



+50/1/23+

## M1 ISFA - Analyse de données et clustering

## Examen du 4 janvier 2022

Documents de cours et TD autorisés

Durée 3h

Vérifiez que votre sujet comporte bien ?? pages. Toutes les feuilles sont à rendre en fin d'épreuve.

Les questions faisant apparaître le symbole ☺ peuvent présenter une ou plusieurs bonnes réponses. Les autres ont une unique bonne réponse.

Vous veillerez à bien noircir au stylo l'intérieur des cases correspondant aux bonnes réponses (et non les cocher ou les entourer simplement) pour limiter les erreurs de correction automatique.

## Exercice kmeans

Sur les quatre graphiques de la figure ??, nous considérons les résultats de procédures de formation de classes reposant sur des méthodes de centres mobiles. Pour chaque graphique, on cherche à déterminer si le nombre de classes fixé est adapté aux données et si les groupes formés semblent satisfaisants.

**Question 1 ☺** Cochez les références des figures pour lesquelles le nombre de classes recherchées semble adapté aux données représentées :

- A [1c]   B [1a]   C [1b]   D [1d]   E Aucune de ces réponses n'est correcte.

**Question 2 ☺** Cochez les références des figures pour lesquelles les classes formées semblent adaptées aux données représentées :

- A [1c]   B [1d]   C [1b]   D [1a]   E Aucune de ces réponses n'est correcte.

**Question 3 ☺** À en juger par les graphiques, pour quelle paire de figures le  $R^2$  sera-t-il le plus mauvais ?

- A [1a]   B [1d]   C [1b]   D [1c]   E Aucune de ces réponses n'est correcte.

**Question 4** La meilleure chance d'améliorer au moins l'un de ces partitionnements, c'est

- A d'augmenter la valeur du paramètre `nstart`  
B d'augmenter la valeur du paramètre `iter.max`  
C de changer la valeur du paramètre `algorithm`  
D d'abaisser la valeur du paramètre `iter.max`

**Question 5 ☺** Pour chacune des sous-figures de la figure ??, le nombre de classes naturelles est indiqué. En tenant compte des centres de gravité de ces classes, cochez les références des figures pour lesquelles la méthode des *kmeans* permettra la distinction correcte de toutes ces classes :

- A [2a]   B [2d]   C [2b]   D [2c]   E Aucune de ces réponses n'est correcte.

**Question 6 ☺** Cochez les références des figures pour lesquelles la méthode des *kmeans* pourrait permettre la distinction correcte d'au moins deux classes :

- A [2b]   B [2c]   C [2d]   D [2a]   E Aucune de ces réponses n'est correcte.

## Exercice

Nous considérons la composition d'eaux minérales. Trois variables sont mesurées pour quatre eaux minérales différentes :

1. magnesium (*Mg*) : teneur en magnésium,
2. calcium (*Ca*) : teneur en calcium,
3. nitrate (*NO<sub>3</sub>*) : teneur en nitrate.

