Université Panthéon-Assas

SESSION : Mai 2021.
ANNEE D'ETUDE : Master 1 Stratégies de l'Entreprise et Economie des Organisations
MATIERE : Data Mining Enseignant : Mr FAKHFAKH
Aucun document autorisé. Calculatrices autorisées. Durée : 2h

Exercice N°1

Nous cherchons à étudier deux types d'outils d'implication des salariés dans l'entreprise : une implication « light » sous forme d'information et une participation à la prise de décision.
Nous disposons pour cela de l'enquête REPONSE (2017), et plus particulièrement le volet concernant les **salariés**. Le tableau suivant donne un descriptif des variables utilisées.

Variable : (modalités : 1 Oui, 2 Non)	N	Moyenne
INFOSAL : bien informé sur les salaires dans l'entreprise	27074	1.38
INFOTW : bien informé sur le temps de travail	27040	1.24
INFON : bien informé sur la situation de l'emploi	26144	1.40
INFOCW : bien informé sur les conditions de travail	26709	1.31
INFOFORM : bien informé sur les possibilités de formation	26789	1.38
PARTIC_POLSAL : participe à la politique salariale	24910	1.76
PARTIC_ORGA : participe à l'organisation du travail	25603	1.51
PARTIC_CONDW : participe à l'élaboration des conditions de travail	25221	1.40
PARTIC_FORMPRO : participe à la politique de formation	24720	1.46

1- Quelles seraient les modalités susceptibles d'influencer les résultats de l'analyse ? Justifier votre réponse.
2- Nous avons procédé à une analyse des correspondances multiples. Les résultats issus de l'analyse de profil des variables montrent que les quatre premières valeurs propres sont respectivement de : 0,455 ; 0,152 ; 0,086 et 0,070.
   a- Quelles tes la valeur de l'inertie totale.
   b- Quel est le nombre d'axes à retenir. Justifier votre réponse.
3- Nous avons ensuite procédé à une CAH sur les CP issues de l'analyse du nuage des profils des lignes. Les derniers regroupements sont données dans le tableau suivant :

Nbr classes	Classes jointes		Fréq	R carré semi-partiel	R carré Lien
10	**CL433**	**CL431**	2845	0.0137	.639
9	**CL22**	**CL19**	2186	0.0158	.623
8	**CL13**	**CL430**	4234	0.0161	.607
7	**CL15**	**CL9**	3850	0.0218	.585
6	**CL7**	**CL10**	6695	0.0317	.553
5	**CL14**	**CL18**	5134	0.0342	.519
4	**CL11**	**CL17**	4485	0.0343	.485
3	**CL6**	**CL5**	11829	0.0917	.393
2	**CL4**	**CL8**	8719	0.1055	.288

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

4- Nous avons opté pour une typologie à quatre classes. La distribution des salariés dans les classes en fonction des variables retenues est donnée dans le tableau suivant :

	Total	CL1		CL2		CL3		CL4	
Variable	**Moyenne**	**N**	**Moyenne**	**N**	**Moyenne**	**N**	**Moyenne**	**N**	**Moyenne**
INFOSAL	1,381	5796	1,877	7818	1,133	2379	1,463	4555	1,130
INFOTW	1,246	5796	1,752	7818	1,025	2379	1,057	4555	1,082
INFON	1,399	5796	1,905	7818	1,152	2379	1,420	4555	1,168
INFOCW	1,318	5796	1,840	7818	1,134	2379	1,139	4555	1,062
INFOFORM	1,380	5796	1,806	7818	1,085	2379	1,905	4555	1,070
PARTIC_POLSAL	1,750	5796	1,952	7818	1,992	2379	1,788	4555	1,059
PARTIC_ORGA	1,517	5796	1,818	7818	1,567	2379	1,482	4555	1,067
PARTIC_CONDW	1,410	5796	1,811	7818	1,398	2379	1,224	4555	1,017
PARTIC_FORMPRO	1,467	5796	1,826	7818	1,361	2379	1,796	4555	1,022

Interpréter chacune des classes retenues : quelles sont les caractéristiques propres à chaque classe ?

