ANALYSE DES DONNEES

Cours 1ère partie – Sylvie Viguier-Pla
Septembre 2009

I. INTRODUCTION

Le but de l'analyse des données est, de façon générale, de décrire une situation à l'aide de mesures relevées. L'intérêt des méthodes réside dans leur pouvoir de résumé, adapté aux grands fichiers de données. Ces mesures sont soit de nature quantitative (dénombrement, évaluation, mesure numérique), soit de nature qualitative (pour laquelle on ne peut calculer de résumé numérique tel une moyenne).

Pour toutes les analyses des données, et quelle que soit la présentation des données dont on dispose, il faut pouvoir construire (parfois formellement) un tableau de données brutes, c'est-à-dire un tableau de n individus sur lequel sont mesurées p variables.

Les méthodes employées diffèrent selon la nature des variables (quantitative ou qualitative), le nombre de variables et la problématique à résoudre.

I.1. Exemples de tableaux de données

I.1.1. Les pays de l'OCDE

Les données proviennent de l'Agence Internationale de l'Energie (IEA=International Energy Agency).
On dispose, pour les 30 pays de l'OCDE, des valeurs suivantes :

popul : population (millions d'habitants, 2002)
CO2 : émission de CO2 par combustion de fuel (Mt, 2002)
TPES : Total Primary Energy supply (millions de tonnes équivalent pétrole, 2002)
electr : consommation d'électricité (teraWatts heure, 2003, 1 TWh=1 millier de milliards de Watts/heure)
import : importation d'électricité (teraWatts heure, 2003)
export : exportation d'électricité (teraWatts heure, 2003)
fuel : production d'électricité d'origine fossile (teraWatts heure, 2003)
nuclear : production d'électricité d'origine nucléaire (teraWatts heure, 2003)
hydro : production d'électricité d'origine hydrolique (teraWatts heure, 2003)
geoth : production d'électricité d'origine géothermique ou autre (teraWatts heure, 2003)
zone : zone géographique (pacasi=Pacifique/Asie, europ=Europe, ameri=Amérique)
climat : climat prédominant (aride, temp=tempéré, arct=arctique/tempéré, medi=méditerranéen, mari=maritime)

NoPays	Pays	popul	CO2	TPES	electr	import	export	fuel	nuclear	hydro	geoth	zone	clim
1	Australie	19,75	342,85	112,71	187,2	0	0	169,2	0	18	0	pacasi	aride
2	Autriche	8,05	66,14	30,44	63,8	19	13,4	22,4	0	31,9	3,9	europ	temp
3	Belgique	10,33	112,55	56,89	87,3	15	8,2	34,3	44,8	1,3	0,1	europ	temp
4	Canada	31,41	531,86	250,03	555	30,5	36,4	158	70,3	332,6	0	ameri	arct
5	Tchéquie	10,21	114,96	41,72	60,4	10,1	26,3	50,5	24,4	1,8	0	europ	temp
6	Danemark	5,38	51,17	19,75	35,2	7	15,6	38,2	0	0	5,6	europ	temp
7	Finlande	5,2	63,5	35,62	84,7	11,9	7	48,6	21,8	9,3	0,1	europ	arct
8	France	61,23	377,07	265,88	471,8	6,2	72,2	55	420,7	62,1	0	europ	temp
9	Allemagne	82,48	837,53	346,35	529,9	45,4	54,1	361,4	156,5	20,7	0	europ	temp
10	Grèce	10,95	90,46	29,02	56,4	4,2	2,1	47,9	0	5,3	1,1	europ	medi
11	Hongrie	10,16	55,45	25,45	40	14,1	7,1	21,9	11	0,2	0	europ	temp
12	Islande	0,29	2,22	3,4	8,4	0	0	0	0	7	1,3	europ	temp
13	Irlande	3,91	42,45	15,3	21,7	1,1	0	20	0	0,5	0	europ	temp
14	Italie	58,03	433,24	172,72	329,2	51,4	0,4	228,8	0	44,1	5,3	europ	medi
15	Japon	127,4	1206,9	516,93	1034	0	0	698	228,2	104,4	3,3	pacasi	medi
16	Corée	47,64	451,55	203,5	322,2	0	0	185,7	129,6	6,8	0	pacasi	temp
17	Luxembourg	0,45	9,28	4,04	7,5	6,2	2,9	3,3	0	0,8	0	europ	temp
18	Mexique	100,4	365,15	157,31	160	0,1	1,2	125,6	10	19,5	6	ameri	aride
19	Pays-Bas	16,15	177,88	77,92	106,1	20,8	3,8	84,2	3,6	0,1	1,2	europ	temp
20	NZélande	3,98	34	8,01	35,4	0	0	10,6	0	22,8	2	pacasi	temp
21	Norvège	4,54	33,06	26,52	114,1	13,4	5,5	0,9	0	105,1	0,1	europ	arct
22	Pologne	38,22	282,9	89,19	127,4	5	14,9	133,8	0	3,3	0,1	europ	temp
23	Portugal	10,37	62,98	26,39	39,1	5,9	3,1	22,4	0	14	0	europ	marit
24	Slovaquie	5,38	37,89	18,55	26,2	4,2	7,2	8,9	17	3,3	0	europ	temp
25	Espagne	40,55	303,41	131,56	231,8	9,5	8,2	124	59,3	47,3	0	europ	temp
26	Suède	8,93	50,12	51,03	145,3	24,3	11,5	13,5	65,5	53	0,6	europ	arct
27	Suisse	7,29	42,83	27,14	62,1	42,4	45,5	2,9	25,9	36,4	0	europ	temp
28	Turquie	69,67	193,05	75,42	133,2	1,1	0,6	37,7	0	34,8	0,1	pacasi	temp
29	Royaume Uni	59,21	529,27	226,51	367,9	5,8	3,2	277,6	81,9	5,9	0	europ	temp
30	Etats-Unis	287,5	5652,3	2290,4	3852	30,3	24,8	2724	766,5	266,6	89	ameri	temp

I.1.2. Les stations de ski de Savoie

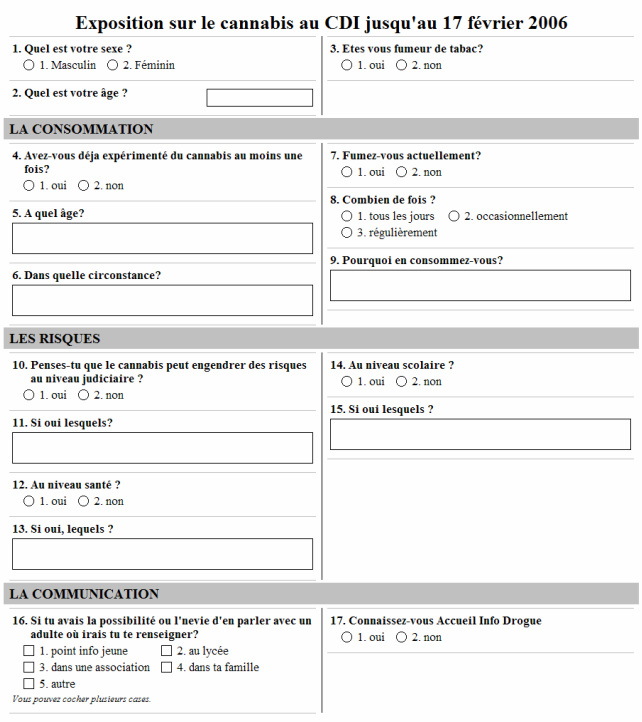
Les données concernent des stations de ski en Savoie. On dispose, pour 32 stations, des variables suivantes (données 1998) :

prixforf : prix du forfait 1 semaine (Euros)
altmin : altitude minimum de la station (m)
altmax : altitude maximum de la station (m)
pistes : nombre de pistes de ski alpin
kmfond : nombre de kilomètres de pistes de ski de fond
remontee : nombre de remontées mécaniques

