

SEMESTRE 5
LICENCE 3 mention ECONOMIE

~~LICENCE 3 mention ECONOMIE et INFORMATIQUE~~

ANALYSE DES DONNEES
(durée 1h30)

Mardi 3 janvier 2012 ~ 13h30 -15h00

S. CASANOVA

Le barème est donné à titre indicatif

1 Questions de cours (4 points)

1. A quel type de données est adaptée une analyse en composantes principales (ACP) ?
2. Combien peut-on calculer de composantes principales au maximum ?
3. Donner la formule de calcul de la contribution de l'individu i à la composante c^j .
4. Donner la formule de calcul de la contribution de la variable x^k à la composante c^j .
5. Donner la formule de la distance du saut de Ward avec les notations du cours.
6. Donner la formule analytique de la décomposition de l'inertie pour une partition en k groupes avec les notations du cours.

2 Analyse en composantes principales (ACP) (9 points)

Nous disposons de données concernant 22 sites de e-commerce.

Les variables considérées sont les suivantes :

- VISITEURS : nombre de visiteurs (en milliers) sur un même site au cours d'une période donnée.
- CA : chiffre d'affaires (en millions d'euros) réalisé par le groupe en une année.
- GOOGLE : nombre d'articles (en millions) référencés sur le moteur de recherche Google en cherchant le site du e-commerce.
- ACHAT : pénétration d'achat (en pourcentage)
- PERFORMANCE : temps nécessaire (en millième de secondes) au chargement complet de la page du site.

On réalise une ACP sur ces 5 variables avec le logiciel SAS. Utiliser les sorties données dans l'annexe 1 pour répondre aux questions suivantes.

1. L'ACP est-elle justifiée ?
2. Calculer les valeurs A, B et C qui ont été supprimées dans les sorties et préciser à quoi correspondent ces valeurs.

3. On souhaite conserver 80% de l'inertie. Donner, en le justifiant, le nombre de composantes principales à retenir pour cette ACP.
4. Interpréter les composantes principales retenues à l'aide des variables initiales.
5. Donner, pour chaque axe retenu, les sites bien représentés.
6. Quels sont les sites ayant fortement contribué au premier axe ?
7. Commenter l'ACP.

3 Analyse typologique (AT) (7 points)

On souhaite réaliser une classification des 22 sites par les deux méthodes d'analyse typologique étudiées en cours. Utiliser les sorties SAS données dans l'annexe 2 pour répondre aux questions suivantes.

3.1 Agrégation autour des moyennes mobiles (AMM)

On réalise une classification des 22 sites en 5 groupes par la méthode d'agrégation autour des moyennes mobiles.

1. Quel est le groupe le plus proche du groupe 5 ? Du groupe 2 ? Justifier.
2. Donner l'effectif de chaque groupe et les sites de chaque groupe.
3. Interpréter le R^2 de la variable GOOGLE et le R^2 global.
4. Donner, en la justifiant, une typologie succincte des sites.

3.2 Classification ascendante hiérarchique (CAH)

On réalise maintenant une classification ascendante hiérarchique des 22 sites.

1. Décrire la partition en 18 groupes (donner l'effectif et les individus de chaque groupe).
2. Donner, en le justifiant, le nombre de groupes à retenir pour la classification.
3. Pour le nombre de groupes retenu à la question 2, donner l'effectif de chaque groupe et les sites de chaque groupe en expliquant votre méthode. Comparer avec la classification par AMM.