Les données mesurées sont données par la table ?? 1

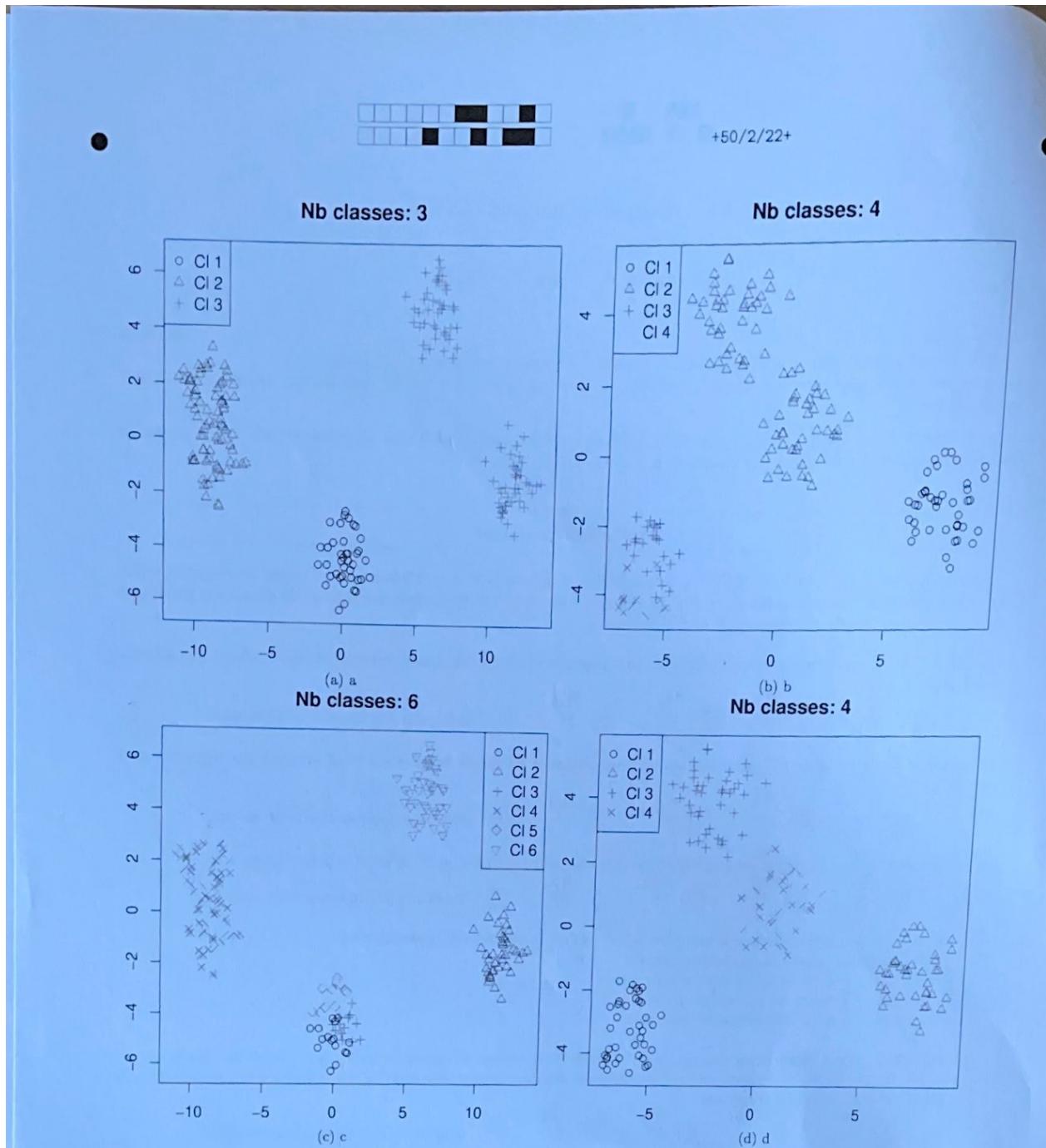


FIGURE 1 – Essais de segmentation

	<i>Mg</i>	<i>Ca</i>	<i>NO<sub>3</sub></i>
<i>eau<sub>1</sub></i>	7	4.5	3
<i>eau<sub>2</sub></i>	4.5	9.5	2.5
<i>eau<sub>3</sub></i>	5.5	9.5	8.5
<i>eau<sub>4</sub></i>	0.5	5.5	2.5

TABLE 1 - Composition des eaux

**Question 1** Notons  $d_\infty$  la distance de Tchebychev,  $d_{\text{Man}}$  la distance de Manhattan et  $d^2$  le carré de la distance euclidienne standard. Cochez toutes les propositions vraies :

- A  $d_{\text{Man}}(e_1, e_2) = 8$        B  $d_{\infty}(e_2, e_4) = 4.5$        C  $d^2(e_3, e_1) = 57.5$        D  $d^2(e_3, e_4) = 77$   
 E  $d_{\infty}(e_2, e_3) = 6$        F  $d_{\text{Man}}(e_1, e_4) = 9$        G Aucune de ces réponses n'est correcte.



