INSTITUT NATIONAL DE STATISTIQUE ET D'ECONOMIE APPLIQUÉE

MASTER SYSTÈMES D'INFORMATIONS ET SYSTÈMES INTELLIGENTS

PROJET ANALYSE DE DONNÉES

ANALYE EN COMPASANTE PRINCIPALE

Réalisé Par:

MOUMAD Hamza

EZZAGRANI Habiba

Encadrée Par :

EL HANNOUN Wafaa

23/12/2022

Année Académique : 2022-2023

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

ACP : Analyse en Composante Principale
KMO : Kaiser-Meyer-Olkin
PIB : produit intérieur brut
HCP : Haut-commissariat au Plan

Listings

INTRODUCTION GÉNÉRALE

Qu'est-ce que l'analyse en composantes principales ?

L'analyse en composantes principales est l'une des méthodes d'analyse de données multivariées les plus fréquemment utilisées. Elle permet d'étudier des ensembles de données multidimensionnelles avec des variables quantitatives. Elle est largement utilisée en biostatistique, en marketing, en sociologie et dans de nombreux autres domaines.

l s'agit d'une méthode de projection car elle projette les observations d'un espace à p dimensions avec p variables vers un espace à k dimensions (où k < p) de manière à conserver le maximum d'information, des dimensions initiales. Les dimensions de l'ACP sont également appelées axes ou facteurs. Si l'information associée aux 2 ou 3 premiers axes représente un pourcentage suffisant de la variabilité totale du nuage de points.

L'ACP peut donc être considérée comme une méthode d'exploration de données car elle permet d'extraire facilement des informations de grands ensembles de données. Elle peut être utilisée à plusieurs fins, notamment :

— L'étude et la visualisation des corrélations entre les variables.

— L'obtention de facteurs non corrélés qui sont des combinaisons linéaires des variables.

— La visualisation des observations dans un espace à deux ou trois dimensions.

Chapitre

1

Aspect théorique de l'ACP

Sommaire

1.1 Introduction

Dans ce chapitre on va entamer la phase théorique, on présentera une introduction à L'ACP, les étapes à suivre dans un projet PCA et finalement on donnera un exemple explicatif comment l'ACP travaille.

1.2 Les etapes de l'ACP

1. Centrer et réduire la matrice des données.
2. Déterminer les valeurs propres de la matrice de corrélation.
3. Classer les valeurs propres selon l'ordre décroissant.
4. Calculer les composantes principales.
5. Calculer les qualités et les contributions des individus dans les composantes principales.
6. Interpréter les résultats obtenus.

1.3 Exemple introductif

1.3.1 Énoncer du problème

Une étude gastronomique à apprécier le service, le prix et la qualité de 4 restaurants. Les résultats sont stockés dans le tableau suivant :

Restaurant	**Services**	**Qualite**	**Prix**
R1	-2	+3	-1
R2	-1	+1	0
R3	+2	-1	-1
R4	+1	-3	+2

TABLE 1.1 – Le résultat du l'étude gastronomique.

1.3.2 Etude des valeurs propres

la matrice des variances-covariances est :

$$V = \begin{pmatrix} 5/2 & -3 & 1/2 \\ -3 & 5 & -5 \\ 1/2 & -2 & 3/2 \end{pmatrix} \tag{1.1}$$

la matrice des Correlation est :

$$T = \begin{pmatrix} 1 & -0.85 & 0.26 \\ -0.85 & 1 & -0.73 \\ 0.26 & -0.73 & 1 \end{pmatrix} \tag{1.2}$$

Pour vérifier simplement que V admet une valeur propre nulle, il suffit de calculer son déterminant (qui doit être nul).

$$\det V = \begin{vmatrix} 5/2 & -3 & 1/2 \\ -3 & 5 & -5 \\ 1/2 & -2 & 3/2 \end{vmatrix}$$

Pour retrouver le déterminant, il suffit de rajouter à la première ligne les deux autres. Et donc, on aura :

$$\det V = \begin{vmatrix} 0 & 0 & 0 \\ -3 & 5 & -5 \\ 1/2 & -2 & 3/2 \end{vmatrix} = 0$$

On sait que la somme es valeurs propres est la trace de la matrice à diagonalise (dans notre cas : la matrice V)

$$\lambda 1 + \lambda 2 + \lambda 3 = 5/2 + 5 + 3/2 = 9, .On$$

en déduit que

$$\lambda 2 = 9 - 30.5/4 = 5.5/4$$

Le pourcentage des inerties :

$$\lambda 1(\%) = (30.5/4)/9 = 84.72\%, \lambda 2(\%) = (5.5/4)/9 = 15.28\%, \lambda 3(\%) = 0/9 = 84.72\%.$$

En examinant le diagramme des valeurs propres, et utilisant le critère du coude qui casse :

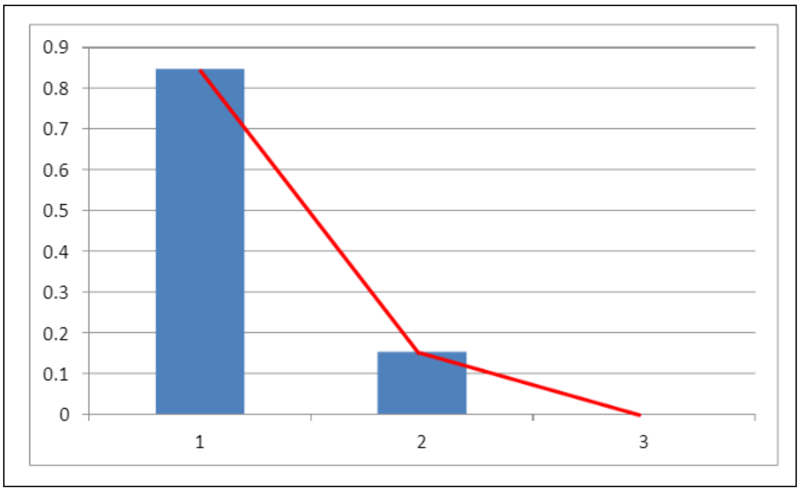


FIGURE 1.1 – Diagramme des valeurs propres

La dimension à retenir est 1 seule.

Le deuxième sera pris pour la frome !!