ANNEXE 1 : Sorties ACP

Procédure PRINCOMP

Observations 22
Variables 5

Statistiques simples

	VISITEUR	PENETRAT	PERFORMA	GOOGLE	CA
Moyenne	5180.045455	15.53636364	3.460909091	10.67577273	1.003000000
StD	3216.605405	12.10700488	1.951453449	20.66711658	1.201279080

Matrice de corrélation

	VISITEUR	PENETRAT	PERFORMA	GOOGLE	CA
VISITEUR	1.0000	0.8436	0.0109	-.1377	-.0938
PENETRAT	0.8436	1.0000	0.0659	-.1380	-.0099
PERFORMA	0.0109	0.0659	1.0000	-.0169	0.0643
GOOGLE	-.1377	-.1380	-.0169	1.0000	-.2244
CA	-.0938	-.0099	0.0643	-.2244	1.0000

Valeurs propres de la matrice de corrélation

	Eigenvalue	Difference	Proportion	Cumulative
1	(A)	0.66195880	0.3781	0.3781
2	1.22854492	0.23612579	0.2457	0.6238
3	0.99241913	0.25443070	0.1985	(B)
4	0.73798843	0.58744464	0.1476	0.9699
5	0.15054380		0.0301	1.0000

Vecteurs propres

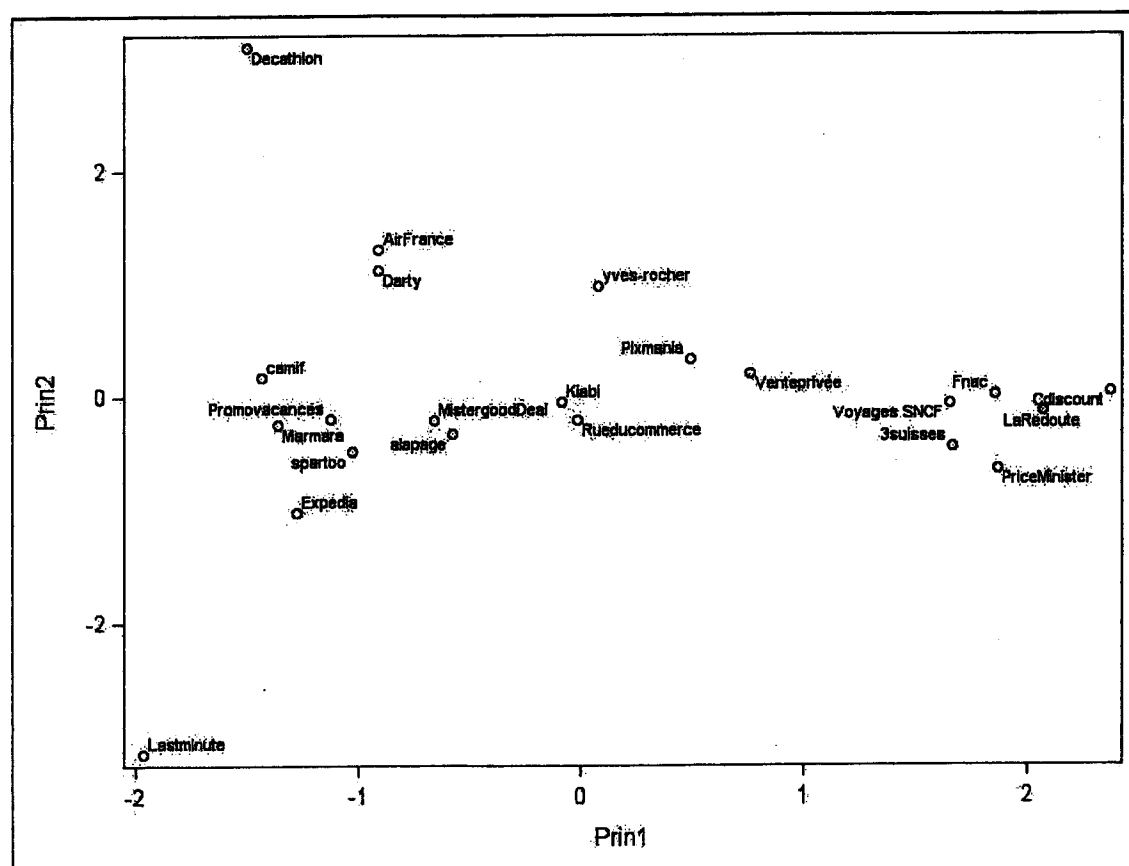
	Prin1	Prin2	Prin3	Prin4	Prin5
VISITEUR	0.689486	-.127259	-.029440	0.083615	0.707500
PENETRAT	0.690366	-.057937	0.027234	0.164719	-.701544
PERFORMA	0.061878	0.229565	0.956343	-.165092	0.040296
GOOGLE	-.208892	-.624967	0.282928	0.696641	0.020601
CA	-.023214	0.732915	-.061250	0.673285	0.072334