+50/3/21+

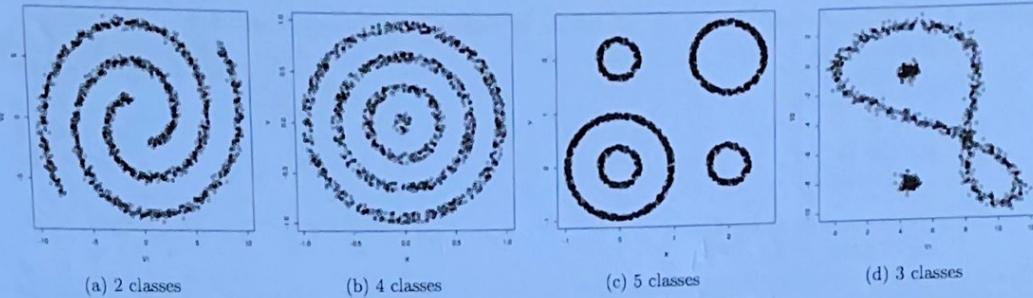


FIGURE 2 – Amas de points

**Question 2** Reportez dans la table ?? les distances pour toutes les paires de points et lorsque la distance choisie est la distance de Tchebychev.

 F  P  A  M  N  L  J

	$e_1$	$e_2$	$e_3$	$e_4$
$e_1$	0			
$e_2$		0		
$e_3$			0	
$e_4$				0

TABLE 2 – Matrice de distances

**Question 3** Nous souhaitons à présent construire une CAH à partir de la matrice de distances précédente (Cf table ??) et en adoptant la stratégie du saut maximum comme critère d'agrégation. Reportez dans les matrices ci-dessous (Table ??) les distances mises à jour au cours de la construction de la CAH.

 F  P  A  M  N  L  J

4	24	
0		
	0	
		0

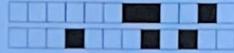
  

		0
		0

TABLE 3 – Matrices de distances

**Question 4** Ecrivez dans la table ?? l'objet `merge` (produit par la fonction `hclust`) correspondant à la CAH construite.

 F  P  A  B  J



+50/4/20+

**Question 5** Reportez dans la table ?? les valeurs de l'ultramétrique correspondant à la CAH construite.

F	P	A	M	N	L	J
---	---	---	---	---	---	---

1 <sup>re</sup> fusion	-2	-4
2 <sup>e</sup> fusion	-1	1
3 <sup>e</sup> fusion	-3	2

(a) Objet merge

	$e_1$	$e_2$	$e_3$	$e_4$
$e_1$	0	3	3	3
$e_2$		0	3	3
$e_3$			0	3
$e_4$				0

(b) Matrice des valeurs de l'ultramétrique

TABLE 4 – Fusions et ultramétrique

### Etude qualitative du risque de ruine

Le jeu de données provient d'une étude qualitative faisant le lien entre différents types de risques avec le risque de ruine d'une entreprise. Votre jeu de données est un échantillon des données disponibles. Les variables disponibles sont les suivantes :

- IndR porte sur le risque industriel.
- MngR porte sur le risque de management.
- FinR porte sur le risque financier.
- CrdR porte sur le risque de crédit.
- CompR porte sur le risque d'image/créabilité.
- OpR porte sur le risque de compétitivité.
- Class porte sur le risque de ruine.

Les 6 premières variables ont trois modalités indiquant la prise en compte du risque en question : P pour "protégé", A pour "acceptable" et N pour "négligé". La variable Class a deux modalités indiquant l'exposition au risque de ruine : B pour banqueroute (ruine) et NB pour l'absence de risque fort de ruine.