1.3.3 Choix des conposantes principales

On donne (aux erreurs d'arrondi près) :

$$u1 = \begin{pmatrix} 0.5 \\ -0.8 \\ 0.3 \end{pmatrix} u2 = \begin{pmatrix} 0.65 \\ 0.11 \\ -0.75 \end{pmatrix} \tag{1.3}$$

$Pourtrouverlescomposantesprincipales, ilfaudrafaireleproduitmatricieldelamatricedesdonnescentr$

$$\tag{1.4}$$

On remarque que les données sont centrées depuis le départ, Alors

$$Xu1 = \begin{pmatrix} -3.7 \\ -1.3 \\ 1.5 \\ 3.5 \end{pmatrix} Xu2 = \begin{pmatrix} -0.2 \\ -0.45 \\ 1.94 \\ -1.18 \end{pmatrix}$$

La représentation des individus dans le plan principal (1,2) :

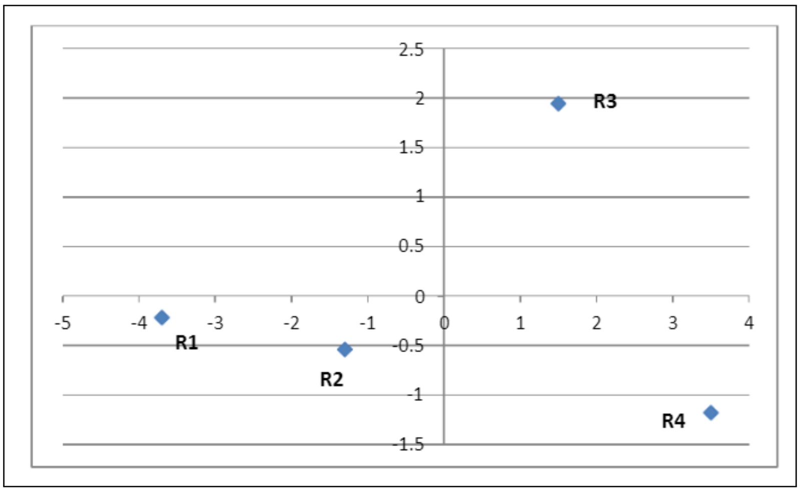


FIGURE 1.2 – Représenter les individus

La corrélation entre les variables et les composantes principales

$$Cor(Vj, ci) = (produitscalairedesdeuxvecteurs) divisparleproduitdeleurscarts - types$$

$$Lacomposantecitantdevariance \lambda 1$$

$$cov(v1, c1) = 0.87$$

De la meme façon on trove les autres corrélations :

	V1	V2	V3
C1	0.87	-0.99	0.68
C2	0.50	0.048	-0.71

TABLE 1.2 – Le résultat du l'étude gastronomique.

1.4 Conclusion

Donc, comme nous avons voyer, ce sont les étapes principale de la méthode PCA, qui peuvent réduire l'ensemble de données de n dimensions à de petites dimensions telles que 2D ou biens 3D, et c'est tellement intéressant car cela vous donne un graphique illustré de vos données, puis vous pouvez l'analyser et extraire ses caractéristiques.

Chapitre

2 Mise en pratique de l'ACP

Sommaire

2.1 Introduction

Afin de mettre en pratique l'ACP Je choisis comme sujet pour y appliquer les données sur la variation des Valeurs ajouter dans le Maroc en quelque domaines depuis l'année 2014 jusqu'à 2021.

2.2 Description des données

Dans le but de mettre en pratique la Méthode HCP avec le langage R nous avons choisir comme données la variation de Valeurs ajouter au Maroc dans les années précédents.

Ces données font partie de de l'ensemble des données publiées par Le Haut-commissariat au Plan (HCP) de Compte de nation. en fait les comptes nationaux constituent l'une des composantes essentielles du système national d'information statistique. Ils donnent une représentation chiffrée détaillée de l'économie nationale aux niveaux annuel et trimestriel.

Nos données sont issues de différentes sources d'informations telles que les recensements et enquêtes réalisés par l'HCP ou encore les statistiques administratives fournies par ses partenaires institutionnels.

Apres le filtrage des données, le **tableau de contingence des données est représenter comme suite :**

	Peche	IE	IT	FPA	FPC	FPP	FO	FME	FM	FMT	Construction	TE	AHR	AFA	AI	RDS	ESH	INS	PIB	PIBNA	SP	SS	ST
14_S1	914	5338	37778	9821	3892	1511	246	915	699	2482	14440	7997	8929	11010	18230	9457	19115	21156	245678	222140	24452	63258	136812
14_S2	1213	5510	38592	10266	4029	1537	252	1097	708	2306	14499	8407	9529	11075	18250	9787	19208	22161	250699	226022	25890	64683	137965
14_S3	1250	5574	39296	10292	4382	1551	261	1086	712	2589	14819	8507	8456	11123	18189	10088	19298	23232	251966	227270	25946	65845	136944
14_S4	1309	5706	38010	10213	4527	1418	262	970	706	2216	14647	8928	8641	11187	18118	10007	19384	24369	253111	227026	27394	64722	136626
15_S1	1383	6084	39434	10062	5184	1480	263	1013	690	2343	14562	9200	8937	11112	18124	10634	19467	26809	264545	235312	30616	67204	139916
15_S2	1437	6351	40441	10705	5122	1431	268	987	733	2601	14750	9465	9180	11185	18209	10880	19547	27726	268243	238480	31201	68788	140529
15_S3	1413	6238	41001	10696	5317	1434	266	1195	746	2885	15220	9598	8808	11256	18307	11027	19623	28431	271721	240500	32634	70077	140578
15_S4	1616	5870	40670	10923	5159	1381	264	992	728	2771	15751	9820	8919	11310	18460	11112	19696	28923	273609	242500	32726	70174	141787
16_S1	1734	5317	40276	10584	4484	1405	257	1066	780	2884	16131	9860	9133	11498	18788	11806	19754	28613	269487	243921	27300	70044	143530
16_S2	1905	5033	39604	10716	3927	1447	256	997	774	2930	16264	9465	9112	11556	18951	11651	19825	28879	270444	244425	27924	69368	144273
16_S3	1734	5012	40603	11156	4231	1413	256	935	774	3391	16234	10027	9805	11633	19181	11886	19898	29130	276116	248966	28883	70788	147315
16_S4	2047	4809	41022	11516	4365	1391	251	967	802	3068	16189	9979	10059	11739	19400	11987	19972	29365	278202	251408	28841	70859	149137
17_S1	2024	5119	41587	12430	4282	1374	267	1076	831	3184	16717	10023	10589	11789	19515	12432	19950	29184	284507	253878	32653	72197	150473
17_S2	1821	5328	42108	12369	4329	1414	284	1022	773	3268	16834	9717	11109	11929	19705	12642	20066	29470	286776	256087	32510	73144	151652
17_S3	1996	5164	41852	12461	4314	1455	283	1002	754	3280	17107	9967	11504	12124	19838	12666	20223	29829	287622	257985	31633	73157	153002
17_S4	2079	5198	42125	12199	4342	1502	283	1240	743	3504	16852	9801	11614	12352	20098	12657	20421	30261	289990	260646	31424	73383	154922
18_S1	2086	5130	43428	12434	5103	1495	285	1132	845	3699	16917	9621	11735	12881	20271	13411	20818	30460	295089	263451	33724	74826	156079
18_S2	2136	5336	43839	12420	5539	1464	285	1335	755	3782	16982	9899	12000	13077	20540	13545	21033	30900	298050	266379	33807	75197	158145
18_S3	2036	5173	44959	12789	6022	1484	287	1137	755	3925	16564	9641	12202	13170	20796	13627	21225	31244	300583	268417	34202	75845	159293
18_S4	1785	5150	45511	12762	6026	1528	292	1425	804	4161	15675	9563	12277	13176	21011	13825	21395	31489	301515	269906	33394	75826	160806
19_S1	1601	5292	44857	12222	5417	1452	294	1330	781	4242	16054	10397	12527	13112	21374	14477	21540	32297	304603	274221	31983	77053	163269
19_S2	1582	5042	46395	12661	5520	1547	277	1365	804	4277	16870	11689	12679	13052	21536	14823	21664	32375	314660	279901	36341	79104	166841
19_S3	1756	5046	45731	12755	5193	1589	278	1352	803	4210	17126	10766	12768	13001	21663	14853	21766	32366	311256	279228	33783	78589	166517
19_S4	1638	5012	46058	12576	5141	1577	290	1642	768	4344	16637	10245	12644	12929	21771	14959	21846	32271	309317	278636	32320	78312	166414

