Université Panthéon-Assas

SESSION : Septembre 2019.
ANNEE D'ETUDE : Master 1 Economie Managériale et Industrielle
MATIERE : DATA MINING Enseignant : Mr FAKHFAKH

Exercice N°1

Nous cherchons dans cet exemple à caractériser les entreprises en fonction des variables quantitatives présentées dans le tableau suivant (en l'absence des salaires, nous retenons les charges salariales) :

	Div : Dividendes par salarié	Marge : marge par salariés	Prod : valeur ajoutée par salarié	K : capital par salarié	Pi : profit par salarié	Amort : amortiss-ement/capital	Part : Bonus par salarié	Charge : charges salariales par employé
Moy.	3,73	23,20	58,52	55,51	12,00	0,31	0,09	12,06
StD	9,865	41,055	26,328	43,322	14,915	0,140	0,534	6,021

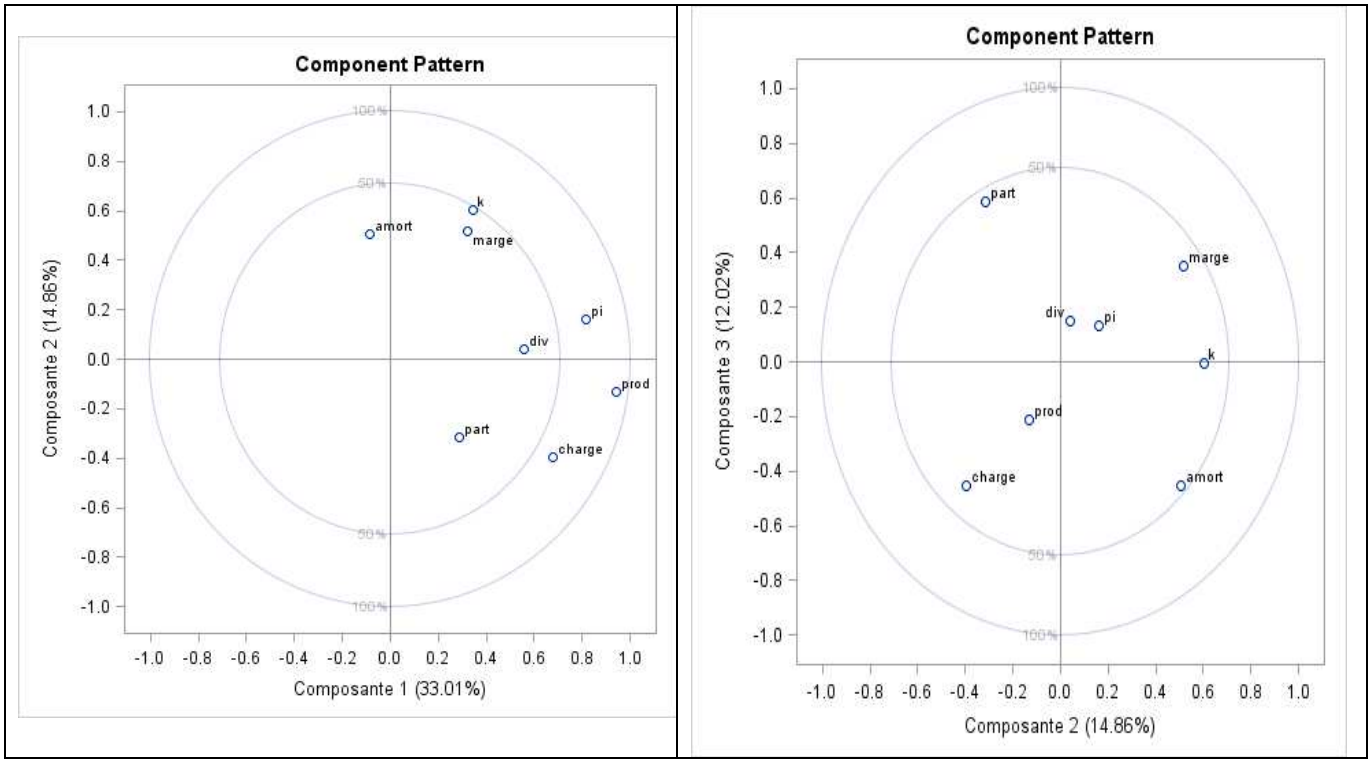
Source : Fare 2008, Part désigne le bonus de la participation aux fruits de l'expansion. StD désigne l'écart-type.