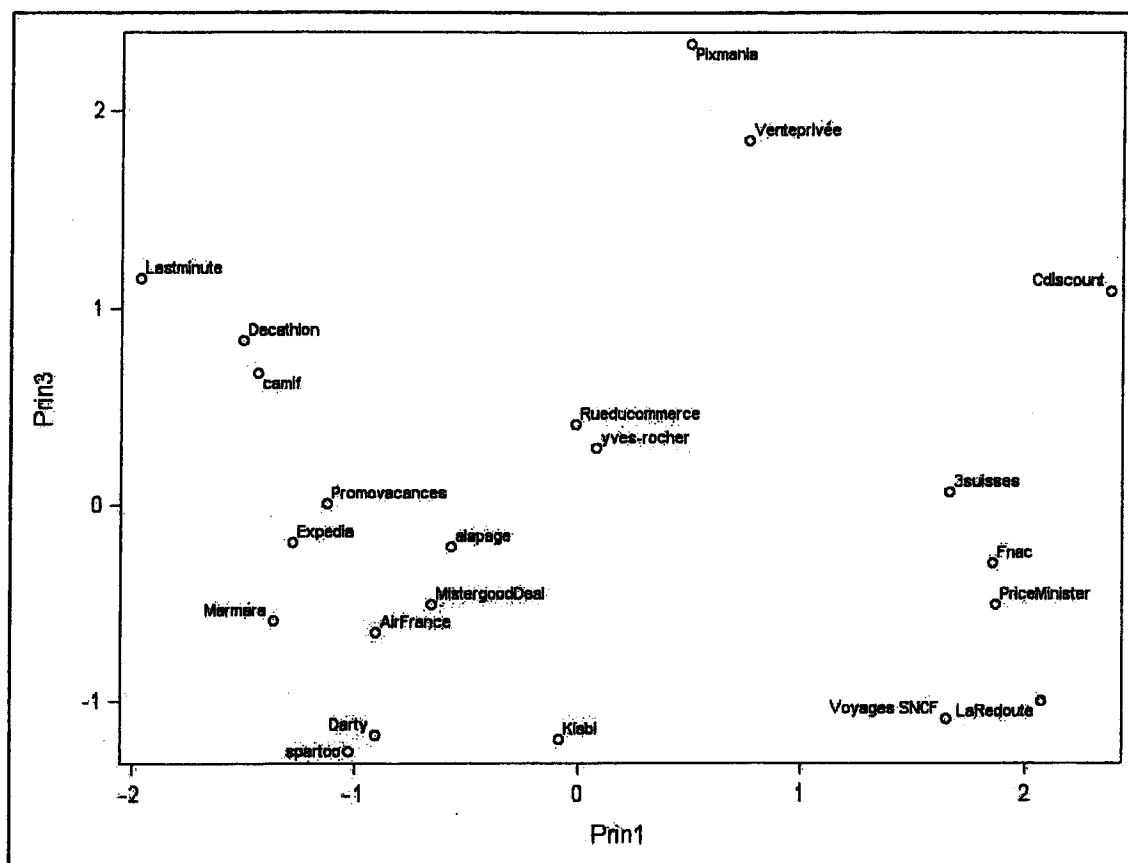
Coefficients de corrélation de Pearson, N = 22