Nom	prixforf	altmin	altmax	pistes	kmfond	remontee
LesAillons	76	900	2000	45	50	22
LesArcs	160	800	3226	117	30	69
Arèches	85	750	2300	30	47	15
Aussois	71	500	2750	21	10	11
Bessans	54	1710	2200	4	80	4
Bonneval	79	1850	3000	16	0	10
LeCorbier	88	1550	1850	36	25	24
Courchevel	140	1100	2707	100	66	67
Crest-Voland	82	1230	1650	26	7	17
Flumet	42	1000	1600	20	12	40
LesKarellis	84	1600	2550	28	30	17
LesMenuires	140	1400	3200	61	28	45
Méribel	140	1100	2950	74	33	60
LaNorma	93	1350	2750	25	6	18
Bellecombe	86	1150	2070	32	8	18
LaPlagne	159	1250	3250	123	73	110
Pralognan	87	1410	2355	22	25	14
LaRosière	122	1150	2640	32	12	19
LesSaisies	95	1150	1941	26	80	24
StFrancois	96	1450	2550	28	50	17
StMartin	140	1450	3200	61	28	48
StSorlin	81	1500	2600	26	16	13
LaTania	140	900	1900	100	8	67
Tignes	160	1550	3450	129	52	96
LaToussuire	98	1800	2400	27	12	19
ValCenis	74	1400	2800	38	6	22
Valfréjus	78	1550	2737	20	2	12
ValdIsère	160	1850	2560	70	24	51
Valloire	107	1430	2600	80	20	34
Valmeinier	107	1450	2600	75	15	33
Valmorel	123	1250	2550	56	20	36
ValThorens	103	1800	3200	54	5	30

I.1.3. Les réponses à une enquête

Une enquête auprès de lycéens visait à mieux connaître leur comportement vis-à-vis du tabac et du cannabis. Voici la fiche questionnaire de l'enquête :



Exposition sur le cannabis au CDI jusqu'au 17 février 2006

1. Quel est votre sexe ? ○ 1. Masculin ○ 2. Féminin

2. Quel est votre âge ?

3. Etes vous fumeur de tabac? ○ 1. oui ○ 2. non

LA CONSOMMATION

4. Avez-vous déja expérimenté du cannabis au moins une fois? ○ 1. oui ○ 2. non

5. A quel âge?

6. Dans quelle circonstance?

7. Fumez-vous actuellement? ○ 1. oui ○ 2. non

8. Combien de fois ? ○ 1. tous les jours ○ 2. occasionnellement ○ 3. régulièrement

9. Pourquoi en consommez-vous?

LES RISQUES

10. Penses-tu que le cannabis peut engendrer des risques au niveau judiciaire ? ○ 1. oui ○ 2. non

11. Si oui lesquels?

12. Au niveau santé ? ○ 1. oui ○ 2. non

13. Si oui, lequels ?

14. Au niveau scolaire ? ○ 1. oui ○ 2. non

15. Si oui lesquels ?

LA COMMUNICATION

16. Si tu avais la possibilité ou l'nevie d'en parler avec un adulte où irais tu te renseigner?
☐ 1. point info jeune ☐ 2. au lycée
☐ 3. dans une association ☐ 4. dans ta famille
☐ 5. autre
Vous pouvez cocher plusieurs cases.

17. Connaissez-vous Accueil Info Drogue ○ 1. oui ○ 2. non

et un extrait des réponses :

no	sexe	âge	fumeur	cannabis	agecan	circonst	fumecan	freqcan	pourquoican	risquejustice	quel_r_justice	risquesanté	quel_r_santé
1	M	15	non	oui	15	ami	non					oui	perte d'endurance
2	M	18	oui	oui	15	ami	oui	tous les jours	pour s'amuser	oui		oui	
3	M	17	non	oui	14 ans	ami	non	occasionnellement		oui	amendes peine de prison...	oui	cancer probleme respiratoire
4	M	18	oui	oui	14	ami	oui	tous les jours	par ce que j'aime sa	oui	la prison	oui	problemes cardiovasculaire
5	M		non	non			non			oui	prison	oui	au lycée
6	M		non	oui	15	ami	non	occasionnellement	pour se faire plaisir	oui		oui	
7	M			oui	13	fête	oui	régulièrement	pour etre bien	non	garde a vue	non	
8	M		oui	oui	16		oui	occasionnellement		oui		oui	beaucoup
9	M	16	oui	non						oui	une forte amende	oui	
10	M	17	non	non			non			oui		oui	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1197	M		non	oui	17	grande fête	non					oui	cerveau
1198	M		non	oui	17	petite fête	non			oui	autre comportement	oui	dépendance
1199	F	17	oui	non							1amende	oui	perte des neurones, on maigrit
1200	F	16	oui	oui	14	pour essayer entre amis	oui		envie de savoir	oui	garde à vue	oui	dépendance, cancer
1201	F	16	oui	oui	14	avec copains	non			oui	amende, prison	oui	réflexes, mémoire
1202	F	16	oui	oui	14	l'été dans une soirée	non			oui	prison, amende	oui	neurones, réflexes
1203	F	16	non	non			non					oui	halucinations, problèmes respiratoires
1204	F	16	non	non			non					oui	
1205	M	16	non	non			non			oui	amendes,saisie du cannabis	oui	cancer, mémoire, dépression
1206	F		non	oui	13	avec copains pour savoir	non			oui	être fiché		

no	risquescolaire	quel_r_scol	renseign1	renseign2	renseign3	renseign4	renseign5	ADI
1	oui	pertes de neurones	dans ta famille					non
2	oui		autre					oui
3	oui	pertes de concentation	autre					non
4	oui	mauvais résultat						non
5			au lycée				hon	
6	oui		dans une association					non
7	non		autre					non
8	oui		dans ta famille					non
9	oui	moins attentif	point info jeune					oui
10	oui							non
...	...	...	...	...	...	...	...	...
1197	oui	mémoire	dans ta famille					non
1198			au lycée	dans une association	dans ta famille			non
1199	oui	exclusion, relachement	dans ta famille					oui
1200	oui	concentration, réflexion, résultats en baisse	dans ta famille					oui
1201	oui	concentration, comprhension	autre					non
1202	oui	ne pas être à fond dans les études	autre					non
1203	oui	difficulté à suivre les cours,à réfléchir	point info jeune					non
1204	oui	pas de concentration en cours	au lycée	dans une association				non
1205	oui	echec scolaire	autre					non
1206	oui	difficulté à suivre les cours,à réfléchi						oui

I.2. Les différentes méthodes et leur but

Les principales méthodes d'analyse des données sont les suivantes :

- ACP (Analyse en Composantes Principales), appliquée à p variables quantitatives, dans le but de résumer les liens entre les variables par l'analyse des covariances ou des corrélations, et dresser une ``carte'' des individus indiquant leur position par rapport à ces liens.

- AFC (Analyse Factorielle des Correspondances), appliquée à 2 variables qualitatives, dans le but de mettre en évidence, graphiquement, le lien entre les deux variables traitées, et accessoirement voir quels individus influencent le plus ce lien.

- AFCM (Analyse Factorielle des Correspondances Multiple), appliquée à p variables qualitatives, avec le même but que l'AFC, mais pour plus de 2 variables.

- AD (Analyse Discriminante), appliquée à une variable qualitative et p variables quantitatives. Son but est à rapprocher de celui de l'ACP, avec la différence que le résumé de l'information doit se faire tout en gardant les groupes formés par la variable qualitative les plus distincts possible. On peut aussi se servir de cette méthode pour classifier des individus (exemple : credit scoring).

- Classification (hiérarchique et non hiérarchique), appliquée à p variables, soit quantitatives, soit qualitatives, rarement les deux types mélangés, dans le but de regrouper les individus les plus ressemblants.

Pour la description des classes formées par une méthode de classification, on peut être amené à utiliser l'AD, ou à étudier des tableaux croisés, ou encore à utiliser la méthode :

- ANOVA (Analyse de la Variance, ou ANalysis Of VAriance), liée à un résultat de classification, pour décrire les classes. Il est à noter que cette dernière technique est utilisée surtout dans d'autres domaines de la statistique tels l'étude des modèles linéaires.

Dans ce cours, nous verrons dans l'ordre l'ACP, l'AFC, l'AFCM, la classification et, s'il reste du temps, l'AD. L'ANOVA, qui sera vue parallèlement dans un autre cours, nous servira pour l'interprétation des résultats de la classification.

II. ANALYSE EN COMPOSANTES PRINCIPALES

II.1. Les données, le but de l'ACP et les étapes

II.1.1. Les données, présentation par l'exemple

L'ACP est une méthode permettant de visualiser de façon synthétique un ensemble de variables quantitatives mesurées sur un ensemble d'individus, et de voir comment les individus se positionnent dans les liens entre ces variables.