Exercice N°2

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	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	58,52	55,51	12,00	0,31	0,09	12,06
StD	9,865	26,328	43,322	14,915	0,140	0,534	6,021

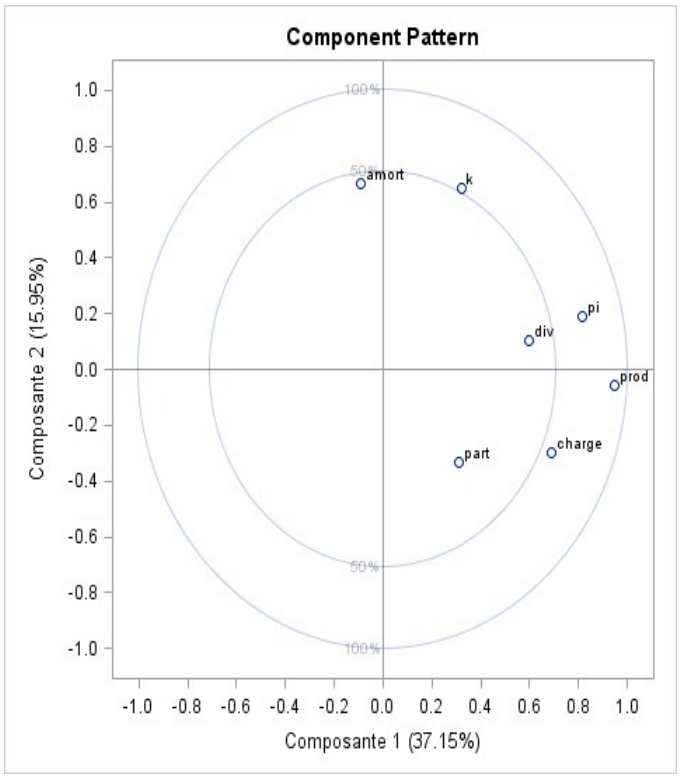
Source : Fare 2008, Part désigne le bonus de la participation aux fruits de l'expansion. StD désigne l'écart-type.

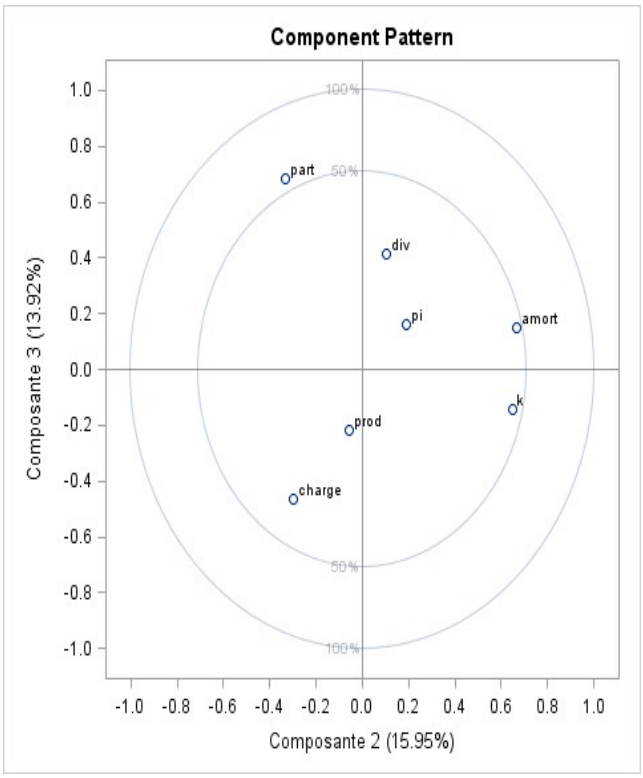
1- Nous proposons de commencer par l'analyse de la matrice de corrélation suivante :

	div	prod	k	pi	amort	part	charge
prod	0,376	1,000					
k	0,087	0,234	1,000				
pi	0,429	0,749	0,268	1,000			
amort	-0,002	-0,073	0,063	-0,039	1,000		
part	0,116	0,181	-0,007	0,178	-0,054		
charge	0,178	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.