Proba > |r| sous H0: Rho=0

	Prin1	Prin2	Prin3
VISITEUR	0.94801	(C)	-0.02933
	<.0001	0.5312	0.8969
PENETRAT	0.94922	-0.06422	0.02713
	<.0001	0.7765	0.9046
PERFORMA	0.08508	0.25445	0.95271
	0.7066	0.2531	<.0001
GOOGLE	-0.28722	-0.69271	0.28185
	0.1950	0.0004	0.2038
CA	-0.03192	0.81236	-0.06102
	0.8879	<.0001	0.7874

Obs	NOM	rap1	rap2	rap3	contri Axel
1	PriceMinister	0.75531	0.09076	0.05642	8.4081
2	Cdiscount	0.78574	0.00025	0.16520	13.6084
3	LaRedoute	0.75568	0.00337	0.17571	10.3373
4	Fnac	0.88109	0.00002	0.02270	8.3216
5	Voyages SNCF	0.59744	0.00095	0.26099	6.5550
6	3suisses	0.91639	0.06446	0.00153	6.6940
7	Venteprivée	0.12872	0.00892	0.75300	1.4006
8	Pixmania	0.03922	0.01700	0.86094	0.5983
9	Kiabi	0.00489	0.00211	0.93498	0.0180
10	Rueducommerce	0.00020	0.07610	0.29353	0.0003
11	MistergoodDeal	0.28487	0.03090	0.17390	1.0336
12	spartoo	0.27512	0.06073	0.41465	2.5287
13	Promovacances	0.67896	0.02018	0.00003	3.0730
14	Darty	0.19879	0.29658	0.33313	1.9890
15	Marmara	0.67819	0.02398	0.12863	4.4627
16	Lastminute	0.18972	0.49030	0.06478	9.2812
17	Expedia	0.59490	0.38359	0.01336	3.9263
18	alpage	0.17942	0.05861	0.02571	0.7870
19	yves-rocher	0.00623	0.82945	0.07332	0.0169
20	camif	0.48447	0.00640	0.10530	4.9771
21	Decathlon	0.15439	0.65008	0.04745	5.4512
22	AirFrance	0.21345	0.43931	0.10919	1.9862





ANNEXE 2 : Sorties AT

The FASTCLUS Procedure

Replace=PARTIAL Radius=0 Maxclusters=5 Maxiter=20 Converge=0.02

Iteration History						
Relative Change in Cluster Seeds						
Iteration	Criterion	1	2	3	4	5
1	0.6353	0.2105	0.3661	0.2312	0	0.4252
2	0.4786	0	0.0749	0	0	0.2092
3	0.4632	0	0	0	0	0

Cluster Listing		
Obs	NOM	CLUSTER
1	PriceMinister	1
2	Cdiscount	3
3	LaRedoute	1
4	Fnac	1
5	Voyages SNCF	1
6	3suisses	1
7	Venteprivée	3
8	Pixmania	3
9	Kiabi	2
10	Rueducommerce	2
11	MistergoodDeal	2
12	spartoo	2
13	Promovacances	2
14	Darty	5
15	Marmara	2
16	Lastminute	4
17	Expedia	2
18	alapage	2
19	yves-rocher	3
20	camif	2
21	Decathlon	5
22	AirFrance	5

Cluster Summary					
Maximum Distance					
Cluster	Frequency	RMS Std Deviation	from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	5	0.3505	1.0189	3	2.3506
2	9	0.4645	1.4077	?	2.5829
3	4	0.6861	1.6133	1	2.3506
4	1	.	0	2	4.5204
5	3	0.7387	1.8151	?	2.7396

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
VISITEUR	1.00000	0.48870	0.806664	4.172355
PENETRAT	1.00000	0.54385	0.760563	3.176467
PERFORMA	1.00000	0.73720	0.560052	1.272993
GOOGLE	1.00000	0.30748	0.923466	12.066031
CA	1.00000	0.46468	0.825201	4.720867
OVER-ALL	1.00000	0.52698	0.775189	3.448185

Distance Between Cluster Centroids					
Nearest Cluster	1	2	3	4	5
1	.	2.888252112	2.350561349	5.409643118	3.664082615
2	2.888252112	.	2.582940517	4.520368069	2.739616305
3	2.350561349	2.582940517	.	5.255552618	3.311147294
4	5.409643118	4.520368069	5.255552618	.	5.425679841
5	3.664082615	2.739616305	3.311147294	5.425679841	.