1- Interpréter ce tableau : quelles seraient les variables les plus pertinentes à l'analyse ?
2- Nous proposons de poursuivre l'analyse grâce à la matrice de corrélation suivante :

	div	marge	prod	k	pi	amort	part	charge
marge	0,088	1,000						
prod	0,376	0,194	1,000					
k	0,087	0,166	0,234	1,000				
pi	0,429	0,241	0,749	0,268	1,000			
amort	-0,002	0,014	-0,073	0,063	-0,039	1,000		
part	0,116	0,022	0,181	-0,007	0,178	-0,054		
charge	0,178	0,049	0,773	0,052	0,251	-0,075	0,124	1,000

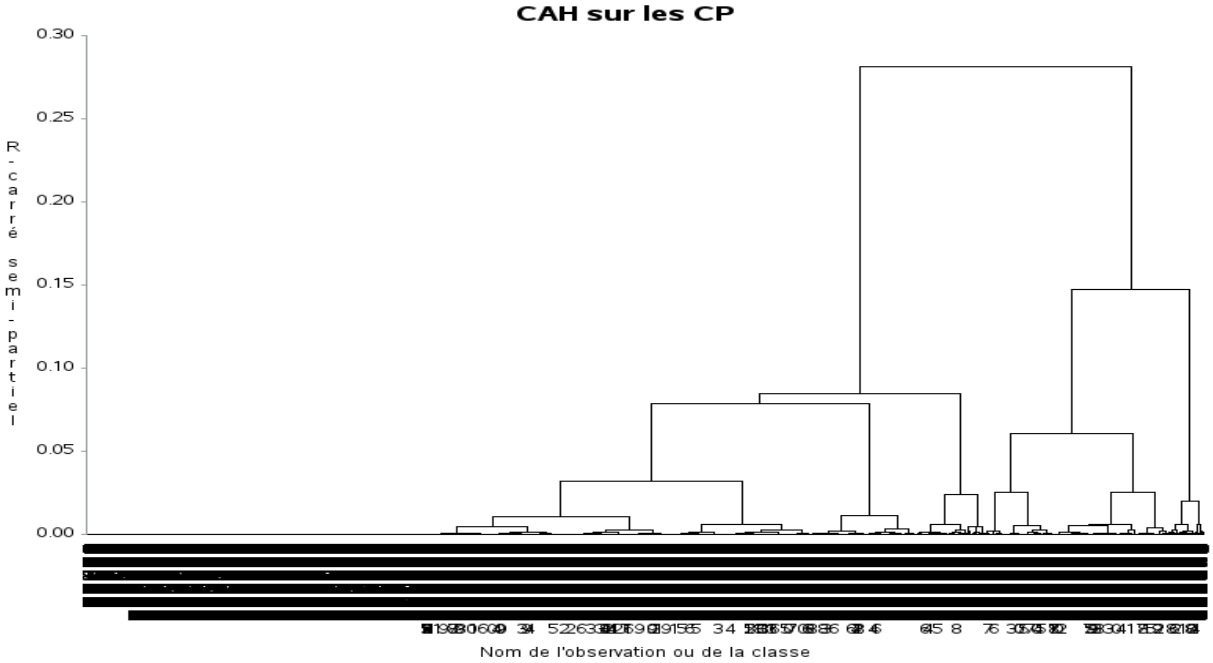
Quelles sont les premières « proximités » que nous pouvons observer entre les variables ? Représenter le dendrogramme des variables. Commenter.

3- Nous avons ensuite effectué une ACP sur l'ensemble de ces variables. Les cinq premières valeurs propres sont : 2.641 ; 1.189 ; 0.962 ; 0.956 et 0.861
   a- Donner la part d'inertie expliquée par chacun de ces axes. Interpréter.
   b- Quel est le nombre maximal que l'on peut retenir. Justifier votre réponse.
   c- Les deux graphiques suivants représentent les deux premiers cercles de corrélations (des variables avec les composantes principales. Interpréter chacun des cercles et donner une interprétation générale aux trois premières composantes principales.