Le tableau des données se présente comme celui destiné à une étude par régression. C'est le but de la méthode qui diffère, puisque, contrairement à la régression, il n'y a pas de variable à expliquer, mais un ensemble de variables à synthétiser.

Considérons pour commencer un exemple simple de tableau de données (avec p=3 et n=4) :

élève	math	français	anglais
e1	0	2	3
e2	0	0	3
e3	4	2	1
e4	4	0	1

Les étapes de l'ACP sont les suivantes

données de départ

E	M	F	A
e1	0	2	3
e2	0	0	3
e3	4	2	1
e4	4	0	1
moy	2	1	2
var	4	1	1

centrage = soustraction de la moyenne

données centrées

E	M_c	F_c	A_c
e1	-2	1	1
e2	-2	-1	1
e3	2	1	-1
e4	2	-1	-1
moy	0	0	0
var	4	1	1

réduction = division par l'écart-type

données réduites

E	M_r	F_r	A_r
e1	-1	1	1
e2	-1	-1	1
e3	1	1	-1
e4	1	-1	-1
moy	0	0	0
var	1	1	1

rotation = multiplication matricielle par une "matrice de rotation"

composantes principales

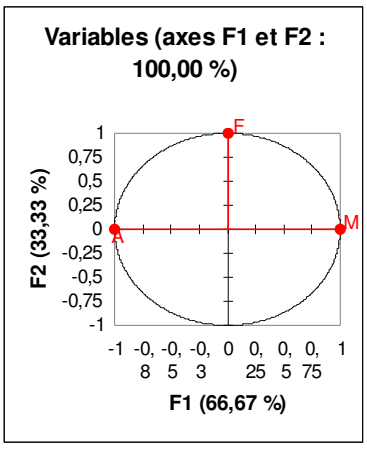
E	C_1	C_2	C_3
e1	-racine(2)	1	0
e2	-racine(2)	-1	0
e3	racine(2)	1	0
e4	racine(2)	-1	0
moy	0	0	0
var	2	1	0

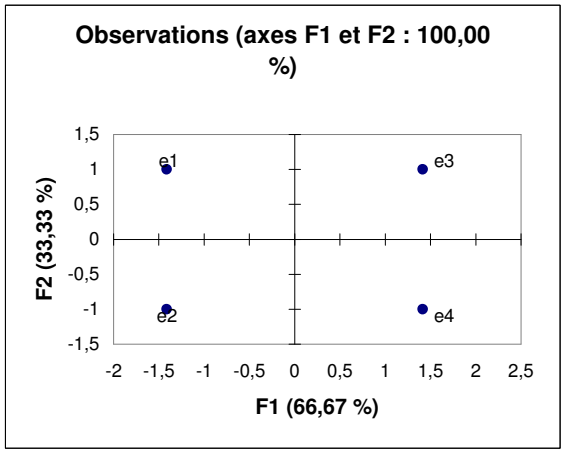
nuage (M,F) : (à tracer)

On retiendra de ces étapes que :
- Des données de départ, où les lignes sont les individus et les colonnes des variables, un ensemble d'étapes fait arriver au tableau des composantes principales, de même nombre de lignes et de colonnes que le tableau de départ, les lignes étant toujours les individus de départ, mais les colonnes n'ont plus la même signification, elles sont chacune un résumé des variables de départ.
- L'inertie totale du nuage des individus (qui est un nuage dans un espace à p dimensions), est égale à la somme des variances des variables. Elle est la même pour le tableau de données réduites (c'est-à-dire p) que pour le tableau des composantes principales.
- Les variances des composantes principales sont appelées valeurs propres. Elles sont ordonnées dans l'ordre décroissant.

On appelle aussi les composantes principales des "axes", "dimensions", "facteurs". Le vocabulaire qu'on emploiera, qui a une signification bien précise dans un contexte plus mathématique, utilisera indifféremment ces termes pour désigner la même notion, c'est-à-dire les différents résumés de l'ensemble des variables.

On représente graphiquement les variables par un cercle des corrélations, et les individus par un nuage de points, ce qui donne les graphes ci-contre.
Une corrélation entre deux variables pouvant être vue comme le cosinus de l'angle entre ces deux variables, les angles entre les variables dans le cercle des corrélations permet de retrouver le fait que les variables A et M ont une corrélation –1 entre elles, et que F est "orthogonal" à A et M, c'est-à-dire de corrélation nulle avec A et M.





L'interprétation des axes à partir d'un cercle des corrélations se fait de la manière suivante :
L'axe 1, qui est corrélé positivement avec M et négativement avec A, est un axe qui oppose les élèves meilleurs en M et moins bons en A (qui seront ceux de droite sur le nuage) aux élèves de caractéristique opposée (à gauche).
L'axe 2, qui est corrélé positivement avec la variable F, ordonne les élèves selon leur importance pour cette variable. Ceux du bas du nuage seront ceux de moins bonne note, et ceux du haut de meilleure note en F.
Dans cet exemple simple, on retrouve sur l'axe 1 le fait que les élèves e1 et e2 sont les moins bons en M et les meilleurs en A, contrairement aux élèves e3 et e4, et sur l'axe 2 le fait que les élèves e2 et e4 sont les moins bons en F, et e1 et e3 les meilleurs.
D'autres éléments d'aide à l'interprétation seront utiles dans des cas plus complexes (et complets).

II.1.2. ACP réduite et ACP non réduite.

Dans l'exemple ci-dessus, l'ACP a consisté à pratiquer une rotation du nuage des individus mesurés avec des variables préalablement réduites. On dit qu'on fait une ACP réduite. Il faut savoir que cette réduction des variables peut être omise, parfois par choix plutôt arbitraire, d'autres fois par nécessité. Voici quelques éléments qui diffèrent entre les deux types d'ACP.

	ACP non réduite	*ACP réduite*
Données	Variables exprimées dans la même unité, Avec des valeurs du même ordre de grandeur	Variables exprimées dans des unités de mesure différentes, ou d'ordre de grandeur trop différentes
Valeurs propres	Somme=somme des variances des variables de départ	Somme=somme des variances des variables réduites, c'est-à-dire p=nombre de variables
Autres résultats	L'ACP réduite se prête mieux que la non réduite à la représentation des variables par cercle des corrélations, puisque les variables ne sont pas réduites au départ, mais l'interprétation telle qu'elle est pratiquée dans ce cours reste valable pour les deux types d'ACP.	

II.2. Les résultats d'une ACP et leur utilisation pour la synthèse de l'information contenue dans un fichier de données

Pour mieux voir la différence d'utilisation d'une ACP avec une méthode de régression, reprenons les données sur les stations de ski de Savoie. Nous verrons au fur et à mesure des résultats comment on fait l'interprétation d'une ACP.

II.2.1. Les statistiques simples et la matrice de corrélation

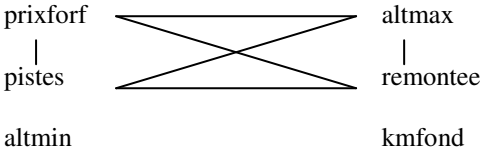
Tout d'abord, l'analyse des statistiques simples (moyenne, écart-type, quartiles, coefficient de variation, asymétrie, aplatissement) permet de voir si les données sont correctement réparties. En effet, des données présentant une asymétrie ou un étalement importants méritent une attention particulière, pour éventuellement détecter des valeurs aberrantes, entre autres.

