

\*\*\*\*\*



## Fouille de données

### Groupe : 11

### Participants :

- BIAKOTA BOMBIA HERBERT CEPHAS
- MILORME PIERRE RUBENS

## TP2 : description des données

### Introduction

D'une manière générale, la fouille de données (data mining) est perçue comme ensemble des techniques d'exploration de données permettant d'extraire des connaissances sous forme de modèles de description. Ainsi divers outils sont utilisés pour faciliter cette exploration parmi lesquels nous avons **Tanagra, Weka, R, SPSS etc.** C'est dans ce contexte qu'il nous est demandé d'utiliser Tanagra en vue de décrire les données utilisées dans ce TP. Pour ce fait nous allons d'abord télécharger un jeu de donnée sur le site <http://archive.ics.uci.edu/ml/datasets.html> afin de répondre à cette perspective.

Le jeu de données que nous avons choisi est : FORESTFIRE.csv. Ce jeu de données relate les données sur le feu de forêt de 1987 à 2005 dans le parc de Montesinho en Portugal. Ainsi pour notre travail, nous avons fixé comme objectif de prédire la surface des zones susceptibles d'être brûlées par des feux de forêt au sein de ce parc

### 1- Description des données

**Notre jeu de données comprend 13 variables :**

- **X** Coordonnées spatiales de l'axe X dans le parc de Montesinho carte: 1 à 9
- **Y** Coordonnées géographiques de l'axe des Y et Y dans le parc de Montesinho carte: 2 à 9
- **month (mois)** mois de l'année (janvier à décembre)
- **day (jour)** jour de la semaine (du lundi au dimanche)
- **FFMC** (18,7 à 96,20) : Le Code d'humidité des combustibles fins (FFMC) est un indice numérique de la teneur en humidité de la litière et d'autres combustibles fins durcis. Ce code est un indicateur de la facilité relative d'allumage et de l'inflammabilité du combustible fin.
- **DMC** (1.1 à 291.3) : Le Duff Moisture Code (DMC) est un indice numérique de la teneur en humidité moyenne des couches organiques compactées de profondeur modérée. Ce code donne une indication de la consommation de carburant dans des couches de duff modérées et des matériaux ligneux de taille moyenne.
- **DC** (7.9 à 860.6) : Le Code de sécheresse (DC) est une estimation numérique de la teneur moyenne en humidité des couches organiques profondes et compactes. Ce code est un indicateur utile des effets saisonniers de la sécheresse sur les combustibles forestiers et de la quantité de fumée dans les couches profondes et les grosses bûches.
- **ISI** (0,0 à 56,10) : L'indice initial de propagation (ISI) est une estimation numérique du taux attendu d'écart d'incendie. Il combine les effets du vent et de la FFMC sur le taux de propagation sans l'influence de quantités variables de carburant.
- **temp** : Température en degrés Celsius **2.2 à 33.30**

- **RH** : Humidité relative en% est le rapport entre la pression partielle de vapeur d'eau "p<sub>v</sub>" et la pression de saturation de la vapeur d'eau "p<sub>vs</sub>" : 15,0 à 10
- **wind (Vent)** Vitesse du vent en km / h: 0.40 à 9.40
- **rain (pluie)** pluie extérieure en mm / m2: 0.0 à 6.4
- **area (région)** la superficie brûlée de la forêt (en Ha) 0,00 à 1090,84 (Cette variable de sortie)

Il existe 2 types de variables dans notre jeu de données :

### 1- Variables qualitatives :

- **Variables ordinales** : day, month

### 2- Variables quantitatives :

- **Variables discrètes**: d'après tanagra, nous avons month et day
- **Variables continues**: X, Y, FPMC, DMC, DC, ISI, temp, RH, wind, rain, area

TP		
Dataset (FORESTFIRE.xls)		
Dataset description		
13 attribute(s) 517 example(s)		
Attribute	Category	Informations
X	Continue	-
Y	Continue	-
month	Discrete	12 values
day	Discrete	7 values
FFMC	Continue	-
DMC	Continue	-
DC	Continue	-
ISI	Continue	-
temp	Continue	-
RH	Continue	-
wind	Continue	-
rain	Continue	-
area	Continue	-
Computation time : 0 ms. Created at 2017-01-13 11:07:01		

Il existe aussi 517 exemples (lignes) dans notre jeu de données

TP		X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
Dataset (FORESTFIRE.xls)		7	5	mar	fri	86,2	26,2	94,3	5,1	9,2	51	6,7	0	0
View dataset 1		7	4	oct	tue	90,6	35,4	669,1	6,7	18	33	0,9	0	0
		7	4	oct	sat	90,6	43,7	686,9	6,7	14,6	33	1,3	0	0
		8	6	mar	fri	91,7	33,3	77,5	9	8,3	97	4	0,2	0
		8	6	mar	sun	89,3	51,3	102,2	9,6	11,4	99	1,8	0	0
		8	6	aug	sun	92,3	85,3	488	14,7	22,2	29	5,4	0	0
		8	6	aug	mon	92,3	88,9	495,6	8,5	24,1	27	3,1	0	0
		8	6	aug	mon	91,5	145,4	608,2	10,7	8	86	2,2	0	0
		8	6	sep	tue	91	129,5	692,6	7	13,1	63	5,4	0	0
		7	5	sep	sat	92,5	88	698,6	7,1	22,8	40	4	0	0
		7	5	sep	sat	92,5	88	698,6	7,1	17,8	51	7,2	0	0
		7	5	sep	sat	92,8	73,2	713	22,6	19,3	38	4	0	0
		6	5	aug	fri	63,5	70,8	665,3	0,8	17	72	6,7	0	0
		6	5	sep	mon	90,9	126,5	686,5	7	21,3	42	2,2	0	0
		6	5	sep	wed	92,9	133,3	699,6	9,2	26,4	21	4,5	0	0
		6	5	sep	fri	93,3	141,2	713,9	13,9	22,9	44	5,4	0	0
		5	5	mar	sat	91,7	35,8	80,8	7,8	15,1	27	5,4	0	0
		8	5	oct	mon	84,9	32,8	664,2	3	16,7	47	4,9	0	0
		6	4	mar	wed	89,2	27,9	70,8	6,3	15,9	35	4	0	0
		6	4	apr	sat	86,3	27,4	97,1	5,1	9,3	44	4,5	0	0
		6	4	sep	tue	91	129,5	692,6	7	18,3	40	2,7	0	0

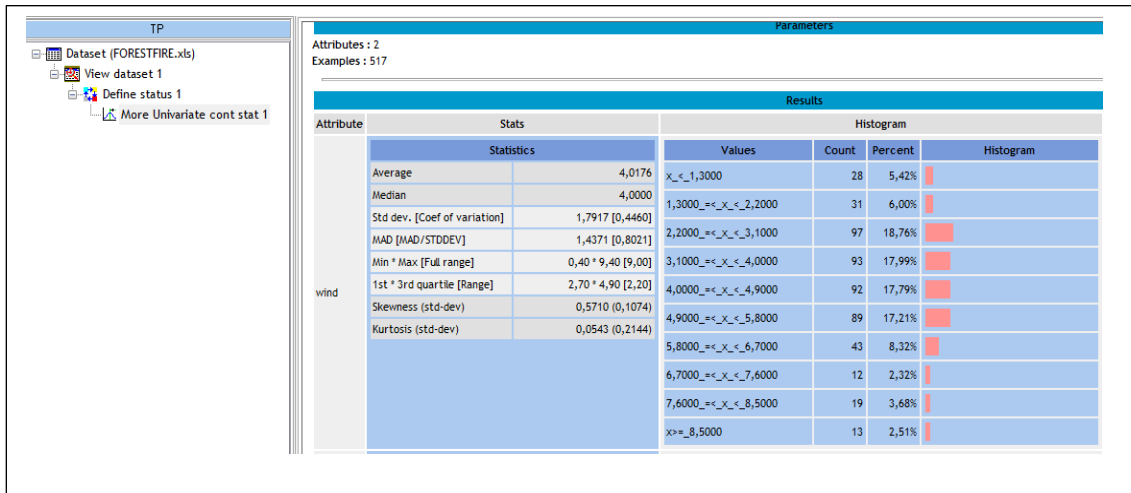
## TP3 : Analyse descriptive d'une ou de deux variables

Dans cet exercice nous avons utilisé premièrement 2 variables continues

Nous allons maintenant mettre **wind** et **rain** en entrée

Resultat :