FIGURE 2.1 – Tableau de contingence des données

2.2.1 Explication des données représenter dans la table

Abr	Explication	Abr	Explication
IE	Industrie d'extraction	**IT**	Industrie de transformation
FPA	Fabrication de produits alimentaires	**FPC**	Fabrication de produits chimiques
FPP	Fabrication de produits pharmaceutiques	**FO**	Fabrication des ordinateurs
FME	Fabrication de matériel électrique	**FM**	Fabrication de machines
FMT	Fabrication de matériel de transport	**TE**	Transports et entreposage
FMT	Fabrication de matériel de transport	**TE**	Transports et entreposage
AHR	Activités hébergements et de restauration	**AFA**	Activités financières et assurances
AI	Activités immobilières	**RDS**	Recherches et développement et services
ESH	Education santé humaine	**INS**	impôts net subventions
PIBNA	PIB non agricole	**SP**	Secteur primaire
SS	Secteur secondaire	**ST**	: Secteur tertiaire

TABLE 2.1 – Table d'escriptif de la Classe Projet.

2.3 Implémentation de l'APC en utilisant R

La première étape pour appliquer l'ACP est d'importer les packages nécessaires, puis se déplacer vers le répertoire où se trouve notre fichier Excel (.csv).

```
# fonctions : PCA, dimdesc (decrire les dimension ou bien les axes
#de notre resultats)
installed.packages("FactoMinerR")
library("FactoMineR")
#setwd permet de changer le repertoire (Set Working Directory)
setwd("E:/tpr/ProjetAD/Version_final")
#lire le fichier excel .cvs en utilisant les option suivant
table <- read.csv("ProjetFinal.csv", header = TRUE, dec = ",",
        row.names = "Saison")
```

Listing 2.1 – importation des packages

2.3.1 Matrice de corrélation

C'est quoi une matrice de corrélation ?

Une **matrice de corrélation** est utilisée pour évaluer la **dépendence** entre plusieurs variables en même temps. Le résultat est une table contenant les coefficients de corrélation entre chaque variable et les autres. L'objectif de cette partie est de vous montrer comment calculer et visualiser une matrice de corrélation dans R.

```
#construction de la matrice de correlation 1 et 2
pairs(table[,1:10])
matrice1 <- cor(table[,1:11])
print(matrice1)
pairs(table[,10:23])
matrice2 <- cor(table[,12:23])
print(matrice2)
```

Listing 2.2 – Matrice de corrélation

Analyse de corrélation dans R

Le tableau suivant montre une partie du résultat de la matrice de corrélation, cette matrice contiens des valeurs négatives, positives et 1. Pour mieux expliquer les rôles de ces valeurs on va tracés des digrammes par la suite.

```
                Peche          IE          IT        FPA       FPC
Peche      1.00000000 -0.03215224  0.63816598  0.5010136 0.5305926
IE        -0.03215224  1.00000000 -0.17617416 -0.5806441 0.1114477
IT         0.63816598 -0.17617416  1.00000000  0.6361248 0.7918310
FPA        0.50101363 -0.58064408  0.63612479  1.0000000 0.2196806
FPC        0.53059260  0.11144773  0.79183100  0.2196806 1.0000000
```

Corrélogramme : visualisation d'une matrice de corrélation

Plusieurs méthodes sont disponibles dans R pour dessiner un corrélogramme. Vous pouvez utiliser soit la fonction ***symnum()***, la fonction ***corrplot()*** ou des nuages de points pour faire le graphique de la matrice de corrélation. Pour nous on va adopter la fonction ***corrplot()***.

```
#fonction pour afficher la matrice de correlation
par(mfrow = c(1, 2))
corrplot(matrice1,type="upper",order="hclust",tl.col="black",tl.srt=45)
corrplot(matrice2,type="upper",order="hclust",tl.col="black",tl.srt=45)
```

Listing 2.3 – Code pour la matrice de corrélation

l'omage suivant present le graphique que j'ai tiré sans trop de difficultés de ma série de données (j'ai travailler seulement avec 5 variables) :

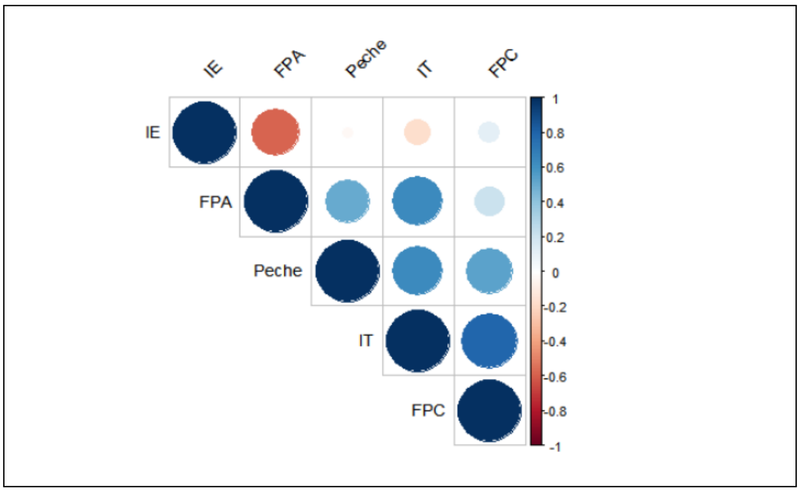


FIGURE 2.2 – Diagramme pour la matrice de corrélation

Interprétation du résultat du graphe de la matrice de corrélation

Les **corrélations positives** sont affichées en **bleu** et les **corrélations négatives** en **rouge**. L'intensité de la couleur et la taille des cercles sont proportionnelles aux coefficients de corrélation. A droite du corrélogramme, la légende de couleurs montre les coefficients de corrélation et les couleurs correspondantes.