Considérez la sortie ci-après :

```
IndR.MngR.FinR.CrdR.CompR.OpR.Class
1          P,P,A,A,A,P,NB
2          N,N,A,A,A,NB
3          A,A,A,A,A,NB
```

**Question 1** Considérez la sortie ci-après. Quelle commande parmi celles proposées a provoqué cette sortie ?

- A head(read.csv('Bankruptcy.data.txt'),n=3)
- B head(read.table('Bankruptcy.data.txt',header=TRUE),n=3)
- C head(read.csv('Bankruptcy.data.txt',header=TRUE),n=3)
- D head(read.table('Bankruptcy.data.txt',sep=',',n=3))
- E head(read.table('Bankruptcy.data.txt',dec=',',n=3))

**Question 2** En vous aidant de la sortie précédente, cochez toutes les commandes ci-dessous qui auraient permis une lecture correcte des données.

- A head(read.table('Bankruptcy.data.txt',sep=',',h=TRUE),n=3)
- B head(read.table('Bankruptcy.data.txt',header=TRUE),n=3)
- C head(read.csv('Bankruptcy.data.txt',header=TRUE),n=3)
- D head(read.csv('Bankruptcy.data.txt'),n=3)
- E head(read.table('Bankruptcy.data.txt',dec=',',n=3))
- F head(read.table('Bankruptcy.data.txt',sep=',',n=3))
- G Aucune de ces réponses n'est correcte.

On suppose les données chargées correctement dans la variable ru. Considérez les tables ci-après :



+50/5/19+

```
attach(ru)
table(Class, IndR)
  IndR
Class A N P
  B 28 53 26
  NB 53 36 54
```

	A	N	P
Class	23	73	11
NB	46	46	51

	A	N	P
Class	4	102	1
NB	70	17	56

	A	N	P
CrdR	17	87	3
NB	60	7	76

detach(ru)

**Question 3** Indiquez les deux risques, parmi les quatre testés ci-dessus, avec lesquels Class est la plus liée, en le justifiant.

 F  P  A  B  J

les deux risques les plus liés à la variable Class sont le risque financier et le risque de crédit, car leur table de contingence respectives sont les plus éloignées de celle liée à une situation d'indépendance.

Considérons la table de contingence de Class avec

**Question 4** Encodez les degrés de liberté de la table croisant Class avec CrdR :

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

**Question 5** Supposons que la valeur approchée de la distance du  $\chi^2$  correspondant à la table croisant Class et IndR soit de 10,61. Encodez la probabilité de dépasser cette valeur (*p-value*), arrondie à la troisième décimale.

0	1	2	3	4	5	6	7	8	9
.									
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

**Question 6** Si nous entreprenons de réaliser une AFC sur l'une de ces tables, quel sera le nombre d'axe maximum ?

0	1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---	---

**Question 7** Justifiez votre réponse à la réponse précédente.

 F  P  A  B  J

comme posant n le nombre de modalités de la variable CrdR et p celui de Class, le rang de la table de contingence est au plus  $\min(n-1; p-1) = \min(2-1; 3-1) = 1$ , le nombre d'axe maximum de l'AFC associée sera de 1.



+50/6/18+

Projection des modalités de Class et FinR sur le premier axe

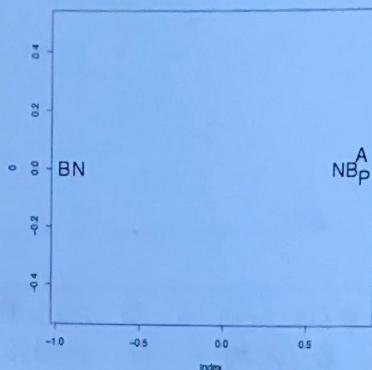


FIGURE 3 – Aperçu des projections des modalités sur le premier axe de projection de l'AFC appliquée à Class et FinR

**Question 8** Interprétez les relations entre les deux variables projetées sur le premier axe, selon le graphique ?? :

 F  P  A  B  J

Comme, selon une projection selon le premier axe, B et N sont proches et à l'opposé du groupe formé par A<sub>P</sub> et NB, la modalité B est souvent associée à la modalité N, tandis que la modalité NB est associée aux modalités A et P.