### Variable wind



### Variable rain

	Statistics		Values	Count	Percent	Histogram
rain	Average	0,0217	x_<_0,6400	512	99,03%	
	Median	0,0000	0,6400_=<_x_<_1,2800	3	0,58%	
	Std dev. [Coef of variation]	0,2960 [13,6617]	1,2800_=<_x_<_1,9200	1	0,19%	
	MAD [MAD/STDDEV]	0,0427 [0,1441]	1,9200_=<_x_<_2,5600	0	0,00%	
	Min * Max [Full range]	0,00 * 6,40 [6,40]	2,5600_=<_x_<_3,2000	0	0,00%	
	1st * 3rd quartile [Range]	0,00 * 0,00 [0,00]	3,2000_=<_x_<_3,8400	0	0,00%	
	Skewness (std-dev)	19,8163 (0,1074)	3,8400_=<_x_<_4,4800	0	0,00%	
	Kurtosis (std-dev)	421,2960 (0,2144)	4,4800_=<_x_<_5,1200	0	0,00%	
			5,1200_=<_x_<_5,7600	0	0,00%	
			x>= 5,7600	1	0,19%	

indication	Description
average	Moyenne
median	Médiane
Std.Dev. [Coef of variation]	Ecart type (échantillon) et coefficient de variation (rapport entre l'écart type et la moyenne, permet la comparaison de la dispersion de variables mesurées sur des unités différentes)
MAD [MAD / STDDEV]	Ecart absolu moyen. Rapport entre l'écart absolu moyen et l'écart type. Lorsque la distribution est normale, ce rapport est proche de 0.8.
Min, Max [Full Range]	Minimum, maximum, étendue

1st * 3rd quartile [Range]	1 <sup>er</sup> et 3 <sup>ème</sup> quartile ; intervalle inter quartile
Skewness (std dev)	Coefficient d'asymétrie et son écart type. Lorsque la distribution est normale, skewness = 0
Kurtosis (std dev)	Coefficient d'aplatissement et son écart type. Lorsque la distribution est normale, kurtosis = 0

Voici les données statistiques relatives aux deux variables continues que nous avons obtenues :

- **wind**

Moyenne : 4,0176

Médiane : 4,000

Coefficient de variance : 1,7917

Valeur Minimale : 0,40

Valeur Maximale : 9,40

Coefficient d'asymétrie : 0,05710

Coefficient d'aplatissement : 0,0543

Nous remarquons qu'il y'a un léger aplatissement et une légère asymétrie de distribution de variable **wind** par rapport à la normale

- **rain :**

Moyenne : 0,0217

Médiane : 0000

Coefficient de variance : 0,2960

Valeur minimale : 0,00

Valeur maximale : 6,40

Coefficient d'asymétrie : 19,8167

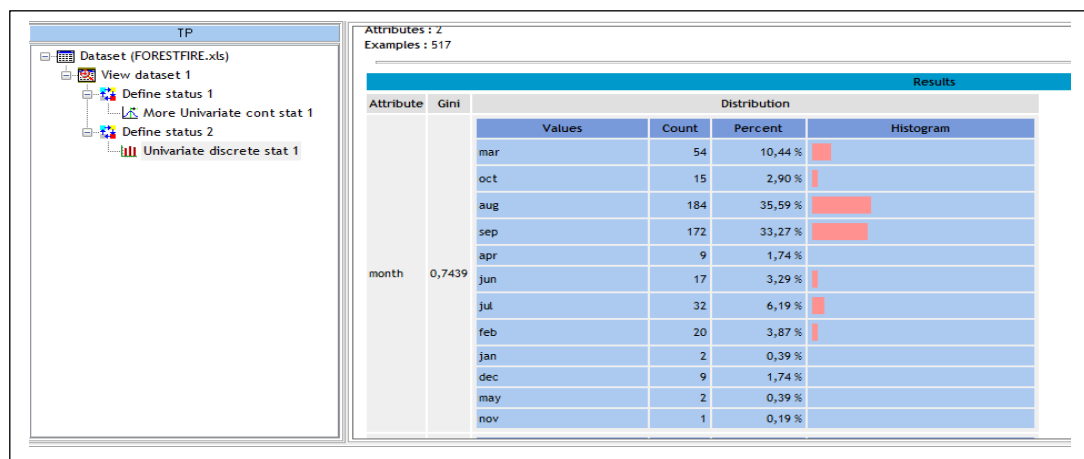
Coefficient d'aplatissement : 421,2960

Nous remarquons qu'il y'a un fort aplatissement et une forte asymétrie de la distribution de variable rain par rapport à la normale ce qui explique une carence de pluie durant la période 2005

**Ensuite nous avons choisie de faire le même traitement pour 2 variables discrètes**








Nous allons maintenant mettre month et day en Input

**Variable month**



D'après les résultats nous remarquons que le mois d'Août et de septembre 2005 sont beaucoup plus touchés par le feu de forêt dans le **parc de Montesinho** car il en résulte 35% de feu de forêt durant le mois d'Août, 33,27% de feu de forêt durant le mois de Septembre. Nous pouvons aussi observer une absence presque totale de la présence du feu de forêt durant le mois Janvier, Décembre, Mai, Novembre.

## Variable day

day	0,8522	Values	Count	Percent	Histogram
		fri	85	16,44 %	
		tue	64	12,38 %	
		sat	84	16,25 %	
		sun	95	18,38 %	
		mon	74	14,31 %	
		wed	54	10,44 %	
		thu	61	11,80 %	

D'après le tableau de fréquence ci-dessus, nous pouvons remarquer que les feux de forêt ont été les plus fréquents pendant les jours de dimanche durant la période de temps de 1987 à 2005 : soit un pourcentage de 18,38% du total de cas d'incendies pour cette même période. Cependant cette valeur n'est pas trop significative si l'on tient compte de sa faible démarcation par rapport aux autres valeurs des autres jours de la semaine : le facteur *jour de la semaine* par conséquent n'a pas été d'une grande influence dans le cas des feux qui ont été déclarés.

## Corrélation entre paire de variables

### - Variables quantitatives

Linear correlation 1					
Parameters					
Cross-tab parameters					
Sort results	non				
Input list	Target (Y) and input (X)				
Results					
Y	X	r	r <sup>2</sup>	t	Pr(> t )
area	X	0,0634	0,0040	1,4413	0,1501
area	DMC	0,0730	0,0053	1,6609	0,0973
area	temp	0,0978	0,0096	2,2311	0,0261
area	RH	-0,0755	0,0057	-1,7187	0,0863
area	wind	0,0123	0,0002	0,2795	0,7799

L'analyse des données nous indique que la variable la plus corrélée avec « area » est « temp » avec  $r=0,0978$  ce qui est bien logique car plus la température est élevée, plus il est probable que la surface brûlée soit grande. La seconde variable la plus corrélée à « area » est « RH » et cette corrélation est négative, ce qui se justifie bien car plus l'humidité est basse, plus la surface brûlée est potentiellement grande. La troisième variable la plus corrélée à « area » est DMC ce qui est aussi normal car si la teneur moyenne en eau dans les couches organique de l'humus est élevée, cela est possible de favoriser pas progression du feu sur une surface grande

Nous évaluons la corrélation entre les différentes paires de variables d'entrées.

Linear correlation 1						
Cross-tab parameters			Parameters			
Sort results	non					
Input list	Cross-input (Y x X)					
						Results
Y	X	r	r <sup>2</sup>	t	Pr(> t )	
X	DMC	-0,0484	0,0023	-1,0993	0,2722	
X	temp	-0,0513	0,0026	-1,1648	0,2447	
X	RH	0,0852	0,0073	1,9411	0,0528	
X	wind	0,0188	0,0004	0,4267	0,6698	
DMC	temp	0,4696	0,2205	12,0704	0,0000	
DMC	RH	0,0738	0,0054	1,6793	0,0937	
DMC	wind	-0,1053	0,0111	-2,4040	0,0166	
temp	RH	-0,5274	0,2781	-14,0867	0,0000	
temp	wind	-0,2271	0,0516	-5,2924	0,0000	
RH	wind	0,0694	0,0048	1,5790	0,1150	

D'après la propriété de  $r$  (coefficient de corrélation) :

- Si  $r$  est égal à 0 il n'existe pas une corrélation linéaire entre Y et X
- si  $r$  est proche de 0, l'intensité de la corrélation linéaire entre Y et X est faible
- si  $r$  est proche de -1 ou 1, l'intensité de la corrélation linéaire entre Y et X est négativement ou positivement forte.

