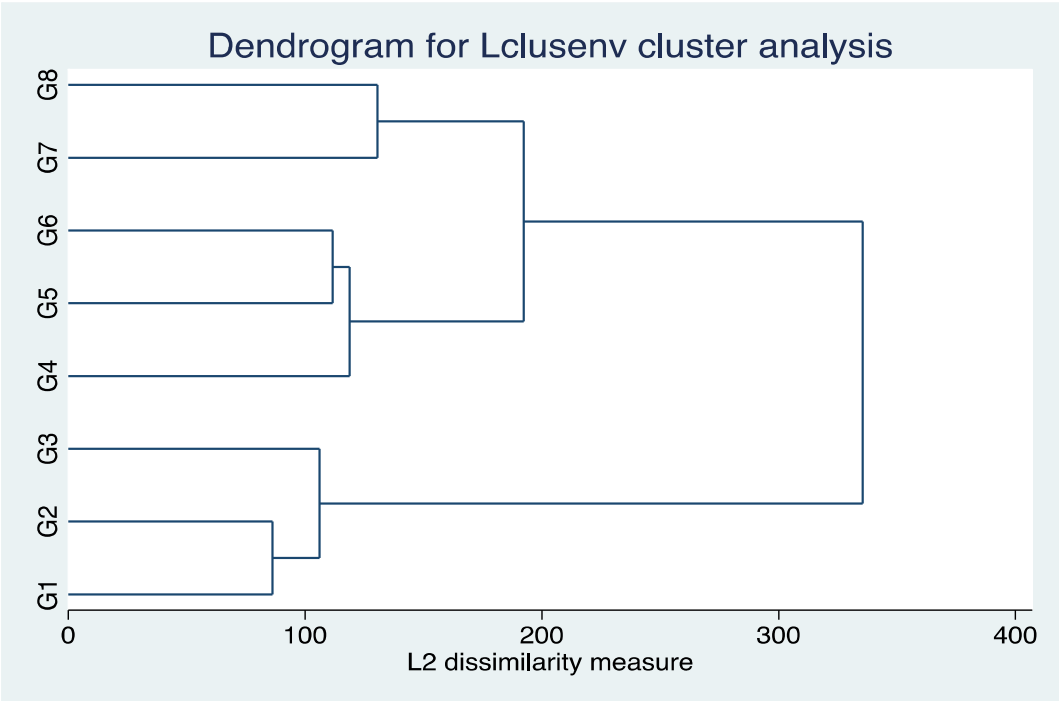
- Trop de sciences mais pas assez de sentiment et de foi (A)
- La science nuit plus qu'elle ne fait du bien (B)
- Tout changement nuit à la nature (C)

- La science va résoudre les problèmes environnementaux (D)

1) Nous avons commencé l'analyse pour une ACM. Les résultats concernant les pourcentage d'inertie expliquée par les trois premières composantes principales sont les suivants :44.9 %; 34.2% et 5.4%. Quel est le nombre d'axes à retenir. Justifier votre réponse.
2) Quelle est la valeur de l'inertie totale et quel est le nombre maximum d'axes qu'on peut obtenir.
3) Nous n'avons pu obtenir que 6 composantes principales. Interpréter ce constat.
4) Le graphique suivant donne la représentation des modalités des 4 variables sur le premier plan factoriel. Interpréter ces résultats : que signifie chacune des deux premières composantes principales ?



5) Nous avons ensuite effectué une CAH. L'arbre associé à cette classification (sur les individus) est le suivant :



a- Sur quelles variables devrait-on effectuer cette classification.
b- Quel est le nombre d'axes à retenir. Justifier votre réponse.

2) Nous avons ensuite décidé de retenir 3 classes. Le tableau suivant donne la distribution des différentes variables entre les classes.

		G1, N=266	G2, N=402	G3, N=203	Total
A	agree strongly	6,02	13,18	24,63	13,66
Trop de sciences mais pas assez de sentiment et de foi	agree	22,93	42,29	44,83	36,97
	neither	16,92	33,83	11,33	23,42
	disagree	40,98	9,45	15,27	20,44
	disagree strongly	13,16	1,24	3,94	5,51
B	agree strongly	0	10,2	14,78	8,15
La science nuit plus qu'elle ne fait du bien	agree	0,38	32,59	20,69	19,98
	neither	9,4	36,07	17,24	23,54
	disagree	51,88	16,42	37,93	32,26
	disagree strongly	38,35	4,73	9,36	16,07
C	agree strongly	0,75	14,68	44,83	17,45
Tout changement nuit à la nature	agree	5,26	47,26	55,17	36,28
	neither	26,69	31,34	0	22,62
	disagree	49,25	5,72	0	17,68
	disagree strongly	18,05	1	0	5,97
D	agree strongly	5,64	11,19	0	6,89
La science va résoudre les problèmes environnementaux	agree	33,46	35,57	0	26,64
	neither	15,04	37,06	6,4	23,19
	disagree	29,32	11,94	49,26	25,95
	disagree strongly	16,54	4,23	44,33	17,34

a- Interpréter chacune des classes.
b- Nous disposons de l'age des individus (de 18 ans et plus) et de leur genre. Comment serai-il possible d'expliquer l'appartenance d'un individu à la classe 3 (par exemple). Donner brièvement la variable endogène et expliquer la méthode d'estimation.
c- Pourquoi il ne serait pas adéquat d'appliquer les MCO ?

3) Les résultats de cette estimation par un logit sont brièvement résumés comme suit (**coefficients associés au variables**)

fem .2100507 ; **age :** 25-34 : .0971647 ; 35-44 : .0794022 ; 45-54 : -.112562 ; 55-64 : -.2595076 , 65et+ : .4848832 .

Interpréter l'ensemble de ces résultats brièvement.

4) Nous cherchons à établir un score d'appartenance à cette classe.
a- Rappeler le principe du scoring par la régression logistique.
b- Appliquer ce principe à cet exemple et donner la grille de scoring associée.

Questions de cours

1. Peut-on obtenir une (ou des) valeur propre négative dans les cas suivants :
   a. Le cas d'une ACP
   b. Le cas d'une AFC
   c. Le cas d'une ACM

2. Nous considérons un échantillon de N individus caractérisés par « m » variables. Chaque variable « j » de ces « m » variables est composée de Pj modalités. Soit P le nombre total de modalités.
   a. Quelle est l'expression de l'inertie totale ?
   b. On suppose que toutes les variables sont binaires. Quelle serait la valeur de l'inertie totale.