La commande suivante permet de réaliser une AFCM sur le tableau disjointif complet issu de ru :

```
afcm<-dudi.acm(ru,scannf=FALSE,nf=6)
```

Avant de tirer des informations de ces graphiques...

**Question 9** Dans le cas le plus général vu en cours, un carré de liaison est construit à partir de valeurs :

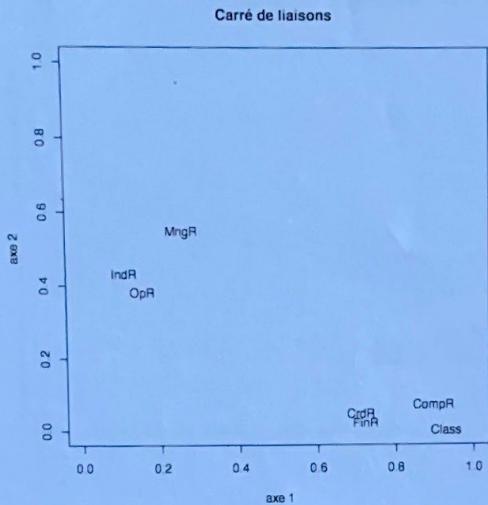
- A qui peuvent mesurer l'intensité de la relation entre une variable quantitative et une variable qualitative
- B qui permettent d'expliquer l'inertie d'un axe à partir des variables
- C qui peuvent mesurer l'intensité de la relation entre deux variables quantitatives
- D nous renseignant sur les liens entre modalités
- E qui peuvent mesurer l'intensité de la relation entre deux variables qualitatives
- F Aucune de ces réponses n'est correcte.

**Question 10** Quelles observations/interprétations pouvez-vous faire à partir de la figure ?? ?

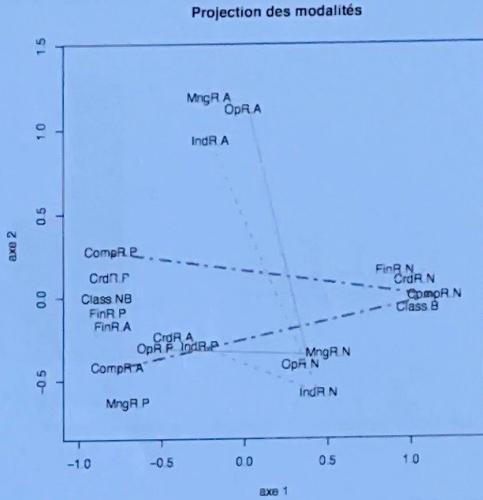
 F  P  A  B  J



+50/7/17+



(a) Pour l'identification de liaisons entre variables



(b) Pour la recherche de lien entre modalités

FIGURE 4 – Sorties graphiques produites avec afcm

**Question 11** Quelles observations/interprétations pouvez-vous faire à partir de la figure ???

F  P  A  B  J

### Exercice

Les données portent sur les mortalités liées aux maladies respiratoires dues à la pollution de l'air, en 2016. Les maladies observées dans 183 pays répartis sur cinq continents (Afrique, Amérique, Asie, Europe, Océanie) sont :

- les infections des voies respiratoires inférieures, LRI (pour *Lower Respiratory Infections*),
- les cancers des poumons, des bronches, de la trachée, TBLC (pour *Tracea, Bronchus, Lung Cancers*),
- les cardiopathies ischémiques, IHD (pour *Ischaemic Heart Disease*),
- les accidents vasculaires cérébraux, S (pour *Stroke*),
- les maladies pulmonaires obstructives chroniques, COPD (pour *Chronic Obstructive Pulmonary Disease*).