**Dans notre cas :**

La corrélation linéaire entre Y (temp) et X (wind) est négativement faible car coefficient de corrélation  $r = -0,2271$  proche de 0.

La corrélation linéaire entre Y (temp) et X (RH) est négativement forte car coefficient de corrélation  $r = -0,5274$  proche de 0.

La corrélation linéaire entre Y (DCM) et X (wind) est négativement faible car coefficient de corrélation  $r = -0,1053$  proche de 0.

La corrélation linéaire entre Y (DCM) et X (RH) est positivement faible car coefficient de corrélation  $r = -0,0738$  proche de 0.

La corrélation linéaire entre Y (DCM) et X (temp) est positivement faible car coefficient de corrélation  $r = -0,4696$  proche de 0.

La corrélation linéaire entre Y (X) et X (RH) est positivement faible car coefficient de corrélation  $r = -0,0852$  proche de 0.

Les informations présentées nous indiquent qu'il existe une corrélation entre certaines paires de variables, ce qui signifie qu'elles ont un effet similaire sur le comportement de la variable de sortie.

### - Variables qualitatives

Pour ce fait, nous devons évaluer le lien existant entre le mois (variable month) d'une part et le jour (variable day) de la semaine d'autre part, en établissant le tableau de contingence suivant avec le composant CONTINGENCY CHI-SQUARE de l'onglet NONPARAMETRIC STATISTICS

Contingency Chi-Square 1												
Parameters												
Cross-tab parameters												
Sort results		non										
Input list		Target (Row) and input (Column)										
Additional information		0										
Contribution threshold		2,0										
Results												
Row (Y)	Column (X)	Statistical indicator		Cross-tab								
month	day	Stat	Value	fri	tue	sat	sun	mon	wed	thu	Sum	
		d.f.	66	mar	11	5	10	7	12	4	5	54
		Tschuprow's t	0,123670	oct	1	2	3	3	4	2	0	15
		Cramer's v	0,143905	aug	21	28	29	40	15	25	26	184
		Phi²	0,124252	sep	38	19	25	27	28	14	21	172
		Chi² (p-value)	64,24 (0,5384)	apr	1	0	1	3	1	1	2	9
		Lambda	0,090090	jun	3	0	2	4	3	3	2	17
		Tau (p-value)	0,0204 (0,0002)	jul	3	6	8	5	4	3	3	32
		U(R/C) (p-value)	0,0384 (0,4567)	feb	5	2	4	4	3	1	1	20
				jan	0	0	1	1	0	0	0	2
				dec	1	1	0	1	4	1	1	9
				may	1	0	1	0	0	0	0	2
				nov	0	1	0	0	0	0	0	1
				Sum	85	64	84	95	74	54	61	517

D'après le résultat Le mois d'aout (aug) est celui où ont été effectués le plus de prélèvements, exactement 184, parmi lesquels 40 ont été effectués le dimanche (sun) ;

- Une partie des prélèvements sur l'année a été aussi effectuée au mois de septembre (sept).

En conclusion, ces deux mois correspondent à la période estivale dans la zone de Montesinho au nord-est du Portugal.

## TP4 : Analyse en Composantes Principales et Analyse Factorielle des correspondances

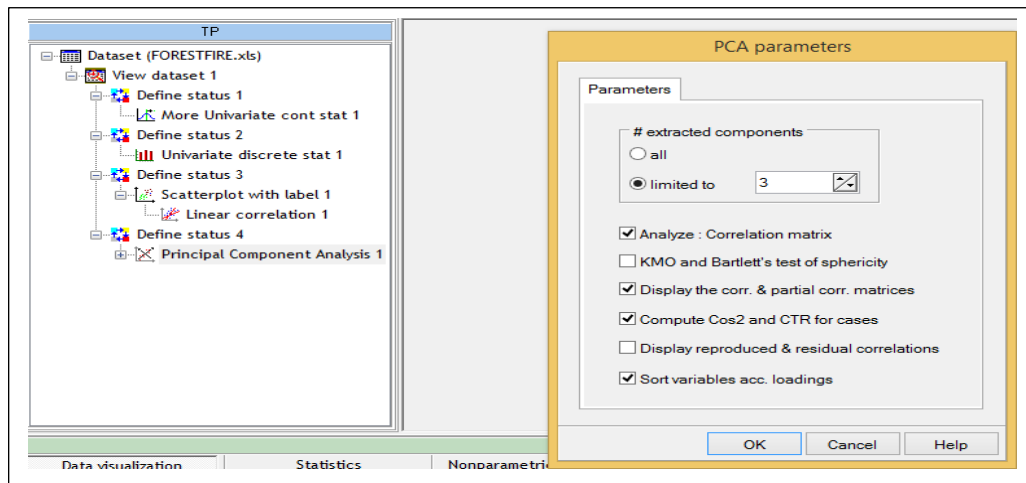
### 1- Analyse en Composantes Principales

L'analyse en composantes principales (ACP) est une technique exploratoire très populaire. Il s'agit de résumer l'information contenue dans un fichier en un certain nombre de variables synthétiques et de combinaisons linéaires des variables originelles.

Nous commençons tout d'abord en mettant en *Input* les données suivantes:FFMC, DCM, DC, ISI

Ensuite, nous ajoutons l'opérateur *Principal Component Analysis* du groupe *Factorial Analysis* et dans les paramètres de l'opérateur, nous prenons:

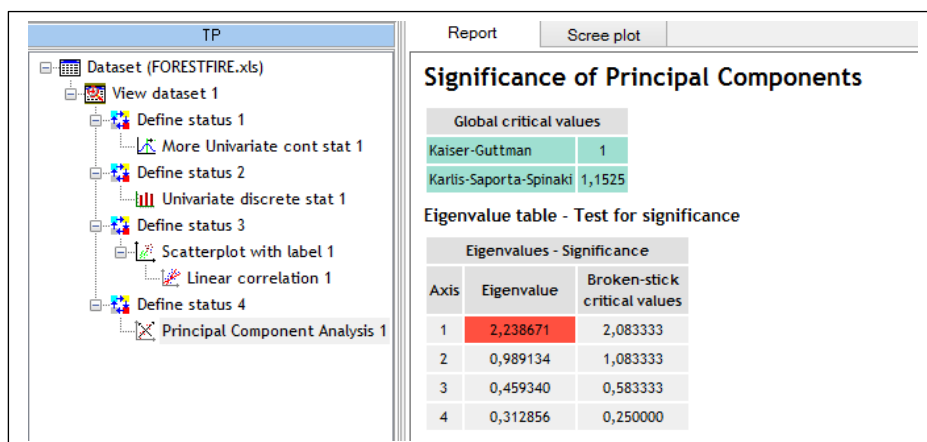
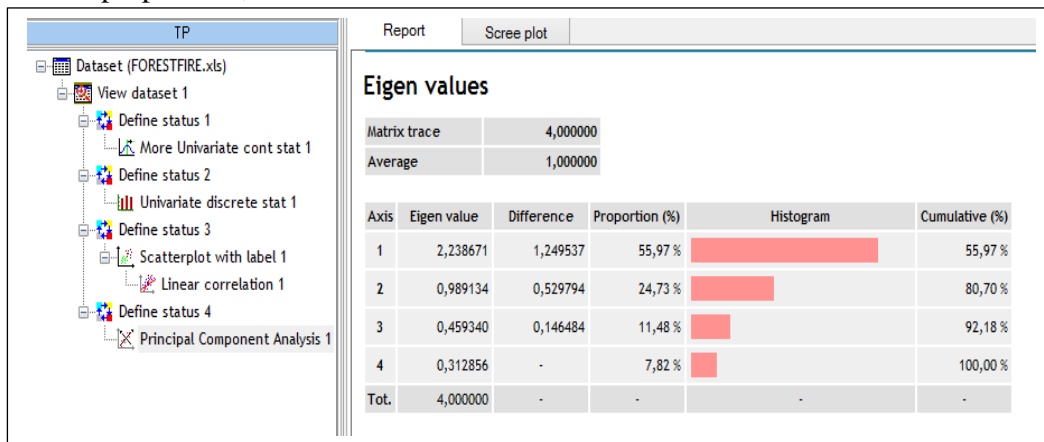




Nous exécutons le composant pour obtenir les résultats ci-dessous :