2.3.2 Matrice d'identitaire

Afin de vérifier qu'il existe bien des corrélations suffisantes entre les variables, il faut utiliser le test de sphéricité de Bartlett. Ce test évalue l'hypothèse nulle selon laquelle les corrélations seraient toutes égales à zéro. Il devrait être significatif au seuil de p<.001 (c'est-à-dire que H0 devrait être rejetée) pour continuer l'analyse.

```
# test pour observer si une matrice identitaire
det(matrice1)
#si H0 : det(matrice) = 1,  H1 : det(matrice) <> 1
cortest.bartlett(matrice1, n = 23)
KMO(matrice1)
```

Listing 2.4 – Code pour la matrice d'identitaire

Le code précèdent donne le résultat suivant :

Un indice KMO > .80 signifie que la « factoriabilité » est bonne, c'est-à-dire que la structure factorielle est intelligible et stable ; un indice KMO entre .60 et .80 correspond à une « factoriabilité » dite moyenne ; si l'indice KMO < .60, la « factoriabilité » est dite mauvaise et la structure factorielle est 14 difficile à interpréter et instable.

```
KMO(matrice)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = matrice)
Overall MSA =  0.7
MSA for each item =
       Peche        IE        IT       FPA       FPC       FPP        FO
         0.53      0.48      0.77      0.69      0.57      0.52      0.26
Construction         TE       AHR       AFA        AI       RDS       ESH
```

```
        0.73     0.59     0.37     0.91     0.79     0.71     0.81
          SP       SS       ST
        0.69     0.76     0.76
```

Interprétation du résultat de la fonction KMO

On observe que notre indice KMO entre .60 et .80 (KMO = 0.7) alors cela correspond à une « factoriabilité » dite moyenne.

2.3.3 utilisation de l'ACP

```
#utiliser PCA pour afficher les 32 lignes existe
#et les 23 colonnes a partir de la colone 1
NewTab1 = PCA(table[1:32,1:23],scale.unit = TRUE,ncp = 5,graph = TRUE)
```

Listing 2.5 – lancenent de l'ACP

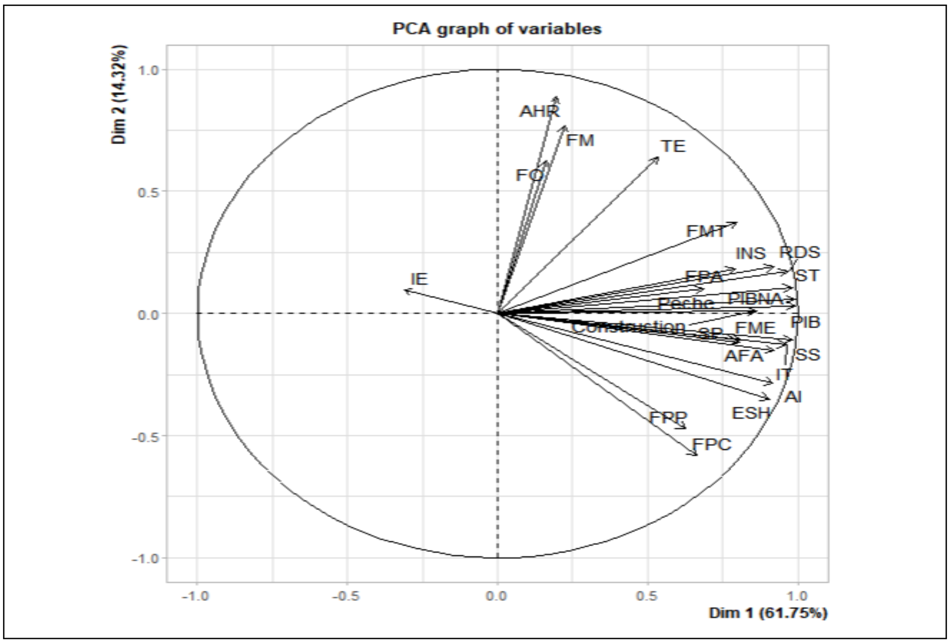


FIGURE 2.3 – résultats de la fonction PCA

2.4 Etudes des variables

2.4.1 Détermination des variables propres

```r
# determination des variables propres (dimension)
ValeurPropre <- get_eigenvalue(NewTab1)
print(ValeurPropre)
# tracer un histogramme pour nous aider a la prise de decision
fviz_eig(NewTab1,addlabels = T, ylim = c(0,50) )
```

Listing 2.6 – Variables propres

Pour déterminer combien d'axe nous avons besoins pout garder le maximum d'information on trace un histogramme pour nous aider à la prise de décision. Histogramme suivant donne une idée sur combien de variables propres (dimension).

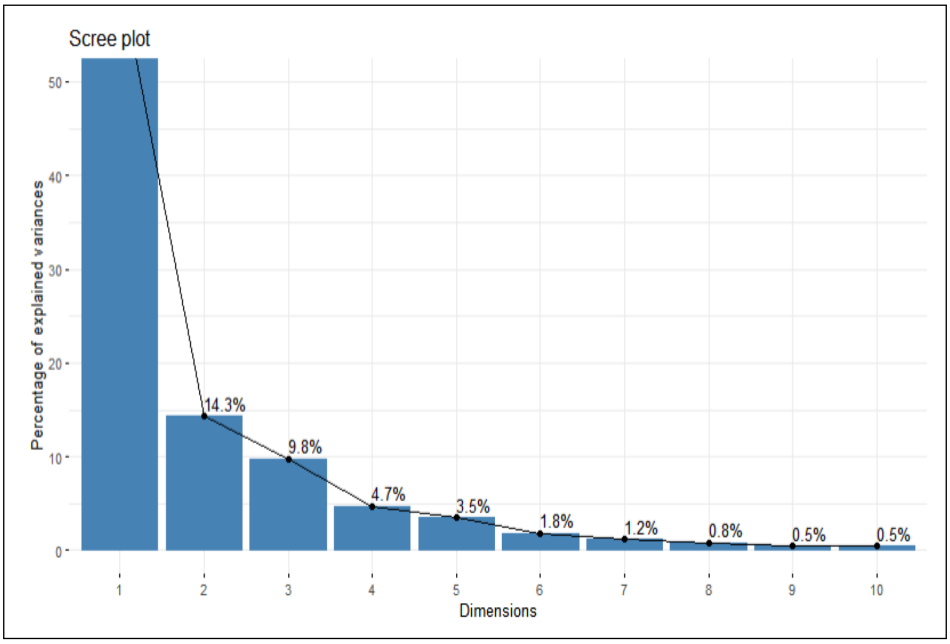


FIGURE 2.4 – Histogramme des variables propres

Analyse d'histogramme :