Echantillon	Minimum	Maximum	1er Quartile	Médiane	3ème Quartile	Moyenne	Ecart-type (n)	Coefficient de variation	Asymétrie (Pearson)	Aplatissement (Pearson)
prixforf	42,000	160,000	81,750	95,500	140,000	104,688	32,096	0,307	0,316	-0,884
altmin	500,000	1850,000	1137,500	1400,000	1550,000	1322,813	328,484	0,248	-0,417	-0,263
altmax	1600,000	3450,000	2275,000	2600,000	2837,500	2566,750	479,913	0,187	-0,210	-0,675
pistes	4,000	129,000	26,000	34,000	71,000	50,063	33,454	0,668	0,953	-0,187
kmfond	0,000	80,000	9,500	22,000	36,500	27,500	22,757	0,828	0,988	-0,069
remontee	4,000	110,000	17,000	23,000	45,750	33,813	25,229	0,746	1,376	1,305

De même, nous avons déjà examiné la matrice de corrélation, que nous rappelons ici :

Matrice de corrélation (Pearson (n)) :						
Variables	prixforf	altmin	altmax	pistes	kmfond	remontee
prixforf	**1**	-0,007	**0,576**	**0,858**	0,212	**0,816**
altmin	-0,007	**1**	0,221	-0,152	-0,110	-0,144
altmax	**0,576**	0,221	**1**	**0,488**	0,025	**0,441**
pistes	**0,858**	-0,152	**0,488**	**1**	0,262	**0,930**
kmfond	0,212	-0,110	0,025	0,262	**1**	0,342
remontee	**0,816**	-0,144	**0,441**	**0,930**	0,342	**1**
Les valeurs en gras sont significativement différentes de 0 à un niveau de signification alpha=0,05						

Elle nous permet de voir les liens les plus significatifs entre variables, positifs ou négatifs. On peut tracer à partir de cette matrice un schéma des corrélations. Puisque toutes les corrélations significatives (en gras) sont positives, on peut faire un schéma comme suit :

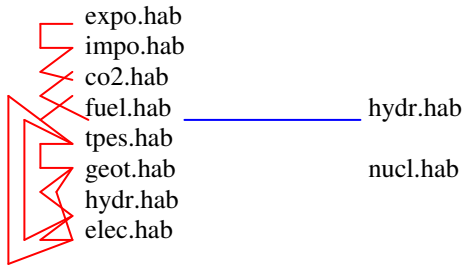


On voit ainsi mieux que les variables prixforf, altmax, pistes et remontee sont toutes corrélées positivement deux à deux, alors que altmin et kmfond ne paraissent pas liées à d'autres caractéristiques. Donc pour cet ensemble de stations, plus l'altitude max est élevée, plus le nombre de pistes, le nombre de remontées et le prix du forfait sont élevés.

Prenons, pour compléter notre idée du schéma des corrélations, un autre exemple, où certaines corrélations sont significativement négatives. C'est le cas pour les données de l'énergie dans les pays de l'OCDE. On a commencé par transformer les variables en valeurs par habitant, pour que l'effet population du pays n'intervienne pas dans les liens entre variables.

Matrice de corrélation :									
	co2.hab	tpes.hab	elec.hab	impo.hab	expo.hab	fuel.hab	nucl.hab	hydr.hab	geot.hab
co2.hab	1	**0,626**	0,345	**0,368**	0,234	**0,809**	-0,044	-0,053	-0,064
tpes.hab	**0,626**	1	**0,866**	0,355	0,228	0,292	0,175	**0,568**	**0,544**
elec.hab	0,345	**0,866**	1	0,313	0,183	0,011	0,125	**0,854**	**0,554**
impo.hab	**0,368**	0,355	0,313	1	**0,857**	0,123	0,015	0,081	-0,110
expo.hab	0,234	0,228	0,183	**0,857**	1	0,045	0,159	0,029	-0,089
fuel.hab	**0,809**	0,292	0,011	0,123	0,045	1	-0,080	**-0,402**	-0,207
nucl.hab	-0,044	0,175	0,125	0,015	0,159	-0,080	1	-0,100	-0,202
hydr.hab	-0,053	**0,568**	**0,854**	0,081	0,029	**-0,402**	-0,100	1	**0,628**
geot.hab	-0,064	**0,544**	**0,554**	-0,110	-0,089	-0,207	-0,202	**0,628**	1
En gras, valeurs significatives (hors diagonale) au seuil alpha=0,050 (test bilatéral)									

Ici il y a des corrélations négatives et positives. Le schéma va dans ce cas être disposé en 2 colonnes de variables. Les liens intra-colonnes représenteront des corrélations positives, les liens entre colonnes représenteront des corrélations négatives.



II.2.2. Les valeurs propres

Elles permettent d'effectuer un choix du nombre de composantes principales à retenir pour l'interprétation. Dans l'exemple simple du début, nous n'avions pas discuté de ce choix car la 3ème composante principale était nulle. Nous savons que les composantes principales ont des valeurs d'autant plus petites qu'on avance dans leur rang. Mais à partir de quel moment décide-t-on que la composante ne mérite plus interprétation ?
Voici les valeurs propres de l'ACP de notre tableau sur les stations :

Valeurs propres :						
	F1	F2	F3	F4	F5	F6
Valeur propre	3,193	1,247	0,855	0,475	0,169	0,061
Variabilité (%)	53,209	20,789	14,254	7,922	2,810	1,016
% cumulé	53,209	73,998	88,252	96,174	98,984	100,000

Le choix du nombre d'axes à interpréter se fait sur la base de règles. On donne ci-après les plus utilisées.

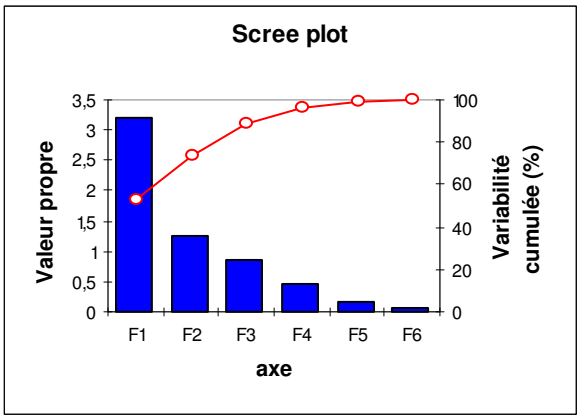
- *La règle de Kaiser.* Elle consiste à retenir les axes pour lesquels les valeurs propres sont supérieures à 1 (1 étant la moyenne de l'ensemble des valeurs propres). Il est à noter qu'on peut aussi avoir des résultats d'ACP dont la somme des valeurs propres n'est pas égale à p (cas de l'ACP non réduite). Dans ce cas, il faut adapter cette règle de Kaiser et retenir les valeurs propres supérieures à la moyenne des valeurs propres, et non plus à 1.
Dans notre exemple, le nombre de variables p=6, est bien la somme des valeurs propres. On retiendra donc 2 axes pour l'interprétation.

- *La part d'inertie.* Dans le tableau des valeurs propres ci-dessus figurent une ligne appelée "Variabilité(%)", et une autre appelée "% cumulé". La première correspond au pourcentage que représente la valeur propre de l'axe par rapport à la somme des valeurs propres, c'est-à-dire p. La seconde correspond au cumul de ces pourcentages jusqu'à l'axe concerné.
Le choix du nombre d'axes se fait par l'exigence d'un certain minimum de variabilité expliquée.

Par exemple, pour l'axe 2, le calcul de la variabilité est : 20,789%=1,247/6 et le % cumulé est 73,998=20,789+53,209. Cela signifie que le deuxième axe comporte 20,789 de la variance (ou variabilité, ou inertie) totale du nuage, et que le plan (1,2) totalise 73,998% de cette variance totale.
Si on souhaite interpréter au minimum 80% de la variance, il faudra interpréter 3 axes. Si 70% minimum suffisent, il faut interpréter 2 axes.

- *La règle de l'éboulis*. Elle consiste à retenir les 2 premiers axes au moins, puis de "couper" l'éboulis (ou scree plot) des valeurs propres entre les valeurs propres dont la différence est maximum.

Dans l'exemple, les différences entre valeurs propres à partir de la deuxième sont :
vp(2)-vp(3)= 0,392
vp(3)-vp(4)= 0,380
vp(4)-vp(5)= 0,307
vp(5)-vp(6)=0,108.
La différence maximum est entre les axes 2 et 3, on retient donc 2 axes.

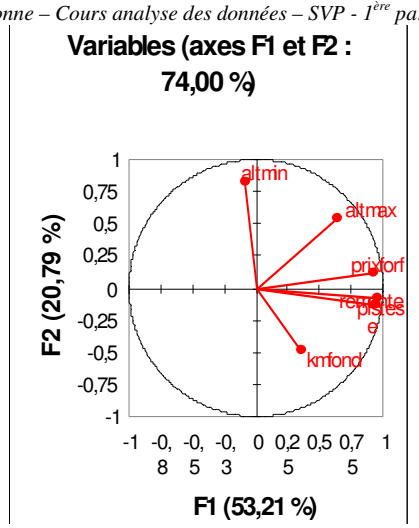


Remarque. Il existe d'autres règles de choix du nombre d'axes. La règle de l'éboulis combinée avec celle de Kaiser est une des meilleures. En effet, on commence par regarder combien de valeurs propres sont supérieures à la moyenne. Puis on regarde si la dernière valeur propre retenue (supérieure à la moyenne) est suffisamment éloignée de celle qui la suit (inférieure à la moyenne). Si oui, on reste sur la décision de la règle de Kaiser, si non, on coupera au saut plus important le plus près.
La prise en compte de la part d'inertie expliquée peut faire pencher la balance vers plus d'axes ou moins d'axes que ce que la règle de Kaiser amène.