- **trace de la matrice 4.000 avec la moyenne de 1.00**

Nous remarquons également que l'axe 1 à lui seul regroupe 55,97% d'inertie (proportion) et une valeur propre de 2,238671



Ensuite nous avons la corrélation variable et axe. La seconde partie des résultats indique la corrélation et le  $\text{COS}^2$  --en % et % cumulé -- des variables avec les axes factoriels

### Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2		Axis_3	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
DMC	-0,81463	66 % (66 %)	0,40904	17 % (83 %)	0,07654	1 % (84 %)
DC	-0,76842	59 % (59 %)	0,51168	26 % (85 %)	0,02445	0 % (85 %)
FFMC	-0,74032	55 % (55 %)	-0,44284	20 % (74 %)	-0,50578	26 % (100 %)
ISI	-0,66068	44 % (44 %)	-0,60324	36 % (80 %)	0,44393	20 % (100 %)
Var. Expl.	2,23867	56 % (56 %)	0,98913	25 % (81 %)	0,45934	11 % (92 %)

En voyant le tableau ci-dessus nous pouvons en déduire :

La variable **DMC** est fortement corrélée négativement sur Axis\_1

La variable **DC** est fortement corrélée négativement sur Axis\_1 et positivement que Axis\_2

La variable **FFMC** est fortement corrélée négativement sur Axis\_1 et sur Axis\_3

La variable **ISI** est fortement corrélée négativement sur Axis\_1 et sur Axis\_2

Ensuite nous avons :

### Factor Score Coefficients

Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3
FFMC	90,6446805	5,5147697	-0,4947966	-0,4452682	-0,7462607
DMC	110,8723405	63,9845115	-0,5444614	0,4112756	0,1129367
DC	547,9400360	247,8261639	-0,5135744	0,5144793	0,0360765
ISI	9,0216634	4,5550654	-0,4415648	-0,6065473	0,6550105

Ensuite nous avons : la matrice de corrélation qui donne l'intensité de corrélation entre les variables. Nous pouvons en déduire du résultat ci-dessous que les parties colorées indiquent qu'il existe une forte corrélation positive entre les variables en question

TP
Dataset (FORESTFIRE.xls)
View dataset 1
Define status 1
More Univariate cont stat 1
Define status 2
Univariate discrete stat 1
Define status 3
Scatterplot with label 1
Linear correlation 1
Define status 4
Principal Component Analysis 1

Report
Scree plot

### Matrices

#### Correlations

	DMC	DC	FFMC	ISI
DMC	1,00000	0,68219	0,38262	0,30513
DC	0,68219	1,00000	0,33051	0,22915
FFMC	0,38262	0,33051	1,00000	0,53180
ISI	0,30513	0,22915	0,53180	1,00000

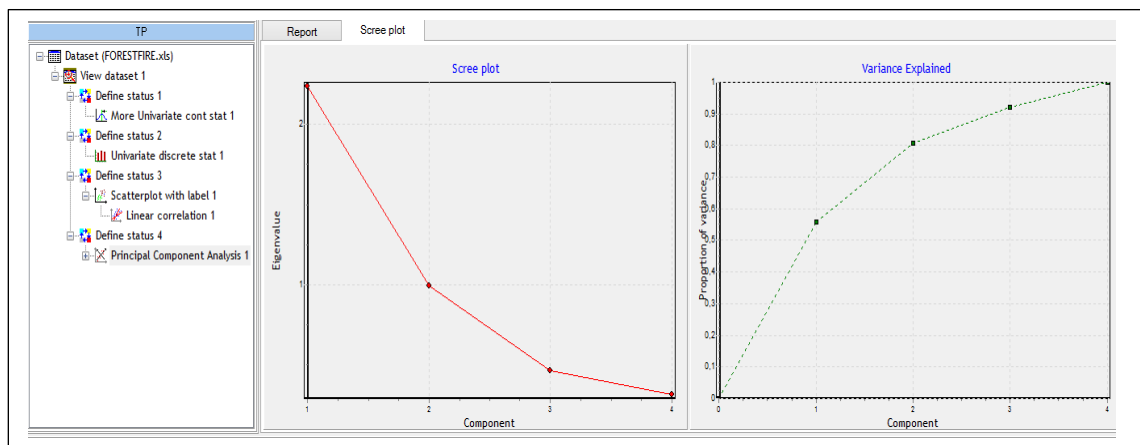
#### Partial Correlations Controlling all other Variables

	DMC	DC	FFMC	ISI
DMC	1,00000	0,63543	0,14599	0,11360
DC	0,63543	1,00000	0,10058	-0,02095
FFMC	0,14599	0,10058	1,00000	0,47135
ISI	0,11360	-0,02095	0,47135	1,00000

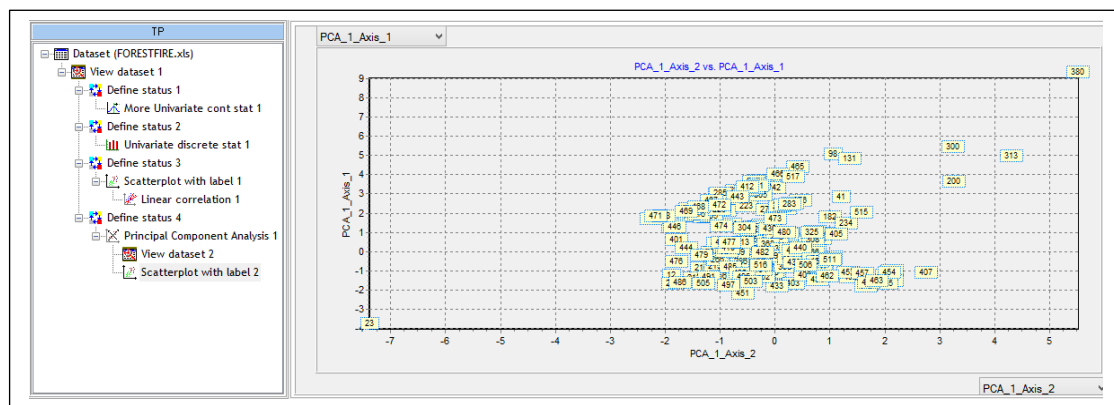
En effet, le composant ACP rajoute automatiquement une série de variables à l'ensemble de données. Il s'agit, pour chaque individu et pour chaque axe demandé, des projections sur les axes, des contributions et des  $\text{COS}^2$ .

TP		wind	rain	area	PCA 1_Axis	PCA 1_Axis	PCA 1_Axis	PCA 1_CTR	PCA 1_CTR	PCA 1_CTR	PCA 1_COS2	PCA 1_COS2	PCA 1_COS2
<div><div><div><div><div><div></div><div>Dataset (FORESTFIRE.xls)</div></div><div><div><div><div><div></div><div>View dataset 1</div></div><div><div><div><div><div></div><div>Define status 1</div></div><div><div><div><div><div></div><div>More Univariate cont stat 1</div></div></div></div><div><div><div><div><div></div><div>Define status 2</div></div><div><div><div><div><div></div><div>Univariate discrete stat 1</div></div></div></div><div><div><div><div><div></div><div>Define status 3</div></div><div><div><div><div><div></div><div>Scatterplot with label 1</div></div><div><div><div><div><div></div><div>Linear correlation 1</div></div></div></div><div><div><div><div><div></div><div>Define status 4</div></div><div><div><div><div><div></div><div>Principal Component Analysis 1</div></div></div></div><div><div><div><div><div></div><div>View dataset 2</div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div></div>	6,7	0	0	2,43954	-0,604921	-0,177961	0,514202	0,071957	0,013936	0,916628	0,0563608	0,00487788	
	0,9	0	0	0,620202	0,0791656	-0,449381	0,032342	0,00122554	0,0827807	0,2035	0,00381566	0,104004	
	1,3	0	0	0,512688	0,169468	-0,42614	0,0227104	0,00561604	0,0764678	0,156796	0,0171319	0,108326	
	4	0,2	0	1,5424	-1,58756	-0,351324	0,205548	0,474397	0,0519745	0,465568	0,074763	0,0241349	
	1,8	0	0	1,49522	-1,2767	0,08509	0,193164	0,318736	0,00380754	0,535185	0,490187	0,00216454	
	5,4	0	0	-0,357156	-1,17858	0,538674	0,0110214	0,271626	0,122187	0,068495	0,745865	0,15581	
	0	0	0	0,197484	-0,314077	-0,345415	0,00336965	0,012898	0,0502409	0,14676	0,371206	0,448977	
	2,2	0	0	-0,658121	0,0544768	0,195315	0,0374223	0,00050856	0,160638	0,849039	0,00581989	0,0747807	
	5,4	0	0	-0,29419	0,660557	-0,284856	0,00747784	0,0853247	0,0341696	0,13812	0,696339	0,129495	
	4	0	0	-0,09776670	0,271834	-0,545834	0,000825850	0,014498	0,125457	0,0121219	0,0937123	0,377842	
	7,2	0	0	-0,09776670	0,271834	-0,545834	0,000825850	0,014498	0,125457	0,0121219	0,0937123	0,377842	
	0	0	0	-1,53115	-1,88159	1,61842	0,20256	0,692314	1,10295	0,238521	0,360198	0,266485	
	6,7	0	0	3,30326	3,27254	2,47372	0,958242	2,09423	2,5015	0,394635	0,381092	0,113391	
	2,2	0	0	-0,247049	0,636684	-0,277508	0,00527336	0,072688	0,0324284	0,106819	0,709459	0,2134781	
	5,4	0	0	-0,72477	0,253156	-0,217883	0,0453858	0,0125323	0,0199904	0,788562	0,0962084	0,0712657	
	5,4	0	0	-1,31313	-0,324521	0,419966	0,148983	0,020594	0,074233	0,840342	0,0513247	0,0859136	
1	0	0	1,63061	-1,37485	-0,518989	0,229732	0,369626	0,11342	0,527749	0,375174	0,0534615		
6,9	0	0	1,52257	1,00619	-0,209411	0,200297	0,197585	0,0184661	0,510439	0,222478	0,0086579		
4	0	0	2,08828	-1,04479	-0,411785	0,376786	0,213459	0,0714029	0,750068	0,187753	0,0291653		