L'histogramme suivant donne une idée sur combien d'axe on doit garder pour résume le maximum d'informations. On constate que la dimension 1 résume 61.8% de l'information, si on ajoute les valeurs des deux autres dimension 2 et 3 on aura 85.8% de l'information ce qui est suffisante pour analyser les données,

```
        eigenvalue     variance.percent     cumulative.variance.percent
Dim.1   1.428190e+01         7.140952e+01             71.40952
Dim.2   2.601721e+00         1.300860e+01             84.41813
Dim.3   1.561314e+00         7.806571e+00             92.22470
Dim.4   4.944629e-01         2.472315e+00             94.69701
Dim.5   3.393252e-01         1.696626e+00             96.39364
```

2.4.2 Analyse des variables par l'ACP

Maintenant on va essayer d'étudier les variables par l'ACP à savoir les coordonnées la contribution et la qualité de représentation (cos2) de chaque varible.

```
# resultats trouver par L'ACP pour les variables
ResVariables <- get_pca_var(NewTab1)
print(ResVariables)
```

Listing 2.7 – Etude des Variables

```
Principal Component Analysis Results for variables
 ==================================================
    Name           Description
1   "$coord"       "Coordinates for the variables"
2   "$cor"         "Correlations between variables and dimensions"
3   "$cos2"        "Cos2 for the variables"
4   "$contrib"     "contributions of the variables"
```

Alors le résultat précèdent montre que l'étude des variables se fait par :

— L'étude des corrélations entre les variables.

— L'étude des coordonnes des variables

— L'étude de la contribution des variables.

— L'étude de la qualité de représentation des variables.

Nous on va s'intersectées à l'étude de la qualité de représentation et la contribution des données car ils sont les plus importants. Les autres indicateurs s'interprètent de la même manière.

L'étude de la qualité de représentation

```
print(ResVariables$cos2)
```

La commande précèdent affiche les résultats suivants :

```
            Dim.1        Dim.2        Dim.3        Dim.4        Dim.5
Peche    0.787273004 1.368135e-01 2.124270e-04 3.705427e-03 1.035850e-02
IE       0.264044519 1.389932e-01 5.280325e-01 1.768666e-02 1.627865e-02
IT       0.918318162 4.519641e-02 1.221766e-02 3.796005e-03 3.720333e-03
FPA      0.918732388 1.007321e-03 5.407437e-03 1.916609e-02 1.755804e-06
FPC      0.270792922 3.073942e-01 3.175360e-01 3.168888e-02 5.807102e-02
FPP      0.001861442 6.576818e-01 2.547746e-01 8.049420e-04 9.916083e-03
FO       0.691522371 1.277772e-01 2.014970e-02 9.352348e-02 6.108975e-03
```

Il est également possible de créer un bar plot du cosinus carré des variables en utilisant la fonction ***fviz_cos2()***

```
# Cos2 total des variables sur Dim.1 et Dim.2
fviz_cos2(NewTab1, choice = "var", axes = 1:2, top = 20)
```

Listing 2.8 – cos2 pour les variables

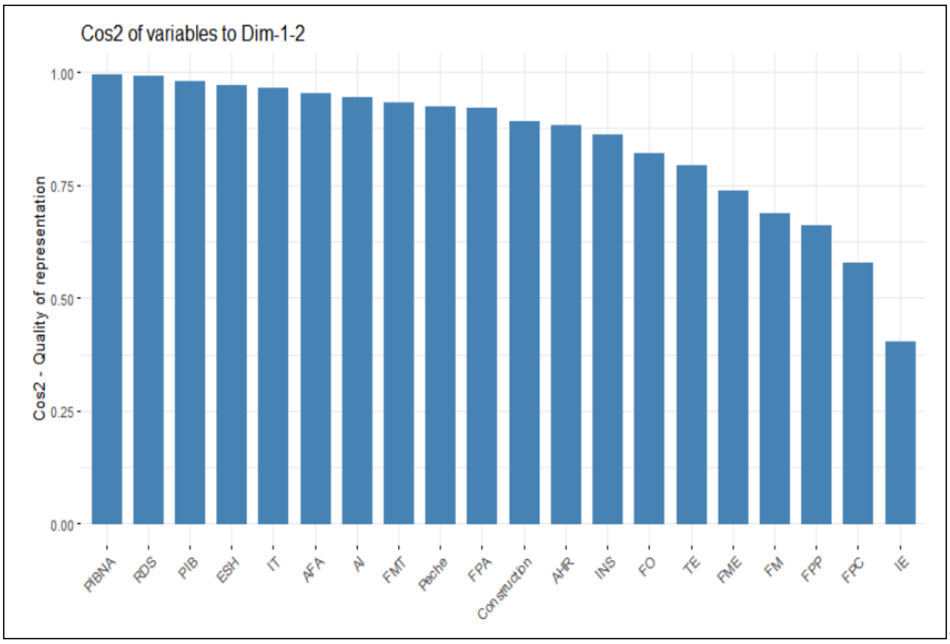


FIGURE 2.5 – Histogramme cos2 des variables

Maintenant on va colorer les variables en fonction de la valeur de leurs cos2, Cela produit un gradient de couleurs. Dans ce cas, l'argument ***gradient.cols*** peut être utilisé pour spécifier une palette de couleur personnalisée. Par exemple, ***gradient.cols = c("white", "blue", "red")*** signifie que :

— Les variables à faible valeur de cos2 seront colorées en “white” (blanc)

— Les variables avec les valeurs moyennes de cos2 seront colorées en “blue” (bleu)

— Les variables avec des valeurs élevées de cos2 seront colorées en “red” (rouge)

```
# Colorer en fonction du cos2: qualit  de repr sentation
fviz_pca_var(NewTab1, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE #  vite  le chevauchement de texte
)
```