Les données sont exprimées en taux de mortalité pour 100 000 habitants standardisés par l'âge. La table ci-dessous fournit la moyenne et la variance descriptives pour chaque variable :

Variable	LRI	TBLC	IHD	S	COPD
Moyenne	11.17	1.55	21.54	9.47	5.36
Variance	175.71	2.14	176.35	35.44	20.25

**Question 1** ☺ À la vue de ces premières informations, vaut-il mieux

- A centrer les données ?  
 B centrer-réduire les données ?  
 C Aucune de ces réponses n'est correcte.



+50/8/16+

Nous réalisons l'ACP adéquate et nous représentons le graphe des valeurs propres (figure 3) :

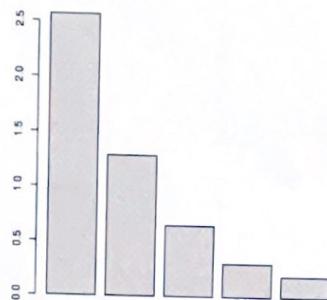


FIGURE 5 – Valeurs propres

Question 2 ☺ Combien d'axes conserver ?

- A 1     B 4     C 3     D 5     E 2     F Aucune de ces réponses n'est correcte.

Nous représentons le cercle des corrélations ainsi que les continents sur le premier plan factoriel (figure 4).

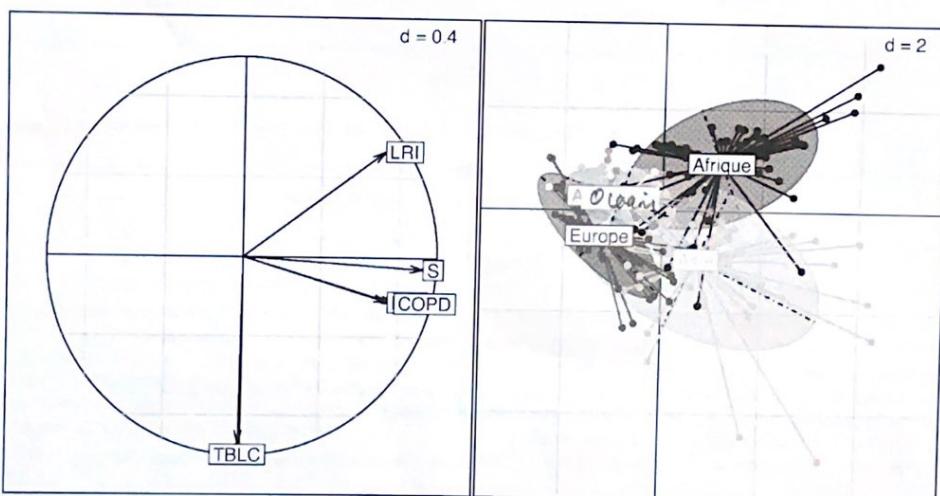


FIGURE 6

Question 3 ☺ Cochez les propositions correctes.

- A IHD et COPD sont fortement corrélées entre elles et expliquent l'axe 1.  
 B Nous observons un effet taille.  
 C Toutes les variables sont corrélées à l'axe 1.  
 D TBLC et LRI sont inversement corrélées sur l'axe 2.  
 E S intervient dans l'interprétation de l'axe 1.  
 F Aucune de ces réponses n'est correcte.



+50/9/15+

**Question 4** Donnez trois phrases caractéristiques de la représentation des continents sur le premier plan factoriel.

F  P  A  B  J

---

### Quelques questions sur le cours

---

**Question 1** Quelle particularité présentent les méthodes de classification basées sur la densité ? Peut-on les voir comme des applications allant de l'ensemble des points à classer vers un ensemble fini ?

F  P  A  B  J

**Question 2** Vous souhaitez réaliser une ACP sur un jeu de données présentant des données manquantes. Proposez trois méthodes à cette fin.

F  P  A  B  J



+50/10/14+

**Question 3** Etant donnés un ensemble de données de volume très important et une machine de calcul précise, proposez une procédure permettant d'estimer le temps nécessaire au partitionnement de ces données - le nombre de classes étant fixé par ailleurs.