II.2.3. Les représentations graphiques des individus et des variables

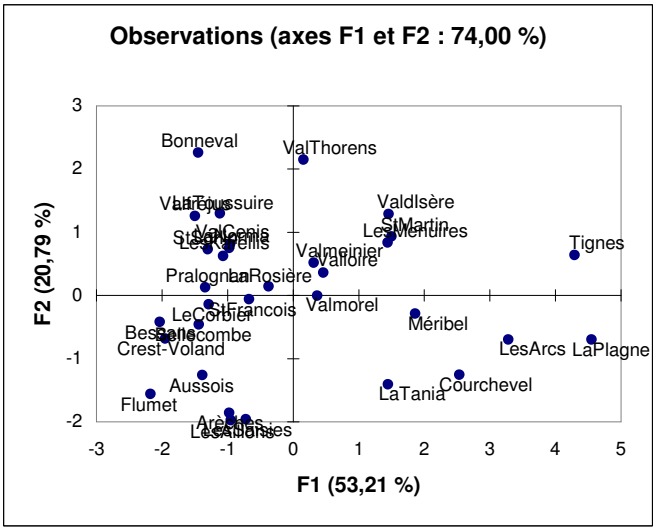
On a déjà vu que les variables étaient représentées par des cercles de corrélation, et les individus par des nuages de points.
Coordonnées des variables et représentation graphique :

Coordonnées des variables :						
	F1	F2	F3	F4	F5	F6
prixforf	0,926	0,107	-0,079	-0,106	0,336	-0,036
altmin	-0,084	0,825	0,475	-0,294	-0,032	0,009
altmax	0,640	0,543	-0,033	0,540	-0,059	-0,004
pistes	0,952	-0,080	-0,123	-0,161	-0,103	0,187
kmfond	0,361	-0,487	0,779	0,158	0,026	0,014
remontee	0,939	-0,133	-0,032	-0,187	-0,200	-0,156



Coordonnées des individus et représentation graphique :

Coordonnées des observations :						
Observation	F1	F2	F3	F4	F5	F6
LesAillons	-0,952	-1,975	0,326	0,151	-0,134	0,342
LesArcs	3,281	-0,699	-1,235	0,668	0,163	0,310
Arèches	-0,973	-1,858	0,004	0,942	0,280	0,101
Aussois	-1,383	-1,259	-1,711	1,801	-0,113	-0,074
Bessans	-2,036	-0,420	2,934	0,314	-0,206	0,119
Bonneval	-1,450	2,259	0,011	0,363	-0,271	-0,086
LeCorbier	-1,287	-0,141	0,429	-1,206	0,023	0,048
Courchevel	2,536	-1,254	0,728	0,031	0,006	0,207
Crest-Voland	-1,953	-0,681	-0,657	-1,125	0,164	-0,049
Flumet	-2,174	-1,560	-0,729	-0,871	-1,172	-0,587
LesKarellis	-1,066	0,623	0,693	0,072	-0,091	0,052
LesMenuires	1,443	0,835	-0,061	0,572	0,397	-0,207
Méribel	1,864	-0,287	-0,398	0,352	0,169	-0,299
LaNorma	-0,976	0,752	-0,614	0,448	0,073	-0,176
Bellecombe	-1,438	-0,460	-0,812	-0,397	0,098	0,022
LaPlagne	4,550	-0,697	0,980	0,079	-0,690	-0,420
Pralognan	-1,338	0,130	0,245	0,010	0,177	-0,049
LaRosière	-0,375	0,143	-0,803	0,390	0,837	-0,179
LesSaisies	-0,720	-1,959	1,854	0,053	0,498	-0,123
StFrancois	-0,673	-0,062	1,166	0,411	0,307	0,030
StMartin	1,498	0,933	0,013	0,475	0,327	-0,277
StSorlin	-1,302	0,731	0,036	0,214	-0,106	0,073
LaTania	1,447	-1,410	-1,671	-1,611	0,130	0,071
Tignes	4,290	0,639	0,650	-0,092	-0,628	0,035
LaToussuire	-1,115	1,300	0,314	-0,695	0,182	-0,102
ValCenis	-0,962	0,811	-0,546	0,422	-0,613	0,110
Valfréjus	-1,496	1,259	-0,381	0,298	-0,211	-0,058
ValdIsère	1,457	1,287	0,445	-1,320	0,798	-0,182
Valloire	0,461	0,360	-0,238	-0,383	-0,224	0,656
Valmeinier	0,313	0,516	-0,370	-0,414	-0,187	0,557
Valmorel	0,367	-0,005	-0,465	-0,161	0,381	-0,030
ValThorens	0,160	2,149	-0,140	0,209	-0,363	0,166



II.2.3. Aides à l'interprétation des axes

Pour savoir quelles variables donnent du sens à chaque axe et quelles variables il est inutile d'interpréter, on examine

- pour les variables, les cosinus carrés, qui ne sont autre que les carrés des coordonnées des variables :

Cosinus carrés des variables :						
	F1	F2	F3	F4	F5	F6
prixforf	0,857	0,012	0,006	0,011	0,113	0,001
altmin	0,007	0,680	0,225	0,086	0,001	0,000
altmax	0,409	0,295	0,001	0,292	0,003	0,000
pistes	0,907	0,006	0,015	0,026	0,011	0,035
kmfond	0,131	0,237	0,606	0,025	0,001	0,000
remontee	0,882	0,018	0,001	0,035	0,040	0,024

- pour les individus, les contributions et les cosinus carrés :

Contributions des observations (%) :							Cosinus carrés des observations :						
	F1	F2	F3	F4	F5	F6		F1	F2	F3	F4	F5	F6
LesAillons	0,887	9,770	0,389	0,149	0,331	6,004	LesAillons	0,179	0,769	0,021	0,004	0,004	0,023
LesArcs	10,54	1,225	5,572	2,938	0,494	4,919	LesArcs	0,807	0,037	0,114	0,033	0,002	0,007
Arèches	0,926	8,653	0,000	5,833	1,452	0,521	Arèches	0,176	0,642	0,000	0,165	0,015	0,002
Aussois	1,871	3,974	10,694	21,320	0,237	0,282	Aussois	0,197	0,164	0,302	0,335	0,001	0,001
Bessans	4,057	0,441	31,448	0,647	0,784	0,720	Bessans	0,317	0,013	0,658	0,008	0,003	0,001
Bonneval	2,059	12,79	0,000	0,864	1,366	0,375	Bonneval	0,283	0,688	0,000	0,018	0,010	0,001
LeCorbier	1,622	0,050	0,674	9,564	0,010	0,117	LeCorbier	0,499	0,006	0,056	0,438	0,000	0,001
Courchevel	6,294	3,942	1,937	0,006	0,001	2,189	Courchevel	0,750	0,183	0,062	0,000	0,000	0,005
Crest-Voland	3,732	1,163	1,575	8,315	0,498	0,122	Crest-Voland	0,635	0,077	0,072	0,211	0,004	0,000
Flumet	4,628	6,097	1,941	4,989	25,47	17,64	Flumet	0,465	0,239	0,052	0,075	0,135	0,034
LesKarellis	1,112	0,972	1,753	0,034	0,153	0,138	LesKarellis	0,562	0,192	0,238	0,003	0,004	0,001
LesMenuires	2,038	1,746	0,013	2,155	2,916	2,201	LesMenuires	0,629	0,210	0,001	0,099	0,048	0,013