4- Nous avons ensuite effectué une Classification ascendante hiérarchique (selon Ward). Les résultats concernant les derniers nœuds ainsi que l'arbre correspondant à l'analyse sont les suivants :

Number	Clusters Joined		Semipartial	R-Square
9	**CL16**	**CL20**	0.0240	.735
8	**CL13**	**CL23**	0.0251	.710
7	**CL31**	**CL19**	0.0254	.684
6	**CL12**	**CL15**	0.0320	.652
5	**CL7**	**CL8**	0.0607	.592
4	**CL6**	**CL11**	0.0787	.513
3	**CL4**	**CL9**	0.0847	.428
2	**CL5**	**CL10**	0.1475	.281
1	**CL3**	**CL2**	0.2808	.000



a- Quel est le nombre de classes à retenir ? Justifier votre réponse.

b- Rappeler le critère de Ward et décrire la démarche à suivre pour établir une telle typologie. Comment détermine-t-on le nombre de classes à retenir

c- Nous avons choisi de retenir une typologie à 6 classes. Les statistiques par classes sont données dans le tableau suivant. Quelles sont les variables pertinentes à l'analyse. Donner une interprétation générale à chacune des classes, ainsi qu'à l'ensemble des classes.

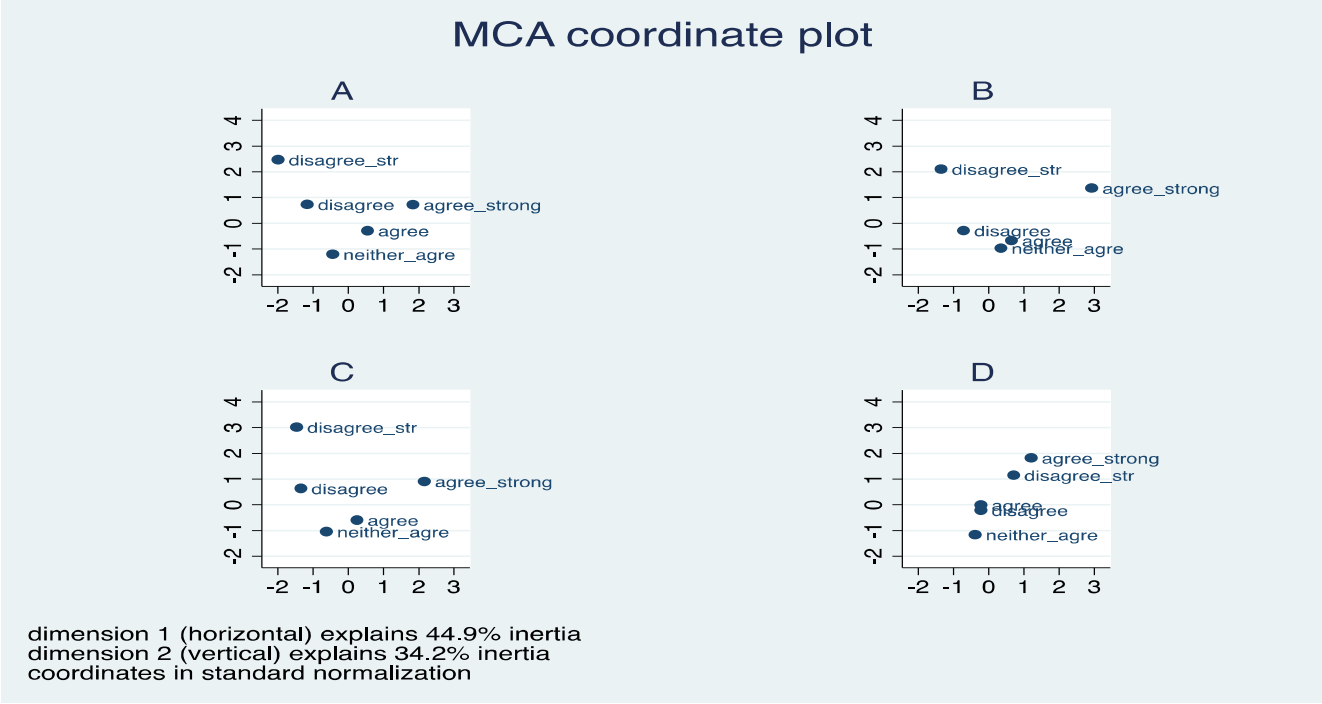
Cluster	N	div	marge	prod	k	pi	amort	part	charge
1	134	13,46	185,74	114,29	41,41	45,45	0,28	0,36	18,88
2	7095	1,98	10,74	55,80	28,14	10,87	0,31	0,08	12,84
3	607	4,30	100,02	68,51	100,43	18,45	0,30	0,07	12,76
4	875	4,40	7,88	70,72	132,00	18,53	0,31	0,09	13,14
5	1389	2,38	0,43	125,08	4,15	59,21	0,26	0,09	25,25
6	26	79,45	3,92	102,43	108,46	41,79	0,26	0,01	16,26
Total	10126	3,73	23,20	58,52	55,51	12,00	0,31	0,09	12,06