Listing 2.9 – Diagramme de cos2

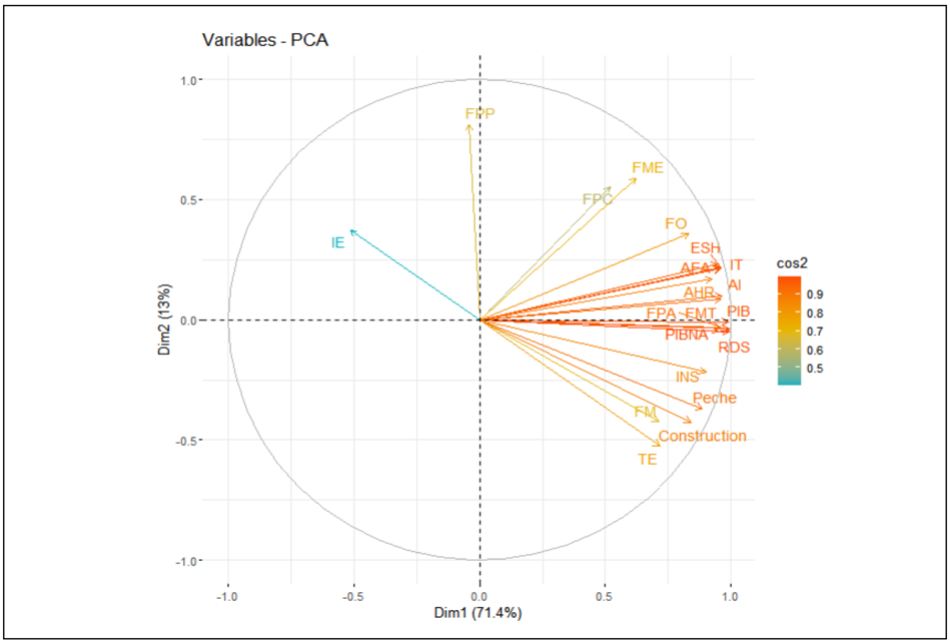


FIGURE 2.6 – Histogramme cos2 des variables

Notez que,

Un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation.

Un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.

<u>L'étude de la contribution</u>

```
# Resultat de la contribution
head(ResVariables$contrib, 5)
```

Listing 2.10 – contribution des variables

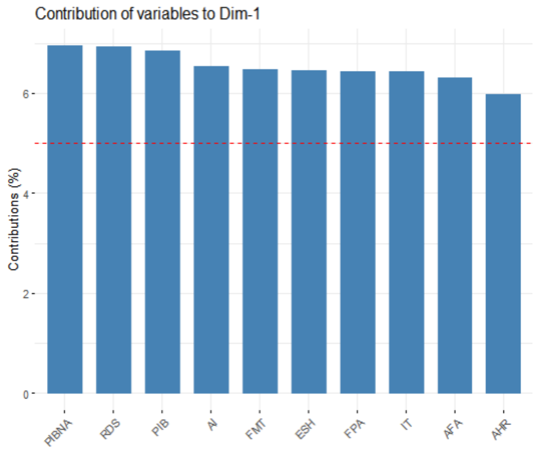
```
> head(ResVariables$contrib, 5)
            Dim.1        Dim.2       Dim.3      Dim.4        Dim.5
Peche   5.512381   5.25857909  0.01360565  0.7493843  3.052677e+00
IE      1.848805   5.34235817 33.81974330  3.5769427  4.797358e+00
IT      6.429942   1.73717364  0.78252436  0.7677026  1.096391e+00
FPA     6.432842   0.03871748  0.34633879  3.8761423  5.174398e-04
FPC     1.896056  11.81503417 20.33773883  6.4087481  1.711368e+01
```

Plus la valeur de la contribution est importante, plus la variable contribue à la composante principale en question.

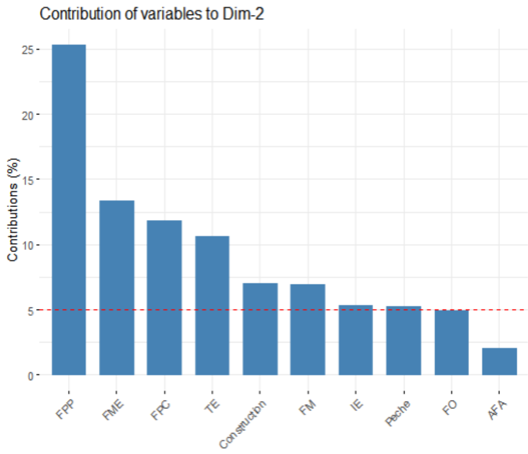
Alors nous avons maintenant créer un bar plot de la contribution des variables. On a nos données contiennent de nombreuses variables, on peut afficher seulement les principales variables contributives. Le code R ci-dessous montre le top 10 des variables contribuant le plus aux composantes principales :

```
#contribution des varibales au Dim1
fviz_contrib(NewTab1, choice = "var", axes = 1, top = 10)
#contribution des varibales au Dim2
fviz_contrib(NewTab1, choice = "var", axes = 2, top = 10)
```

Listing 2.11 – Contributions des variables



(a) Contribution Variables Dim1



(b) Contribution Variables Dim1

FIGURE 2.7 – Contribution des Variables pour les deux dimension

l'interpretation du graphe

La ligne en pointillé rouge, sur le graphique ci-dessus, indique la contribution moyenne attendue. Pour une composante donnée, une variable avec une contribution supérieure à ce seuil pourrait être considérée comme importante pour contribuer à la composante.

On peut voir que les variables PIBNA, RDS et PIB - contribuent le plus aux dimensions 1 par contre FPP et FME.

Les variables les plus importantes (ou, contributives) peuvent être mises en évidence sur le graphe de corrélation comme suit :

```
# les varibales les plus contibutives sur le cercle de correlation
fviz_pca_var(NewTab1, col.var = "contrib",
             gradient.cols = c("blue", "yellow", "red")
)
```