Méribel	3,400	0,207	0,579	0,815	0,530	4,569	Méribel	0,878	0,021	0,040	0,031	0,007	0,023
LaNorma	0,932	1,417	1,379	1,321	0,098	1,595	LaNorma	0,446	0,265	0,177	0,094	0,002	0,015
Bellecombe	2,024	0,531	2,407	1,037	0,177	0,026	Bellecombe	0,666	0,068	0,212	0,051	0,003	0,000
LaPlagne	20,27	1,216	3,510	0,041	8,816	9,026	LaPlagne	0,908	0,021	0,042	0,000	0,021	0,008
Pralognan	1,752	0,042	0,220	0,001	0,580	0,123	Pralognan	0,942	0,009	0,032	0,000	0,016	0,001
LaRosière	0,138	0,051	2,356	1,002	12,99	1,643	LaRosière	0,083	0,012	0,381	0,090	0,414	0,019
LesSaisies	0,507	9,617	12,557	0,018	4,595	0,770	LesSaisies	0,064	0,476	0,426	0,000	0,031	0,002
StFrancois	0,443	0,010	4,971	1,111	1,742	0,045	StFrancois	0,218	0,002	0,654	0,081	0,045	0,000
StMartin	2,198	2,181	0,001	1,485	1,982	3,926	StMartin	0,637	0,247	0,000	0,064	0,030	0,022
StSorlin	1,660	1,339	0,005	0,300	0,210	0,273	StSorlin	0,739	0,233	0,001	0,020	0,005	0,002
LaTania	2,050	4,979	10,200	17,064	0,314	0,258	LaTania	0,221	0,209	0,294	0,274	0,002	0,001
Tignes	18,01	1,022	1,543	0,056	7,311	0,064	Tignes	0,937	0,021	0,022	0,000	0,020	0,000
LaToussuire	1,216	4,234	0,361	3,179	0,615	0,536	LaToussuire	0,349	0,475	0,028	0,136	0,009	0,003
ValCenis	0,905	1,649	1,091	1,170	6,962	0,615	ValCenis	0,378	0,269	0,122	0,073	0,153	0,005
Valfréjus	2,192	3,972	0,530	0,585	0,825	0,175	Valfréjus	0,545	0,386	0,035	0,022	0,011	0,001
ValdIsère	2,077	4,150	0,724	11,452	11,82	1,705	ValdIsère	0,332	0,259	0,031	0,273	0,100	0,005
Valloire	0,208	0,325	0,207	0,965	0,934	22,05	Valloire	0,207	0,126	0,055	0,143	0,049	0,419
Valmeinier	0,096	0,668	0,501	1,125	0,645	15,92	Valmeinier	0,096	0,262	0,135	0,168	0,034	0,305
Valmorel	0,132	0,000	0,791	0,171	2,694	0,046	Valmorel	0,257	0,000	0,414	0,050	0,278	0,002
ValThorens	0,025	11,58	0,071	0,286	2,449	1,411	ValThorens	0,005	0,949	0,004	0,009	0,027	0,006

II.2.4. Interprétation des axes

Signification des axes grâce aux statistiques sur les variables.

On remarque que la somme des cosinus carrés pour chaque ligne égale 1, ce qui fait une moyenne des cosinus carrés égale à 1/6.

Si on considère les variables dont le cosinus carré est supérieur à 1/6, nous pouvons les citer dans le tableau suivant, en relevant aussi le signe de leur coordonnée :

signe coordonnée	-	+
axe 1		prixforf, altmax, pistes, remontee
axe 2	kmfond	altmin, altmax

On peut en déduire que l'axe 1 est un axe d'échelle qui ordonne les stations de ski selon leur importance pour les valeurs de prixforf, altmax, pistes et remontee.

De même, l'axe 2 oppose les stations pour lesquelles kmfond est élevé et altmin, altmax sont faibles (stations de coordonnée négative) aux stations pour lesquelles kmfond est faible et altmin, altmax sont élevées (stations de coordonnée positive sur axe 2).

Remarque : l'axe 1 reflète le schéma des corrélations. On retrouve les 4 variables inter-corrélées positivement du même côté.

Interprétation de la position des individus.

La somme des contributions de l'ensemble des individus sur chaque axe égale 1 (ou 100%). Ce qui signifie qu'avec tous les individus on arrive à 100% de la variance de l'axe. C'est pour cette raison que la moyenne des contributions égale 1/n.

Dans notre exemple, n=32 individus, ou stations, donc la contribution moyenne égale 1/32.

Relevons les individus dont la contribution est supérieure à 1/32 sur les axes 1 et 2.

signe coordonnée	-	+
axe 1	Bessans, Crest-Voland, Flumet	LesArcs, Courchevel, Méribel, LaPlagne, Tignes
axe 2	Les Aillons, Arèches, Aussois, Courchevel, Flumet, Les Saisies, LaTania	Bonneval, LaToussuire, ValFréjus, ValdIsère, ValThorens

On peut en déduire que, l'axe 1 doit le sens qu'on a expliqué plus haut (c'est-à-dire ordonne les stations selon leur grandeur, globalement), au fait que les stations LesArcs, Courchevel, Méribel, LaPlagne et Tignes ont dans l'ensemble des valeurs élevées pour les variables prixforf, altmax, pistes, kmfond et remontee, alors qu'à l'opposé les stations Bessans, Crest-Voland et Flumet ont de faibles valeurs.

Le fait que l'axe 2 est l'axe d'opposition qu'on a décrit est surtout dû au fait que les stations Bonneval, LaToussuire, ValFréjus, ValdIsère et ValThorens sont des stations d'altitude avec peu de kmfond, alors qu'à l'opposé les stations Les Aillons, Arèches, Aussois, Courchevel, Flumet, Les Saisies et LaTania sont moins en altitude, et avec beaucoup de kmfond.

La prise en compte des cosinus carrés permet de terminer l'interprétation, en citant les individus qui ont une position interprétable dans la signification donnée aux axes.

Pour les individus, les cosinus carrés ont la même propriété que pour les variables, à savoir que pour chaque individu ligne, la somme des cosinus carrés égale 1. Il en découle que le cosinus carré moyen égale 1/6, quand il y a 6 axes.

Citons ces individus dont le cosinus carré dépasse la moyenne :

signe coordonnée	-	+
axe 1	LesAillons, Arèches, Aussois, Bessans, Bonneval, LeCorbier, Crest-Voland, Flumet, Les Karellis, LaNorma, Bellecombe, Pralognan, StFrançois, StSorlin, LaToussuire, ValCenis, ValFréjus	LesArcs, Courchevel, LesMénuires, Méribel, LaPlagne, StMartin, LaTania, Tignes, ValdIsère, Valloire, Valmorel
axe 2	Les Aillons, Arèches, Courchevel, Flumet, Les Saisies, LaTania	Bonneval, LesKarellis, LesMenuires, LaNorma, StMartin, StSorlin, LaToussuire, ValCenis, ValFréjus, ValdIsère, Valmeinier, ValThorens

Les stations de signe négatif sur l'axe 1 sont celles qui se positionnent comme à faible valeur dans cet ordre selon les variables prixforf, altmax, pistes, kmfond et remontee, celles qui sont de coordonnée positive se positionnent comme à fortes valeurs pour ces mêmes variables. A noter que toutes les stations qui contribuent sont bien reconstituées, c'est-à-dire ont un cosinus carré supérieur à la moyenne.

Les stations de coordonnée négative sur l'axe 2 sont celles qui ont beaucoup de km de fond et sont à altitude moyenne à faible, alors que celles de coordonnée positive sont à altitude élevée et avec peu de kmfond. A noter que toutes celles qui contribuent sont bien reconstituées, sauf une, Aussois. Cette station contribue à l'inertie de l'axe 2 comme station ayant une coordonnée élevée sur cet axe, donc avec des caractéristiques marquées pour altmin, altmax et kmfond, mais qui ne se positionne pas par rapport à cet axe. En regardant les données, on voit qu'elle a une altmin très faible (500m), mais en même temps une altmax supérieure à la moyenne (2750m), et un nombre de kmfond supérieur à la moyenne. Si on interprétait sa position, négative sur l'axe 2, on serait amené à dire que cette station a de faibles altmin et max, alors que ce n'est pas le cas. C'est ce qui explique son cosinus carré faible.

De même, pour chaque station mal représentée sur un axe, on peut expliquer pourquoi cette mauvaise reconstitution. On peut en conclure que cette aide à l'interprétation que sont les cosinus carrés est indispensable à une interprétation raisonnable des axes.

Ces axes peuvent être considérés comme de nouvelles variables synthétiques, l'axe 1 résumant l'importance de la station, l'axe 2 résumant l'orientation piste ou fond de la station, quand orientation il y a (ce qui n'est pas le cas de toutes les stations, on a vu le cas pour la station d'Aussois).

III. ANALYSE FACTORIELLE DES CORRESPONDANCES

III.1. Les tableaux de données

Le but de l'Analyse Factorielle des Correspondances (AFC) est de détecter des liens entre variables qualitatives, et de positionner les individus par rapport à ces liens.