Ensuite nous avons la courbe de pourcentage d'inertie

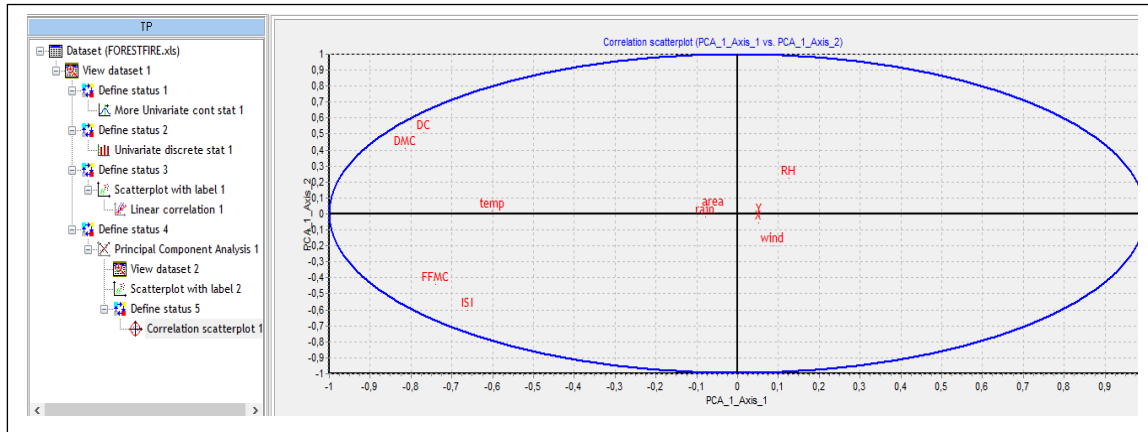


Nous constatons une démarcation des individus 23 et 380 du regroupement, ainsi qu'une légère démarcation des individus 300, 313, 200



### Le cercle de corrélation linéaire :

Il est une représentation graphique en 2D de variables à l'intérieur d'un cercle. Ainsi, pour commencer nous devons mettre en input toutes les variables continues et les premiers résultats de l'ACP en target



D'après la propriété de la corrélation :

- [-1 ; -0,5] la corrélation est négativement forte
- [-0,5 ; 0] la corrélation est négativement faible
- [0 ; 0,5] la corrélation est positivement faible
- [0,5 ; 1] la corrélation est positivement forte

Interprétation des résultats :

Sur l'axe PCA\_1\_Axis\_1 [-1 ; 1]

La corrélation de la variable **rain** est négativement faible ;  
La corrélation de la variable **wind** est négativement faible ;  
La corrélation de la variable **X** est positivement faible ;  
La corrélation de la variable **Y** est positivement faible ;  
La corrélation de la variable **RH** est positivement forte ;  
La corrélation de la variable **area** est négativement faible.  
La corrélation de la variable **temp** est négativement forte ;  
La corrélation de la variable **FFMC** est négativement forte ;  
La corrélation de la variable **ISI** est négativement forte ;  
La corrélation de la variable **DMC** est négativement forte ;  
La corrélation de la variable **DC** est négativement forte ;  
La corrélation de la variable **wind** est positivement faible ;

- Sur l'axe PCA\_1\_Axis\_2 [-1 ; 1]

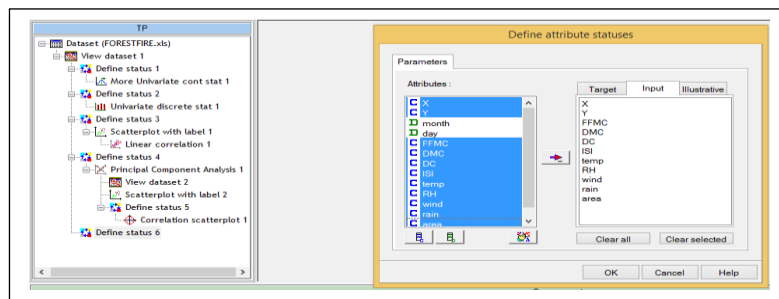
La corrélation de la variable **rain** est positivement faible ;  
La corrélation de la variable **wind** est positivement faible ;  
Il n'y pas de corrélation entre **X** et PCA\_1\_Axis2 ;  
La corrélation de la variable **Y** est positivement faible ;  
La corrélation de la variable **RH** est positivement forte ;

La corrélation de la variable **area** est positivement faible.  
 La corrélation de la variable **temp** est positivement faible;  
 La corrélation de la variable **FFMC** est négativement faible ;  
 La corrélation de la variable **ISI** est négativement forte ;  
 La corrélation de la variable **DMC** est positivement faible ;  
 La corrélation de la variable **DC** est positivement forte;

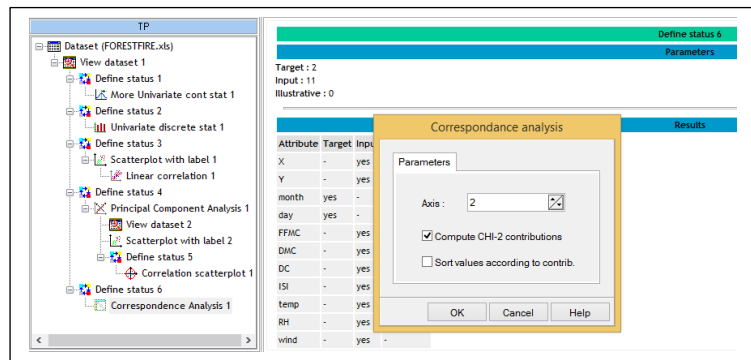
## 2- Analyse factorielle des correspondances

L'analyse factorielle des correspondances est une méthode statistique de réduction de dimension. Elle propose une vision synthétique de l'information intéressante d'un tableau de contingence. Son pouvoir de séduction repose en grande partie sur les représentations graphiques qu'elle propose. Elles nous permettent de situer facilement les similarités (dissimilarités) et les attractions (répulsions) entre les modalités.