Exercice N°2

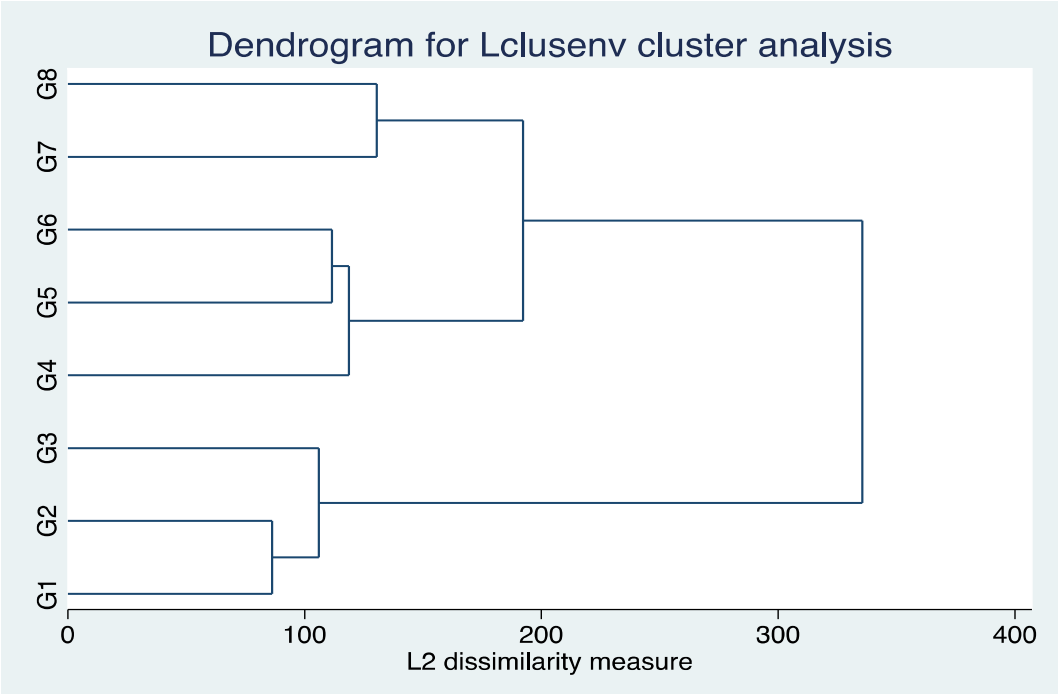
Nous cherchons dans cette application à cerner les attitudes des individus face aux effets de la science. Plus précisément, nous nous intéressons aux quatre effets suivants (les choix de réponse étant tout à fait d'accord : strongly agree ; d'accord : agree ; neutre : neither (agree or disagree): pas 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 pourcentages 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 de classes 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'âge des individus (de 18 ans et plus) et de leur genre. Comment serai-il possible d'expliquer l'appartenance d'un individu à une classe (la 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,** tous les coefficients sont significatifs)

fem 0.210 ; **age :** 25-34 : 0.097 ; 35-44 : 0.079 ; 45-54 : -0.113 ; 55-64 : -0.259 ; 65et+ : 0.485 .

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 un score négatif en appliquant le scoring par la régression ? Justifier votre réponse.
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.