Listing 2.12 – Contributions des variables

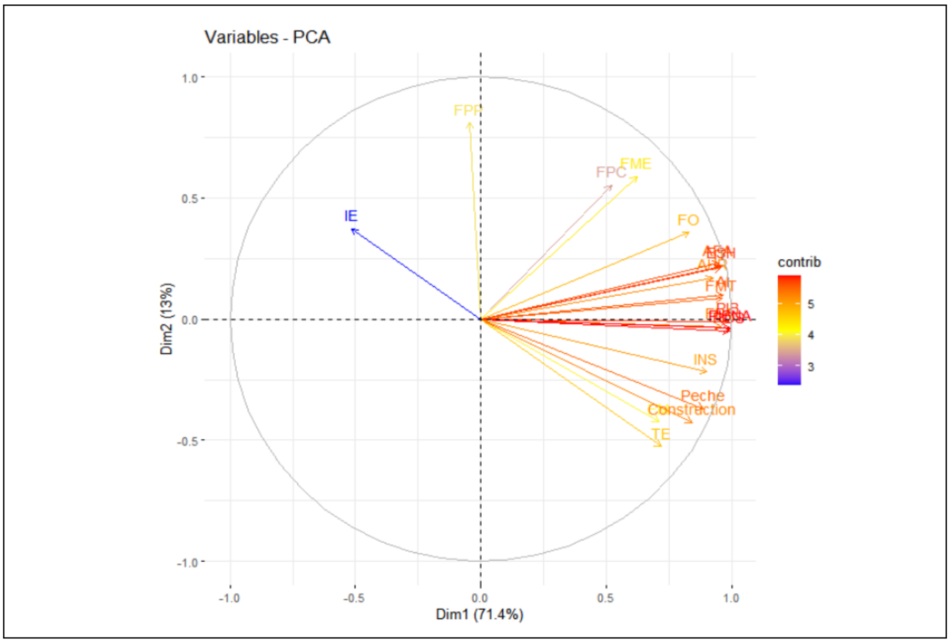


FIGURE 2.8 – variables les plus contributives

2.5 Etudes des individus

L'analyse des individus suivis les mêmes étapes que les variables. Les résultats, pour les individus, peuvent être extraits à l'aide de la fonction ***get_pca_ind()*** retourne une liste de matrices contenant tous les résultats pour les individus (coordonnées, corrélation entre individus et axes, cosinus-carré et contributions)

```
ResIndividus <- get_pca_ind(NewTab1)
head(ResIndividus$coord, 5)
head(ResIndividus$cos2, 5)
head(ResIndividus$contrib, 5)
```

Listing 2.13 – Etude des individus

```
> head(ResIndividus$coord, 5)
          Dim.1      Dim.2      Dim.3       Dim.4      Dim.5
14_S1 -6.455169 1.02594984 -2.2102972 -0.01520276 -0.8558741
14_S2 -5.166807 1.59290052 -1.7194405 -0.07071612  0.4997962
14_S3 -4.582765 1.82049765 -1.1759988 -0.11728862  0.4734538
14_S4 -4.772240 0.05898898  0.4775466  0.32247446 -0.3827657
15_S1 -3.738631 1.26417161  1.5148770  0.39443121 -0.2148088
```

```
> head(ResIndividus$cos2, 5)
           Dim.1        Dim.2       Dim.3        Dim.4       Dim.5
14_S1 0.8545622 0.0215864222 0.100191241 4.739948e-06 0.015022708
14_S2 0.8123565 0.0772110637 0.089965613 1.521735e-04 0.007601303
14_S3 0.7910472 0.1248324502 0.052090801 5.181541e-04 0.008443109
14_S4 0.9323333 0.0001424519 0.009335941 4.257134e-03 0.005997809
15_S1 0.7439238 0.0850580260 0.122140017 8.280294e-03 0.002455882
```

```
> head(ResIndividus$contrib, 5)
           Dim.1       Dim.2      Dim.3       Dim.4      Dim.5
14_S1 14.588114 2.022840223 15.6451970 0.002337122 10.7937824
14_S2  9.346055 4.876257556  9.4679075 0.050567691  3.6807785
14_S3  7.352567 6.369268715  4.4288756 0.139106686  3.3030032
14_S4  7.973122 0.006687304  0.7303166 1.051542742  2.1588370
15_S1  4.893382 3.071294029  7.3491048 1.573181483  0.6799204
```

L'étude de la qualité de représentation

Comme le cas des variables il est possible d'étudier la qualité de représentation des individus et les organiser dans un diagramme colorer en fonction de leurs valeurs de cos2.

```
# qualite de representation des individus
fviz_pca_ind(NewTab1, col.ind = "cos2", gradient.cols
             = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping (slow if many points)
)
```

Listing 2.14 – cos2 des individus

Vous remarquerais sur le graphe suivant que les individus qui sont similaire sont regrouper sur le graphique.

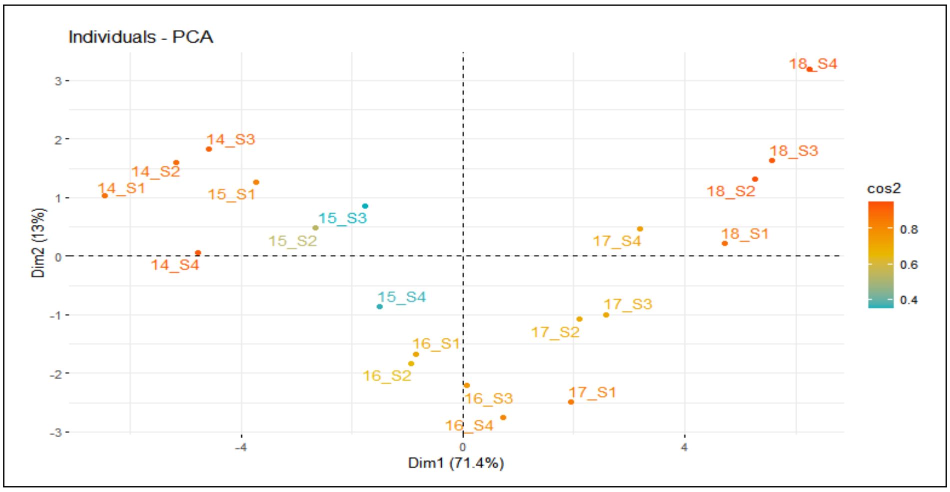


FIGURE 2.9 – qualité de représentation des individus avec couleurs

Dans le cas d'individus, au lieu d'utiliser des couleurs, nous pouvons également utiliser la taille du point en fonction de cos2, comme indiqué dans le schéma suivant :

```
fviz_pca_ind (NewTab1 , pointsize = "cos2",
             pointshape = 21, fill = "#E7B800",
             repel = TRUE #  vite  le chevauchement de texte
)
```