Nous allons maintenant définir les 2 variables qualitatives en cible (month et day) et les autres variables continues en entrées

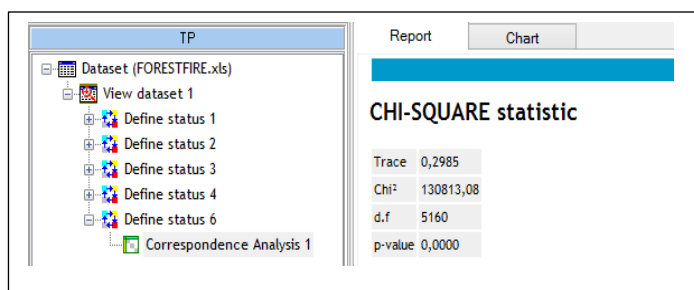


Ensuite nous allons paramétrer l'opérateur correspondance analysis



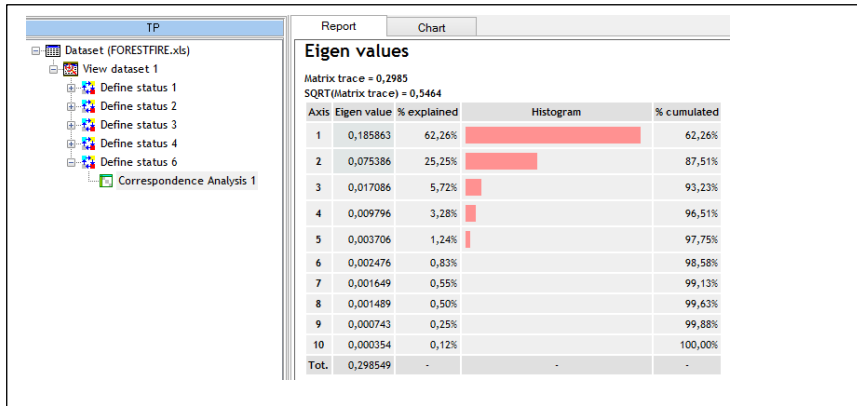
### KHI-2 (global) de l'écart à l'indépendance

D'après le tableau ci-dessous, nous avons  $\text{Chi}^2 = 130813,08$ , avec un degré de liberté égal à 5160, la liaison est très significative ( $p\text{-value} < 0.0001$ )

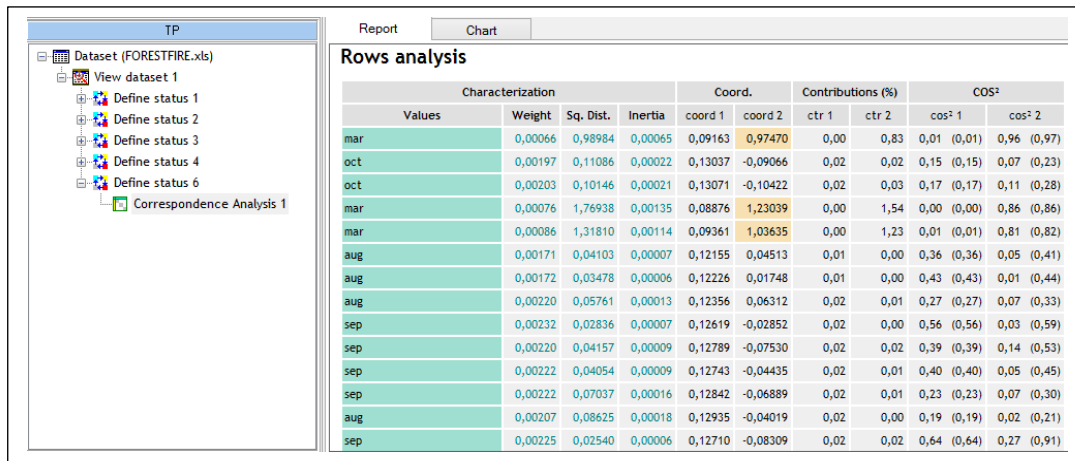


## Valeurs propres :

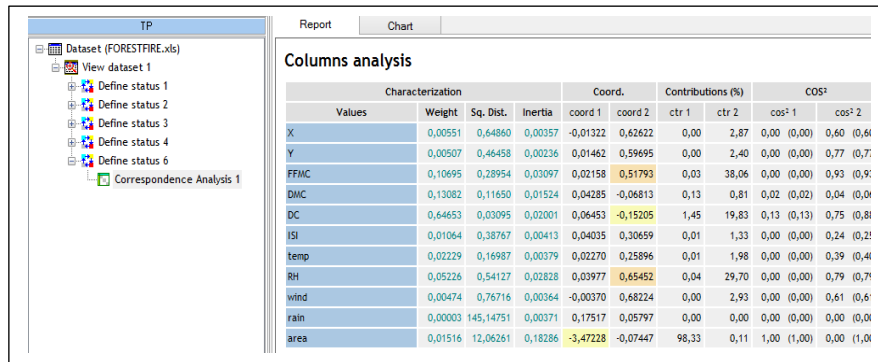
D'après les résultats, nous observons la valeur propre calculée, le pourcentage d'inertie associé à chaque axe et le pourcentage cumulé qui permet de se donner une idée du nombre d'axes à retenir. Dans notre exemple, les deux premiers axes résument 62.26% de l'information disponible.



## Représentation des lignes

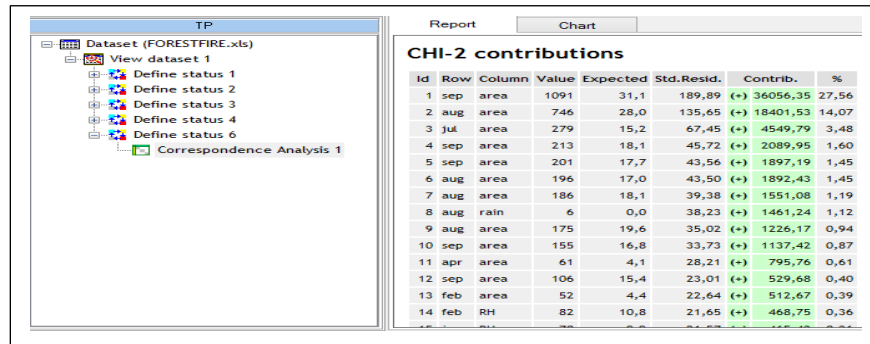


## Représentation des colonnes



## Contributions au KHI-2

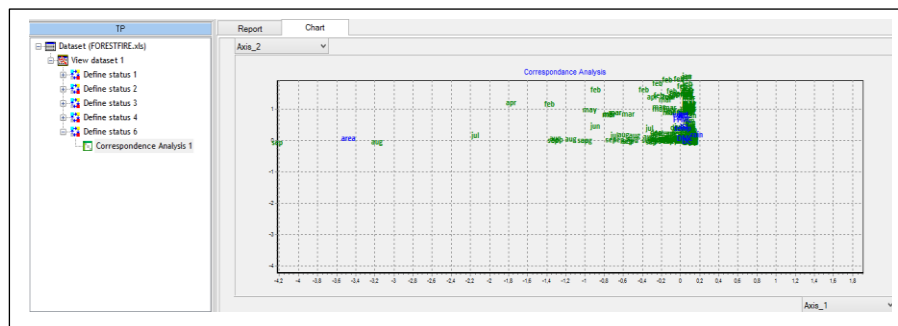
Le résultat ci-dessous recense la contribution au CHI-2 de chaque case du tableau de contingence, en confrontant la valeur observée et la valeur espérée sous l'hypothèse d'indépendance. Il s'agit d'une autre manière de détecter les informations importantes



The screenshot shows the TP software interface. On the left, a tree view under 'Dataset (FORESTFIRE.xls)' includes 'View dataset 1', 'Define status 1' through 'Define status 6', and 'Correspondence Analysis 1'. The main window displays a table titled 'CHI-2 contributions' with columns: Id, Row, Column, Value, Expected, Std.Resid., Contrib., and %. The table contains 14 rows of data for various months and weather conditions.

Id	Row	Column	Value	Expected	Std.Resid.	Contrib.	%
1	sep	area	1091	31,1	189,89	(+) 36056,35	27,56
2	aug	area	746	28,0	135,65	(+) 18401,53	14,07
3	jul	area	279	15,2	67,45	(+) 4549,79	3,48
4	sep	area	213	18,1	45,72	(+) 2089,95	1,60
5	sep	area	201	17,7	43,56	(+) 1897,19	1,45
6	aug	area	196	17,0	43,50	(+) 1892,43	1,45
7	aug	area	186	18,1	39,38	(+) 1551,08	1,19
8	aug	rain	6	0,0	38,23	(+) 1461,24	1,12
9	aug	area	175	19,6	35,02	(+) 1226,17	0,94
10	sep	area	155	16,8	33,73	(+) 1137,42	0,87
11	apr	area	61	4,1	28,21	(+) 795,76	0,61
12	sep	area	106	15,4	23,01	(+) 529,68	0,40
13	feb	area	52	4,4	22,64	(+) 512,67	0,39
14	feb	RH	82	10,8	21,65	(+) 468,75	0,36

## Représentation graphique



## TP5 : Méthodes de classification (clustering)

Faire une classification de vos données en combinant les méthodes d'analyse factorielle et les méthodes de classification HAC et K-means. Interpréter les résultats par les individus et les variables.