 F  P  A  B  J

**Question 4** Quelles réserves doivent accompagner les résultats d'une classification obtenue à l'issue d'une procédure de type k-means ?

 F  P  A  B  J

Fin



+50/11/13+

### Annexe - Table de valeurs du $\chi^2$

La table ci-dessous donne, en fonction de la valeur du nombre de degré de liberté (abrégé par ddl dans la suite) et en fonction de  $P$ , la valeur du  $\chi^2$  telle que la probabilité pour une variable aléatoire suivant une loi du  $\chi^2$  de dépasser cette valeur est  $P$ .

ddl	1 - $P$													
	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
1	0.000	0.000	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.69	13.82
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.07	12.83	15.09	16.75	20.51
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	0.980	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.603	3.053	3.816	4.575	5.578	7.584	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	3.074	3.571	4.404	5.226	6.304	8.438	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.565	4.107	5.009	5.892	7.041	9.299	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.075	4.660	5.629	6.571	7.790	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.601	5.229	6.262	7.261	8.547	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.142	5.812	6.908	7.962	9.312	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.697	6.408	7.564	8.672	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.265	7.015	8.231	9.390	10.86	13.68	17.34	21.60	25.09	28.87	31.53	34.81	37.16	42.31
19	6.844	7.633	8.907	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.434	8.260	9.591	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.31
21	8.034	8.807	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.643	9.542	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.260	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.886	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.65	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
31	14.46	15.66	17.54	19.28	21.43	25.39	30.34	35.89	41.42	44.99	48.23	52.19	55.00	61.10
32	15.13	16.36	18.29	20.07	22.27	26.30	31.34	36.97	42.58	46.19	49.48	53.49	56.33	62.49
33	15.82	17.07	19.05	20.87	23.11	27.22	32.34	38.06	43.75	47.40	50.73	54.78	57.65	63.87
34	16.50	17.79	19.81	21.66	23.95	28.14	33.34	39.14	44.90	48.60	51.97	56.06	58.96	65.25
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27	66.62
36	17.89	19.23	21.34	23.27	25.64	29.97	35.34	41.30	47.21	51.00	54.44	58.62	61.58	67.98
37	18.59	19.96	22.11	24.07	26.49	30.89	36.34	42.38	48.36	52.19	55.67	59.89	62.88	69.35
38	19.29	20.69	22.88	24.88	27.34	31.81	37.34	43.46	49.51	53.38	56.90	61.16	64.18	70.70
39	20.00	21.43	23.65	25.70	28.20	32.74	38.34	44.54	50.66	54.57	58.12	62.43	65.48	72.06
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.93	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.1	118.5	124.3	129.6	135.8	140.2	149.4
120	83.85	86.92	91.57	95.70	100.6	109.2	119.3	130.1	140.2	146.6	152.2	159.0	163.6	173.6
140	100.7	104.0	109.1	113.7	119.0	128.4	139.3	150.9	161.8	168.6	174.6	181.8	186.8	197.4
160	117.7	121.3	126.9	131.8	137.5	147.6	159.3	171.7	183.3	190.5	196.9	204.5	209.8	221.0
180	134.9	138.8	144.7	150.0	156.2	166.9	179.3	192.4	204.7	212.3	219.0	227.1	232.6	244.4
200	152.2	156.4	162.7	168.3	174.8	186.2	199.3	213.1	226.0	234.0	241.1	249.4	255.3	267.5
240	187.3	192.0	199.0	205.1	212.4	224.9	239.3	254.4	268.5	277.1	284.8	293.9	300.2	313.4
300	240.7	246.0	253.9	260.9	269.1	283.1	299.3	316.1	331.8	341.4	349.9	359.9	366.8	381.4
400	330.9	337.2	346.5	354.6	364.2	380.6	399.3	418.7	436.6	447.6	457.3	468.7	476.6	493.1