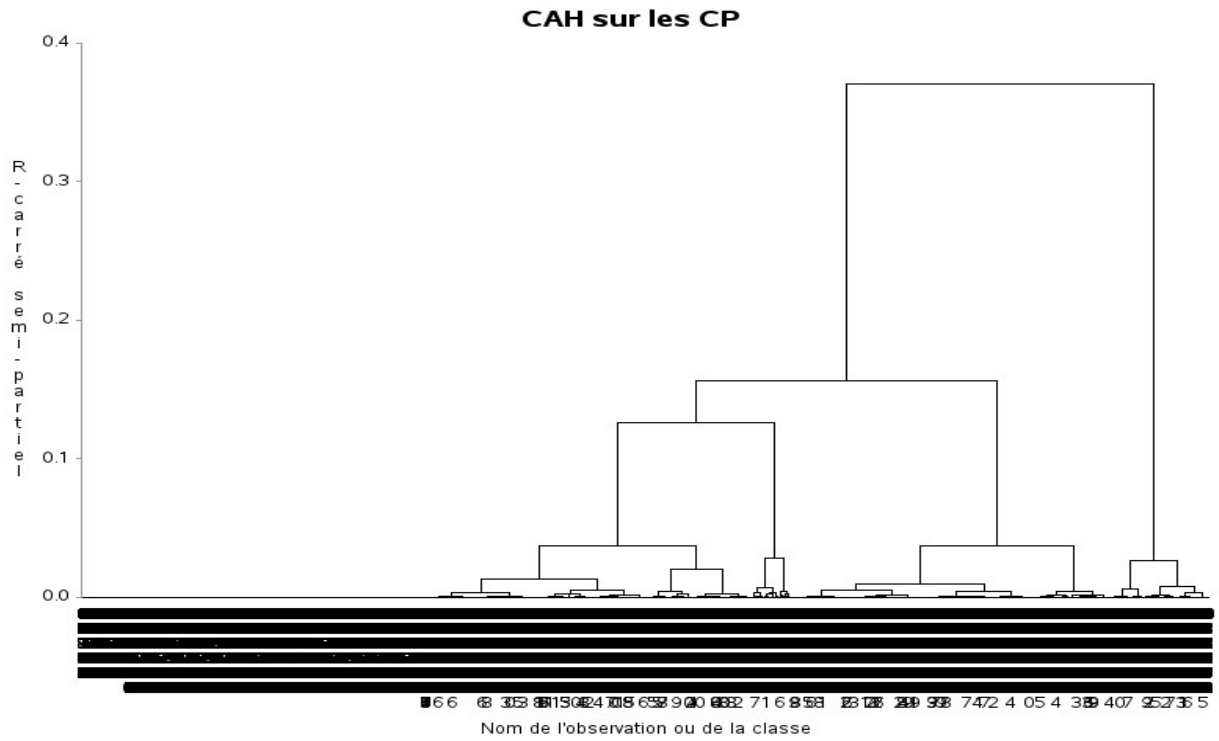
2- Nous avons ensuite effectué une ACP sur l'ensemble de ces variables. Les cinq premières valeurs propres sont : 2.60 ; 1.12 ; 0.97 ; 0.91 et 0.85
   a- Donner la part d'inertie expliquée par chacun de ces axes. Quel est le nombre maximal d'axes que l'on peut retenir.
   b- Les deux graphiques suivants 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.





3- Nous avons ensuite effectué une Classification ascendante hiérarchique. L'arbre correspondant à l'analyse est le suivant :

4-



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

b- Nous avons choisi de retenir une typologie à 4 classes. La description des classes est donnée dans le tableau suivant. Quelles sont les variables pertinentes à l'analyse. Donner une interprétation à chacune des classes.

Cluster	N	Div	Prod	k	pi	amort	part	charge
1	244	14,24	128,82	109,02	48,44	0,30	0,31	21,46
2	7176	2,06	56,83	26,50	11,43	0,31	0,08	13,05
3	1373	3,24	61,11	122,36	14,27	0,31	0,07	11,68
4	1318	2,38	126,22	3,15	60,69	0,23	0,10	25,14
Total	***10111***	***2,56***	***68,19***	***38,46***	***19,13***	***0,31***	***0,08***	***14,64***

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. Un pays en développement, ayant des difficultés à acheter des vaccins à toute sa population, décide de procéder à une analyse lui permettant de cibler les individus prioritaires. Nous supposons que ce pays dispose d'une base de données de tous les individus atteints du Covid. Ces derniers sont observés en fonction de leur caractéristiques individuelles (age, genre, IMC, morbidité1 , morbidité 2, …, morbidité 5). Nous avons une indicatrice signalant le décès ou non des malades.

En vous inspirant des modèles de scoring, décrire minutieusement les différentes étapes à effectuer (y compris le modèle économétrique) pour obtenir un score permettant de classer les individus en fonction de l'urgence de les vacciner.