On considère par exemple un ensemble de 18282 individus pour lesquels on connaît la CSP (modalités agriculteur AGRI, cadre supérieur CADR, inactif INAC, et ouvrier OUVR) et le choix de l'hébergement pour les vacances HEB (modalités camping CAMP, HOTEL, location LOCA, et résidence secondaire RESI). Le tableau des données brutes serait de la forme :

individu	CSP	HEB
1	OUVR	CAMP
2	INAC	CAMP
3	AGRI	HOTEL
…	…	…
18281	INAC	RESI
18282	CADR	LOCA

Cependant, les identifications des individus ne nous important peu dans l'étude de ce lien, on préfèrera présenter ces données sous forme de données groupées :

en colonnes ("déplié") :

CSP	HEB	effectif
AGRI	CAMP	239
AGRI	HOTEL	155
AGRI	LOCA	129
AGRI	RESI	0
CADR	CAMP	1003
CADR	HOTEL	1556
CADR	LOCA	1821
CADR	RESI	1521
INAC	CAMP	682
INAC	HOTEL	1944
INAC	LOCA	967
INAC	RESI	1333
OUVR	CAMP	2594
OUVR	HOTEL	1124
OUVR	LOCA	2176
OUVR	RESI	1038

ou sous forme de tableau de contingence :

CSP\HEB	CAMP	HOTEL	LOCA	RESI	Total
AGRI	239	155	129	0	523
CADR	1003	1556	1821	1521	5901
INAC	682	1944	967	1333	4926
OUVR	2594	1124	2176	1038	6932
Total	4518	4779	5093	3892	18282

Dans cet exemple, le but de l'AFC sera de représenter les éventuels liens entre la CSP et le type d'hébergement choisi HEB.
Cette étude de lien peut se passer de l'analyse par AFC, via le calcul de la statistique du khi-deux, ou le calcul des profils-lignes et des profils-colonnes. D'autres méthodes d'étude de ce type de lien existent.
Cependant, l'avantage qu'a l'AFC par rapport à ces méthodes est la hiérarchisation des différents types de lien, la représentation, s'il y a lieu, des individus par rapport à ces liens, et des représentations graphiques qui facilitent la communication de l'information.

III.2. L'étude des liens entre deux variables qualitatives sans AFC

Rappelons les deux principales méthodes citées plus haut.

III.2.1. La statistique du khi-deux et le test associé

On appelle effectifs observés (EFO) les effectifs du tableau de contingence. Les totaux des lignes et des colonnes s'appellent les marges. Pour calculer la statistique du khi-deux, il faut calculer les effectifs théoriques (EFT) en faisant les produits des marges divisés par l'effectif total général n.

Effectifs théoriques :					
	CAMP	HOTEL	LOCA	RESI	Total
AGRI	129,248	136,715	145,697	111,340	523
CADR	1458,304	1542,549	1643,901	1256,246	5901
INAC	1217,354	1287,679	1372,285	1048,681	4926
OUVR	1713,094	1812,057	1931,117	1475,733	6932
Total	4518	4779	5093	3892	18282

La statistique du khi-deux est alors la somme des valeurs (EFO-EFT)²/EFT. Ce calcul donne la valeur 2067,911 pour ces données. Le test du khi-deux consiste à comparer cette valeur observée du khi-deux (khi²obs) à une valeur théorique (khi²théo). dans cet exemple, khi²théo(5%)=16,919. la règle du test est alors la suivante :
si khi²obs<khi²théo(alpha), alors on ne peut pas affirmer la dépendance entre les variables. Dans notre exemple, on a khi²obs>khi²théo(alpha), donc on peut affirmer, avec un risque alpha=5% de se tromper, qu'il y a dépendance entre CSP et HEB. Autrement dit, le choix de l'hébergement pour les vacances semble dépendre de la CSP.
Une autre façon de faire le test est de comparer la p-value (calculée par les logiciels qui font le test) à 5%. Ici, p-value<0,0001.
La règle de test, équivalente à la première, est alors :
si p-value>alpha, alors on ne peut pas affirmer la dépendance entre les variables.

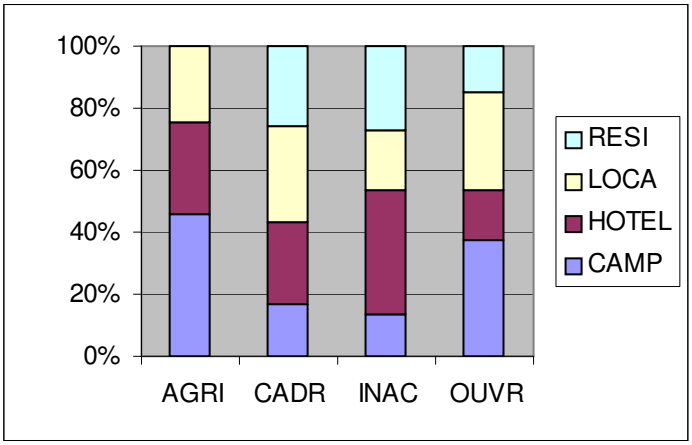
III.2.2. Les profils-ligne et colonne

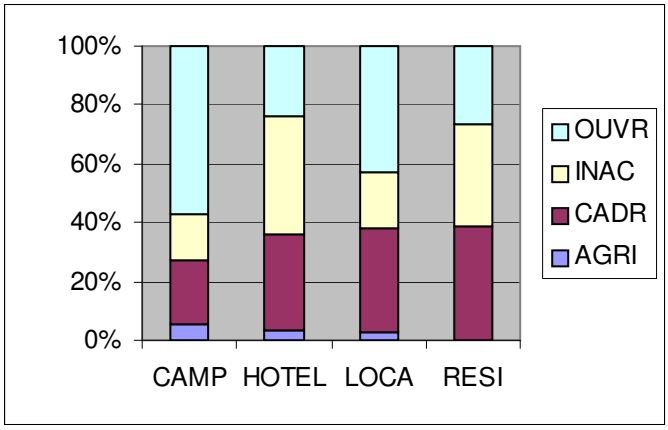
Ce sont les pourcentages (ou taux) d'individus d'une catégorie d'une des deux variables répartis selon les modalités de l'autre variable. On donne ci-dessous les profils-lignes et colonnes pour l'exemple.

Proportions / Ligne :					
	CAMP	HOTEL	LOCA	RESI	Total
AGRI	0,457	0,296	0,247	0,000	1
CADR	0,170	0,264	0,309	0,258	1
INAC	0,138	0,395	0,196	0,271	1
OUVR	0,374	0,162	0,314	0,150	1
Total	0,247	0,261	0,279	0,213	1

Proportions / Colonne :					
	CAMP	HOTEL	LOCA	RESI	Total
AGRI	0,053	0,032	0,025	0,000	0,029
CADR	0,222	0,326	0,358	0,391	0,323
INAC	0,151	0,407	0,190	0,342	0,269
OUVR	0,574	0,235	0,427	0,267	0,379
Total	1	1	1	1	1

Ils permettent de décrire les liens entre variables.
Par exemple, on voit qu'il y a 24,7% d'individus qui choisissent CAMP sur l'ensemble. Dans la catégorie AGRI, il y en a 45,7%. Cela signifie que les agriculteurs choisissent plus souvent le camping que les individus des autres CSP.
De même, dans la catégorie INAC, seulement 13,8% choisissent CAMP. Donc les inactifs choisissent moins souvent le camping que l'ensemble. Par une analyse de tous les profils-ligne, on décrirait ainsi les liens (préférences/non préférences) entre les CSP et les HEB choisis.
Une analyse des profils-colonne donnerait une même conclusion, mais dite différemment.
Ces profils font souvent l'objet de représentations graphiques :





III.2.3. Autre indicateur : les significativités par case

Elles permettent, dans un lien entre deux variables, de savoir où se situent les liens les plus significatifs.
Dans le tableau ci-contre, il y a le signe à apposer entre EFO et EFT (< signifie que EFO<EFT, et > signifie que EFO>EFT). Les signes sont en gras et surlignés si cette différence est significative, c'est-à-dire suffisamment grande pour mériter d'être mentionnée.

	CAMP	HOTEL	LOCA	RESI
AGRI	**>**	>	<	**<**
CADR	**<**	>	**>**	**>**
INAC	**<**	**>**	**<**	**>**
OUVR	**>**	**<**	**>**	**<**

On peut voir par exemple que le choix de l'hôtel est moins spécifique à telle ou telle CSP que celui du camping ou de la résidence secondaire.