### 1- K-means

La méthode des K-Means (méthode des centres mobiles) est une technique de classification automatique (clustering en anglais). Elle vise à produire un regroupement de manière à ce que les individus du même groupe soient semblables et que les individus dans des groupes différents soient dissemblables

Nous allons insérer dans un nouveau un define status7 au niveau de l'opérateur *Principal Component Analysis* ensuite mettre toutes les données issues de l'analyse factorielle en entrée.

Nous allons maintenant insérer le composant K-MEANS (onglet CLUSTERING) et paramétrer son menu contextuel PARAMETERS Le nombre de clusters demandé est 20 (Number of Clusters). Nous ne réalisons que 3 essais d'optimisation (Number of trials = 3), avec un nombre d'itération maximum à 10 (Max iterations).

**K-Means parameters**

Clusters	20
Max Iteration	10
Trials	3
Distance normalization	none
Average computation	McQueen
Seed random generator	Standard

**Global evaluation**

Within Sum of Squares	178,7033
Total Sum of Squares	3408,9185
R-Square	0,9476

**R-Square for each attempt**

Number of trials	3
Trial	R-square
1	0,864475
2	0,947578
3	0,808486

**Cluster centroids**

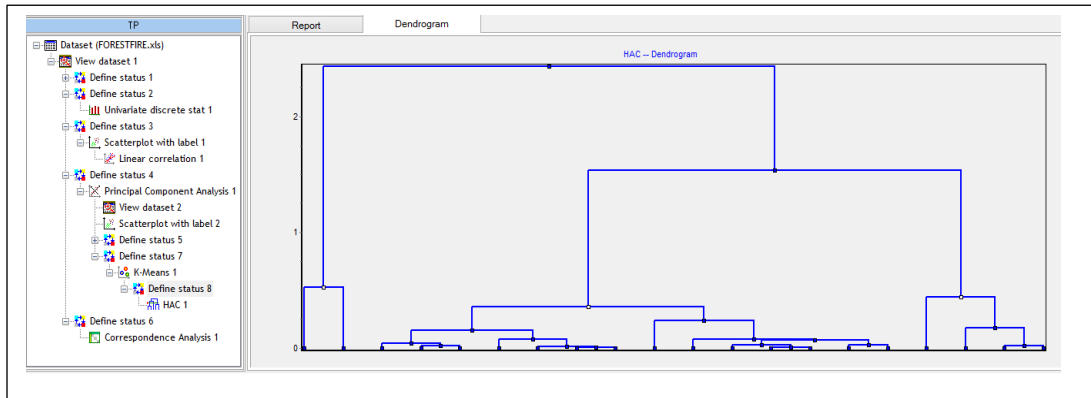
Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4	Cluster n°5	Cluster n°6	Cluster n°7	Cluster n°8	Cluster n°9	Cluster n°10	Cluster n°11	Cluster n°12	Cluster n°13	Cluster n°14	Cluster n°15	Cluster n°16	Cluster n°17
PCA_1_Axis_1	-0,668147	1,134509	-1,779888	-0,832219	-0,290897	2,774847	-1,269998	-4,046661	9,043972	-0,251425	-0,758329	-0,019525	1,522479	-1,652365	0,749498	4,124095	-0,674706
PCA_1_Axis_2	0,284049	-1,960008	-1,541624	-0,337768	-1,046066	-0,239961	-1,140745	-7,380012	5,522929	0,180802	0,805682	0,635570	1,284074	-0,252252	-0,124777	3,525748	0,040740
PCA_1_Axis_3	-0,191768	-0,010385	1,309843	-0,038095	0,211845	-0,161671	0,567895	6,198772	8,189713	-0,280189	-0,224453	-0,327088	0,051329	0,455148	-0,438731	3,168452	0,274354
PCA_1_CTR_1	0,042149	0,127822	0,275571	0,062262	0,015055	0,699646	0,143166	1,414858	7,067042	0,007013	0,052532	0,007159	0,217042	0,248799	0,054018	1,526668	0,041926
PCA_1_CTR_2	0,017755	0,756320	0,473496	0,024900	0,224080	0,057374	0,262260	10,650464	5,964765	0,011738	0,131360	0,081799	0,336858	0,027641	0,018654	2,471070	0,002055
PCA_1_CTR_3	0,022959	0,039862	0,732414	0,026926	0,067285	0,098504	0,149331	16,180296	28,243113	0,039916	0,026135	0,059791	0,013124	0,116667	0,093856	4,462236	0,039047
PCA_1_COS2_1	0,586150	0,239396	0,434054	0,665003	0,103204	0,919937	0,467569	0,149845	0,456009	0,164958	0,348836	0,092487	0,490436	0,829945	0,425933	0,421247	0,714621
PCA_1_COS2_2	0,100716	0,719142	0,317306	0,127361	0,757263	0,034058	0,363587	0,498384	0,170057	0,091821	0,408997	0,505862	0,373199	0,040262	0,076410	0,321771	0,007879
PCA_1_COS2_3	0,071983	0,016902	0,230313	0,068744	0,079361	0,022501	0,090447	0,351610	0,373932	0,208350	0,046305	0,179381	0,010555	0,093214	0,177330	0,244635	0,099021

## 2- classification HAC



Classification Ascendante Hiérarchique (CAH) est une technique de classification automatique (clustering en anglais). Elle vise à produire un regroupement des individus de manière à ce que les individus du même groupe soient semblables.

Nous souhaitons maintenant réaliser la CAH (Classification Ascendante Hiérarchique) en prenant comme groupes de départ, ceux qui sont produits par les K-MEANS. Nous insérons encore une fois le composant DEFINE STATUS. Nous plaçons en TARGET la variable indicatrice des sous-groupes CLUSTER\_KMEANS\_1, produite par le composant KMEANS. Ensuite nous définissons une 3 classe pour la répartition de nos données. Ces paramètres nous donne le résultat suivant



Pour bien comprendre l'appartenance aux groupes, nous aurons besoin du composant GROUP CHARACTERIZATION afin de pouvoir les interpréter. Nous introduisons tout d'abord le composant DEFINE STATUS. Nous plaçons en TARGET la variable à caractériser CLUSTER\_HAC\_1; en INPUT les variables originelles et ensuite Nous insérons le l'opérateur **Group Characterization**

TP

Define status 2

Univariate discrete stat 1

Define status 3

Scatterplot with label 1

Linear correlation 1

Define status 4

Principal Component Analysis

View dataset 2

Scatterplot with label 2

Define status 5

Correlation scatterplot

Define status 7

K-Means 1

Define status 8

HAC 1

Define statu

Group c

Cluster\_HAC\_1=c\_hac\_1

Examples [ 0,4 % ] 2

Att - Desc Test value Group Overall

Continuous attributes : Mean (StdDev)

ISI 5,91 28,05 (39,67) 9,02 (4,56)

RH 2,40 72,00 (39,60) 44,29 (16,32)

X 0,51 5,50 (2,12) 4,67 (2,31)

Y 0,23 4,50 (0,71) 4,30 (1,23)

rain -0,10 0,00 (0,00) 0,02 (0,30)

area -0,29 0,00 (0,00) 12,85 (63,66)

wind -1,04 2,70 (2,55) 4,02 (1,79)

DMC -1,37 48,70 (67,32) 110,87 (64,05)

temp -1,41 13,10 (11,17) 18,89 (5,81)

DC -2,07 185,70 (20,22) 547,94 (248,07)

FFMC -8,76 56,50 (53,46) 90,64 (5,52)

Discrete attributes : [Recall] Accuracy

month=jan 11,31 [ 50,0 % ] 50,0 % 0,4 %

month=jun 3,71 [ 5,9 % ] 50,0 % 3,3 %

month=nov -0,06 [ 0,0 % ] 0,0 % 0,2 %

month=may -0,09 [ 0,0 % ] 0,0 % 0,4 %

month=dec -0,19 [ 0,0 % ] 0,0 % 1,7 %

month=apr -0,19 [ 0,0 % ] 0,0 % 1,7 %

month=oct -0,24 [ 0,0 % ] 0,0 % 2,9 %

month=feb -0,28 [ 0,0 % ] 0,0 % 3,9 %

month=jul -0,36 [ 0,0 % ] 0,0 % 6,2 %

month=mar -0,48 [ 0,0 % ] 0,0 % 10,4 %

month=sep -1,00 [ 0,0 % ] 0,0 % 33,3 %

month=aug -1,05 [ 0,0 % ] 0,0 % 35,6 %

Cluster\_HAC\_1=c\_hac\_2

Examples [ 79,1 % ] 409

Att - Desc Test value Group Overall

Continuous attributes : Mean (StdDev)