Cluster=1					
Procédure MEANS					
Variable	N	Mean	Std Dev	Minimum	Maximum
VISITEUR	5	9605.00	1483.50	7287.00	10954.00
PENETRAT	5	31.1600000	3.0046630	27.80	36.00
PERFORMA	5	2.3840000	0.8688095	1.46	3.52
GOOGLE	5	8.5734000	5.4631774	0.98	14.20
CA	5	1.0942000	0.3220950	0.75	1.47

Cluster=2					
Variable	N	Mean	Std Dev	Minimum	Maximum
VISITEUR	9	3237.33	1513.30	810.00	5323.00
PENETRAT	9	7.4111111	5.7533565	1.40	17.00
PERFORMA	9	2.8511111	1.2855198	1.09	5.00
GOOGLE	9	6.2877778	8.0805675	0.16	26.30
CA	9	0.2398889	0.2546515	0.01	0.83

Cluster=3					
Variable	N	Mean	Std Dev	Minimum	Maximum
VISITEUR	4	6855.75	1849.65	4500.00	9022.00
PENETRAT	4	23.0000000	11.4600756	15.00	40.00
PERFORMA	4	6.4350000	1.7082642	4.35	8.24
GOOGLE	4	5.5675000	3.6554377	2.73	10.50
CA	4	0.9747500	0.6893018	0.51	2.00

Cluster=4					
Variable	N	Mean	Std Dev	Minimum	Maximum
VISITEUR	1	2391.00	.	2391.00	2391.00
PENETRAT	1	7.1000000	.	7.10	7.10
PERFORMA	1	3.2100000	.	3.21	3.21
GOOGLE	1	99.4000000	.	99.40	99.40
CA	1	0.0250000	.	0.025	0.025

Cluster=5					
Variable	N	Mean	Std Dev	Minimum	Maximum
VISITEUR	3	2328.67	1519.89	1000.00	3986.00
PENETRAT	3	6.7333333	4.5883911	3.60	12.00
PERFORMA	3	3.2033333	2.2571073	1.45	5.75
GOOGLE	3	4.5800000	1.5256802	3.39	6.30
CA	3	3.5040000	1.2120264	2.72	4.90

Procédure CLUSTER

Analyse de classification de la variance minimum de Ward

Les données ont été normalisées à la moyenne 0 et la variance 1

Cluster History

NCL	-----Clusters	Joined-----	FREQ	SPRSQ	RSQ
21	Venteprivée	Pixmania	2	0.0019	.998
20	Promovacances	Marmara	2	0.0020	.996
19	Fnac	3suisses	2	0.0025	.994
18	MistergoodDeal	CL20	3	0.0027	.991
17	PriceMinister	LaRedoute	2	0.0028	.988
16	Darty	AirFrance	2	0.0049	.983
15	CL17	CL19	4	0.0057	.977
14	CL18	spartoo	4	0.0067	.971
13	Rueducommerce	alapage	2	0.0079	.963
12	CL14	Expedia	5	0.0096	.953
11	CL15	Voyages SNCF	5	0.0124	.941
10	Kiabi	CL13	3	0.0139	.927
9	CL10	yves-rocher	4	0.0173	.910
8	CL12	camif	6	0.0197	.890
7	CL11	Cdiscount	6	0.0308	.859
6	CL9	CL8	10	0.0339	.825
5	CL16	Decathlon	3	0.0471	.778
4	CL7	CL21	8	0.1217	.656
3	CL6	CL5	13	0.1483	.508
2	CL3	Lastminute	14	0.1901	.318
1	CL4	CL2	22	0.3179	.000