III.3. l'AFC

Elle consiste à faire une ACP bien choisie du tableau des profils-ligne, ce qui est équivalent à l'ACP du tableau des profils-colonne. Le vocabulaire employé sera donc assez voisin de celui de l'ACP. Cependant, leur calcul et usage diffèrera.
Les résultats, pour un tableau de contingence à r lignes et c colonnes est le suivant.

III.3.1. Les valeurs propres

Elles sont au maximum au nombre de min(r,c)-1.
Dans l'exemple, r=c=4, donc nous aurons 3 axes.
Leur valeur étant toujours inférieure à 1, la règle de Kaiser pour choisir le nombre d'axes doit toujours être adaptée, en choisissant les valeurs propres supérieures à la moyenne.

Valeurs propres et pourcentages d'inertie :			
	F1	F2	F3
Valeur propre	0,098	0,014	0,001
Les lignes dépendent des colonnes (%)	86,855	12,256	0,889
% cumulé	86,855	99,111	100,000

Pour les autres méthodes de choix du nombre d'axes à interpréter, les règles sont les mêmes que pour l'ACP.
Dans cet exemple, comme il est recommandé de choisir au moins 2 axes, et qu'il y en a 3 en tout, le choix se porte sur 2 axes quelle que soit la méthode.

Propriété. La statistique du khi-deux égale la somme des valeurs propres multipliée par n.

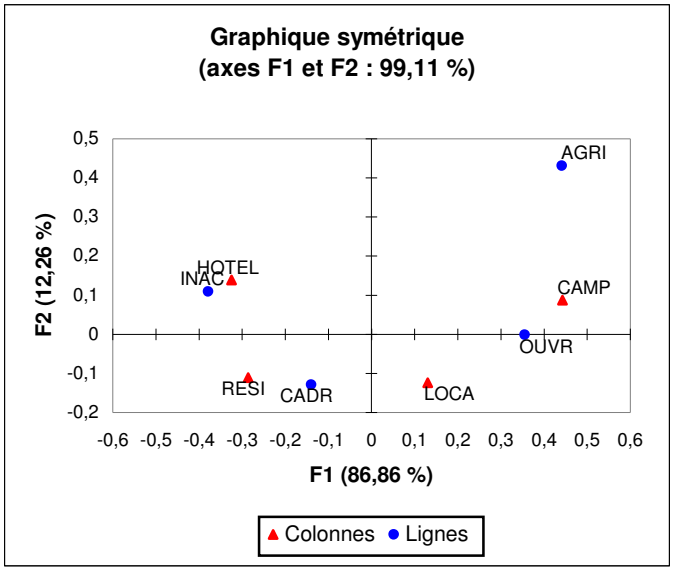
Cela fait pour cet exemple la relation suivante : 18282*(0,098+0,014+0,001)=2067,911.

III.3.2. Les coordonnées des modalités et leur représentation graphique

Une des particularités de l'AFC par rapport à l'ACP est de pouvoir représenter sur un même graphique les lignes et les colonnes du tableau.

Coordonnées principales (lignes) :			
	F1	F2	F3
AGRI	0,441	0,431	0,137
CADR	-0,140	-0,129	0,027
INAC	-0,379	0,109	-0,020
OUVR	0,355	-0,001	-0,019

Coordonnées principales (colonnes)			
	F1	F2	F3
CAMP	0,443	0,088	-0,022
HOTEL	-0,325	0,139	0,019
LOCA	0,130	-0,124	0,036
RESI	-0,286	-0,110	-0,045



Ces coordonnées sont centrées, comme en ACP, mais en tenant compte de la pondération de chaque modalité par son effectif total.
Exemple avec la dimension 1 des coordonnées des lignes :
(523*0,441+5901*(-0,140)+4926*
(-0,379)+6932*0,355)/18282=0.
Cela entraîne le fait qu'il y a toujours des modalités de part et d'autre de chaque axe (coordonnées négatives et positives).

La variance de chaque axe est aussi, comme en ACP, égale à la valeur propre de l'axe. Cette variance est, comme pour la moyenne pondérée. Cela donne la relation, par exemple :
(523*0,441²+5901*(-0,140)²+4926*(-0,379)²+6932*0,355²)/18282=0,098.

IV.3.3. Les aides à l'interprétation : contributions et cosinus carrés

Contributions (lignes) :				Cosinus carrés (lignes) :		
	F1	F2	F3	F1	F2	F3
AGRI	0,057	0,383	0,531	0,488	0,465	0,047
CADR	0,064	0,385	0,228	0,532	0,449	0,019
INAC	0,393	0,232	0,105	0,921	0,077	0,003
OUVR	0,486	0,000	0,135	0,997	0,000	0,003

Contributions (colonnes) :				Cosinus carrés (colonnes)		
	F1	F2	F3	F1	F2	F3
CAMP	0,494	0,137	0,122	0,960	0,038	0,002
HOTEL	0,281	0,366	0,092	0,842	0,155	0,003
LOCA	0,048	0,310	0,364	0,504	0,457	0,039
RESI	0,178	0,187	0,422	0,852	0,127	0,021

Comme en ACP, pour chaque tableau, la somme des contributions d'une colonne égale 1 (ou 100%). Ce qui entraîne une moyenne des contributions égale à 1/r pour les modalités ligne, et 1/c pour les modalités colonne.
Dans notre exemple, r=c=4, donc les moyennes des contributions sont égales à ¼=0,25.
De même, les sommes des cosinus carrés pour une ligne égalent 1. Donc la moyenne des cosinus carrés égale 1/nombre d'axes total = 1/3=0,333 dans l'exemple.
Citons, en indiquant le signe des coordonnées, les modalités qui contribuent pour plus que la moyenne à l'inertie des axes 1 et 2.

signe coordonnée	-	+
axe 1	INAC, HOTEL	OUVR, CAMP
axe 2	CADR, LOCA	AGRI, HOTEL

Donc l'axe 1 est basé sur l'opposition entre d'une part le fait que les inactifs choisissent plus souvent l'hôtel et moins souvent le camping, alors que c'est l'inverse chez les ouvriers.
L'axe 2 est construit sur la base du fait que pour un sous-groupe d'individus, les cadres vont plus souvent en location et moins à l'hôtel, et en cela ils s'opposent aux agriculteurs, qui ont, pour un sous-groupe d'entre eux, un comportement inverse.
Citons de même les modalités qui sont reconstituées pour plus que la moyenne sur les axes 1 et 2 :

signe coordonnée	-	+
axe 1	CADR, INAC, HOTEL, RESI	AGRI, OUVR, CAMP, LOCA
axe 2	CADR, LOCA	AGRI

Sur l'axe 1, les cadres sup ont un comportement voisin des inactif, choisissant plus volontiers l'hôtel, ou la résidence secondaire. A l'opposé, les agriculteurs ont un comportement voisin des ouvriers, choisissant plus le camping ou la location.
Sur l'axe 2, la modalité HOTEL disparaît par rapport à ce que nous avons dit avec les contributions. Cela signifie que le sous-groupe concerné par l'axe 2 comporte moins de personnes choisissant l'hôtel que les autres modes d'hébergement cités.

III.3.4. Quelques plus de l'AFC

Reprenons le graphique des modalités. Il permet de voir que les modalités sont disposées en arc de cercle. Ce phénomène est connu sous le nom d'*effet Guttman*. Il apparaît quand un ordre sous-tend les modalités. Dans notre exemple, l'ordre est le suivant : hotel, inac, resi, cadr, loca, ouvr, camp, agri. On peut sans trop s'avancer soupçonner un ordre dû au coût de ces hébergements, et au moyen financier consacré par chaque type de CSP. D'autre part, un regard plus attentif permet de

constater que, si l'axe 1 est celui des moyens financiers consacrés aux vacances, l'axe 2 est plutôt celui du type de vacances choisi. Les modes d'hébergement côté négatif (RESI, LOCA) sont plutôt de type sédentaire, ceux côté positif sont plutôt destinés à des vacances itinérantes. Ces constatations, dont l'explication est grandement aidée par la représentation graphique, sont particulières à l'AFC, et auraient été plus difficilement décelables par une autre analyse de ce tableau.
On peut facilement généraliser la méthode AFC par l'AFCM, qui consiste à étudier plusieurs variables qualitatives, et à en faire une représentation graphique comme en AFC.