DC 19,99 660,12 (112,18) 547,94 (248,07)

DMC 14,88 132,43 (53,56) 110,87 (64,05)

temp 13,48 20,66 (4,56) 18,89 (5,81)

FFMC 10,53 91,96 (2,26) 90,64 (5,52)

ISI 7,38 9,78 (3,80) 9,02 (4,56)

area 1,18 14,55 (71,19) 12,85 (63,66)

rain 0,78 0,03 (0,33) 0,02 (0,30)

RH -1,44 43,76 (15,30) 44,29 (16,32)

X -1,58 4,59 (2,40) 4,67 (2,31)

Y -1,81 4,25 (1,22) 4,30 (1,23)

wind -5,20 3,81 (1,64) 4,02 (1,79)

Discrete attributes : [Recall] Accuracy

month=aug 8,23 [ 98,9 % ] 44,5 % 35,6 %

month=sep 7,78 [ 98,8 % ] 41,6 % 33,3 %

month=jul 2,10 [ 93,8 % ] 7,3 % 6,2 %

month=oct 2,02 [ 100,0 % ] 3,7 % 2,9 %

month=jun -0,88 [ 70,6 % ] 2,9 % 3,3 %

month=nov -1,95 [ 0,0 % ] 0,0 % 0,2 %

month=jan -2,75 [ 0,0 % ] 0,0 % 0,4 %

month=may -2,75 [ 0,0 % ] 0,0 % 0,4 %

month=apr -5,88 [ 0,0 % ] 0,0 % 1,7 %

month=dec -5,88 [ 0,0 % ] 0,0 % 1,7 %

month=feb -8,87 [ 0,0 % ] 0,0 % 3,9 %

month=mar -15,10 [ 0,0 % ] 0,0 % 10,4 %

Cluster\_HAC\_1=c\_hac\_3

Examples [ 20,5 % ] 106

Att - Desc Test value Group Overall

Continuous attributes : Mean (StdDev)

wind 5,40 4,86 (2,07) 4,02 (1,79)

Y 1,79 4,49 (1,26) 4,30 (1,23)

X 1,51 4,97 (1,94) 4,67 (2,31)

RH 1,08 45,81 (19,11) 44,29 (16,32)

rain -0,77 0,00 (0,02) 0,02 (0,30)

area -1,15 6,51 (12,76) 12,85 (63,66)

ISI -8,34 5,72 (3,32) 9,02 (4,56)

FFMC -9,26 86,22 (7,35) 90,64 (5,52)

temp -13,36 12,17 (5,02) 18,89 (5,81)

DMC -14,77 28,86 (17,23) 110,87 (64,05)

DC -19,81 121,94 (137,04) 547,94 (248,07)

Discrete attributes : [Recall] Accuracy

month=mar 15,28 [ 100,0 % ] 50,9 % 10,4 %

month=feb 8,97 [ 100,0 % ] 18,9 % 3,9 %

month=apr 5,95 [ 100,0 % ] 8,5 % 1,7 %

month=dec 5,95 [ 100,0 % ] 8,5 % 1,7 %

month=may 2,79 [ 100,0 % ] 1,9 % 0,4 %

month=nov 1,97 [ 100,0 % ] 0,9 % 0,2 %

month=jan 1,03 [ 50,0 % ] 0,9 % 0,4 %

month=jun 0,31 [ 23,5 % ] 3,8 % 3,3 %

month=oct -1,99 [ 0,0 % ] 0,0 % 2,9 %

month=jul -2,06 [ 6,3 % ] 1,9 % 6,2 %

month=sep -7,68 [ 1,2 % ] 1,9 % 33,3 %

month=aug -8,12 [ 1,1 % ] 1,9 % 35,6 %

Interprétation :

(C\_HAC\_1) :

Nous remarquons que : dans le premier groupe (C\_HAC\_1) la moyenne d'indice de propagation du feu de forêt (ISI) est un peu élevé (9,02), ce qui signifie une propagation rapide suite à une présence considérable des combustibles légers (les feuilles mortes, les débris végétaux) entraîné par un effet de vent considérable. Nous avons aussi l'humidité relative avec une moyenne de 44,29 (l'air peut absorber encore beaucoup de vapeur d'eau puisqu'il n'est qu'à 44,29 % de la saturation). Ces phénomènes sont beaucoup plus fréquents au mois de janvier (50%) et juin (50%) dans le groupe

(C\_HAC\_2) :

Nous remarquons que : la moyenne d'indice de sécheresse (DC) est très élevée (547,94) ce qui signifie que la présence de la teneur en eau des épaisse couches organiques compactés est presque inexistante. Cela pourrait entraîner une combustion rapide des matières organiques séchées. Ceci est plus fréquent au mois de d'Aout 44,5% et Septembre 41,6% dans le groupe

(C\_HAC\_3)

Nous remarquons que : la moyenne de la vitesse du vent est faible (4,02). Ce résultat signifie une montée verticale de la fumée. Ce phénomène est fréquemment observé au mois de Mars et Février

### Tableau de contingence sur K-Means

un **tableau de contingence** (également appelé **table croisée** ou **tableau croisé**) est un type de tableau dans un format matriciel qui affiche la distribution de fréquences (multivariée) des variables. Ils sont fortement utilisés dans la recherche d'enquêtes, de business intelligence, d'ingénierie et de recherche scientifique. Ils fournissent une image de base de l'interrelation entre deux variables et peuvent aider à trouver des interactions entre eux. Pour ce fait nous allons mettre **en target la variable month et en input la variable Cluster\_KMeans**  
**résultat :**

Stat	Value	c_kmeans_1	c_kmeans_2	c_kmeans_3	c_kmeans_4	c_kmeans_5	c_kmeans_6	c_kmeans_7	c_kmeans_8	c_kmeans_9	c_kmeans_10	c_kmeans_11	c_kmeans_12
d.f.	209	0	6	0	0	0	16	0	0	0	0	0	0
Tschuprow's t	0,448726	0,00%	11,11%	0,00%	0,00%	0,00%	29,63%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
Cramer's v	0,514424	0	0	0	0	0	0	0	0	0	0	0	0
Phi²	2,910952	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
Chi² (p-value)	1504,96 (0,0000)	9 4,89%	1 0,54%	8 4,35%	10 5,43%	8 4,35%	0 0,00%	18 9,78%	0 0,00%	0 0,00%	3 1,63%	21 11,41%	0
Lambda	0,459459	51 29,65%	0 0,00%	1 0,58%	6 3,49%	0 0,00%	0 0,00%	9 5,23%	0 0,00%	0 0,00%	21 12,21%	14 8,14%	0
Tau (p-value)	0,3916 (0,0000)	0 0,00%	2 22,22%	0 0,00%	0 0,00%	0 0,00%	7 77,78%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0
U(R/C) (p-value)	0,5219 (0,0000)	0 0,00%	2 11,76%	0 0,00%	1 5,88%	2 11,76%	0 0,00%	1 5,88%	1 5,88%	0 0,00%	0 0,00%	0 0,00%	0
		1 3,13%	0 0,00%	0 0,00%	2 6,25%	7 21,88%	1 3,13%	1 3,13%	0 0,00%	0 0,00%	3 9,38%	0 0,00%	0
		0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	20 100,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0
		0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	1 5,00%	0 0,00%	0 0,00%	1 5,00%	0 0,00%	0 0,00%	0
		0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	9 100,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0
		0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0 0,00%	0

### Interprétation du tableau de contingence

D'après le tableau ci-dessus, nous remarquons que dans la 1<sup>ère</sup> catégories des individus de c\_kmean\_1 nous avons 29% de feux de forêts sont plus fréquents au mois de septembre et 4% de feu plus fréquents au mois d'Août durant la période de 1987 à 2005

## Références

[1]

**PauloCORTEZetAnibalMORAIS.«ADataMiningApproachtoPredictForestFiresusing Meteorological Data. ». In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN13 978-989-95618-0-9. Available at : [http ://www.dsi.uminho.pt/pcortez/fires.pdf](http://www.dsi.uminho.pt/pcortez/fires.pdf)**