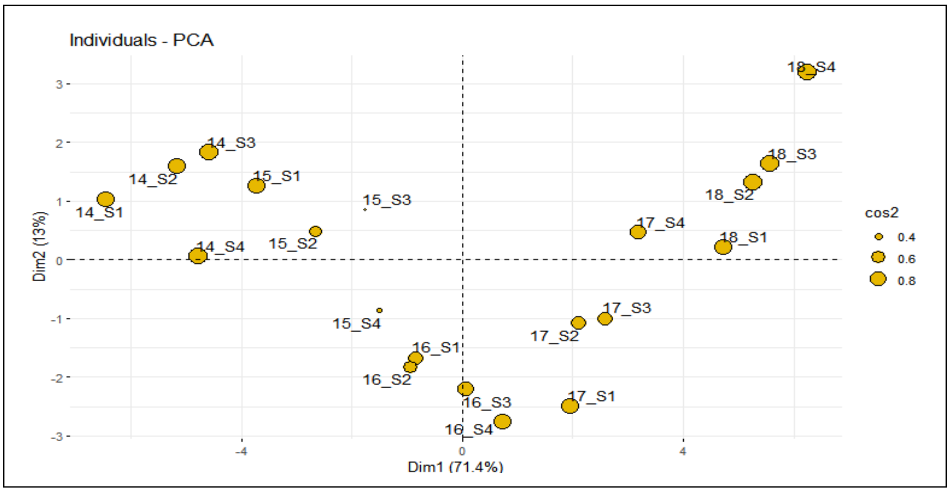
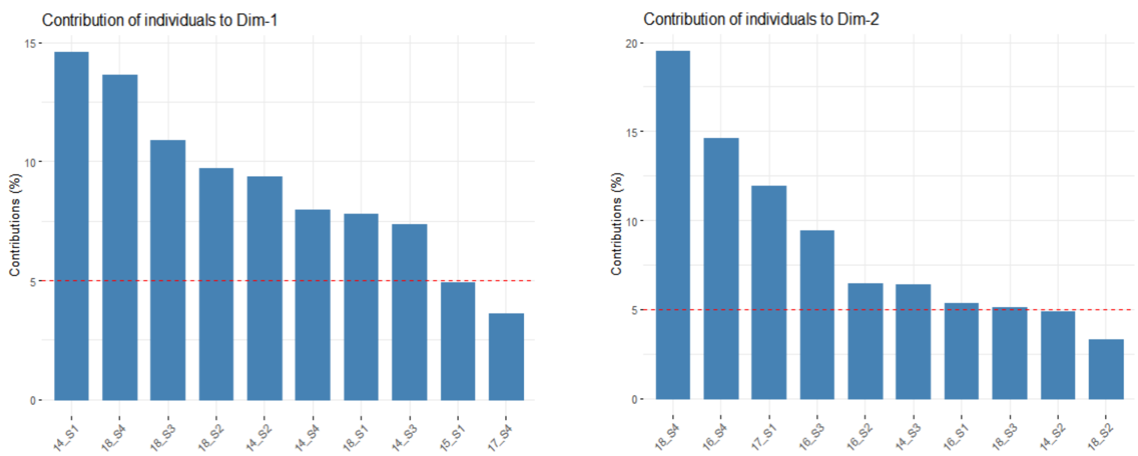


FIGURE 2.10 – qualité de représentation des individus avec des points

L'étude de la contribution

```
# 2- contribution des individus aux axes Dim1 et Dim2
fviz_contrib(NewTab1, choice = "ind", axes = 1, top = 10)
fviz_contrib(NewTab1, choice = "ind", axes = 2, top = 10)
```

Listing 2.15 – Contributions des individus



(a) Contribution individus Dim1 **(b) Contribution individus Dim1**

FIGURE 2.11 – Contribution des individus pour les deux dimension

2.6 Etudes des groups

Dans cette partie on colorer les individus par groupes, en utilisant des ellipses de concentration et des ellipses de confiances par grappes. Le jeu de données iris ressemble à ceci.

```
head(iris, 5)
# La variable Species (index = 5) est supprim e
print(PCA(iris [, - 5], graph = FALSE))
```

Listing 2.16 – Etude par groupes

```
> head(iris, 5)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
```

```
> print(PCA(iris [, - 5], graph = FALSE))
  name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. for the variables"
4  "$var$cor"           "correlations variables - dimensions"
5  "$var$cos2"          "cos2 for the variables"
6  "$var$contrib"       "contributions of the variables"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$call"              "summary statistics"
12 "$call$centre"       "mean of the variables"
13 "$call$ecart.type"   "standard error of the variables"
14 "$call$row.w"        "weights for the individuals"
15 "$call$col.w"        "weights for the variables"
```

Diagramme pour les groups

```
fviz_pca_ind(IrisPca,
             geom.ind = "point", # Montre les points seulement
             col.ind = iris$Species, # colorer by groups
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, # Ellipses de concentration
             legend.title = "Groups"
)
```

Listing 2.17 – Diagramme des groupes

Dans le code R ci-dessous : l'argument habillage ou col.ind peut être utilisé pour spécifier la variable à utiliser pour colorer les individus par groupes.

Pour ajouter une ellipse de concentration autour de chaque groupe, spécifiez l'argument addEllipses = TRUE. L'argument palette peut être utilisé pour changer les couleurs du groupe.

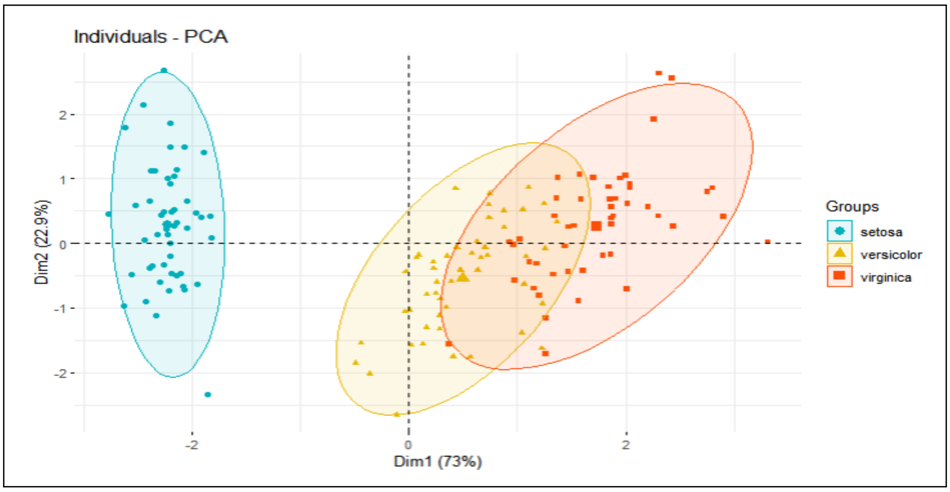


FIGURE 2.12 – Graphe des Groupes

2.7 Conclusion

Alors nous avons présenté axes possibles pour étudier l'ACP à savoir l'étude de la contribution et la qualité de représentation des variables et des individus. Dans le chapitre suivant on va interpréter les résultats obtenus et les liées avec nos donnes.

Chapitre

3

Interprétations et Conclusions

Sommaire

3.1 Introduction

Dans ce chapitre on va interpréter les résultats obtenus des variables et des individus, et finalement on va essayer de sortir avec des conclusions concernant nos données.

3.2 Interprétation des variables

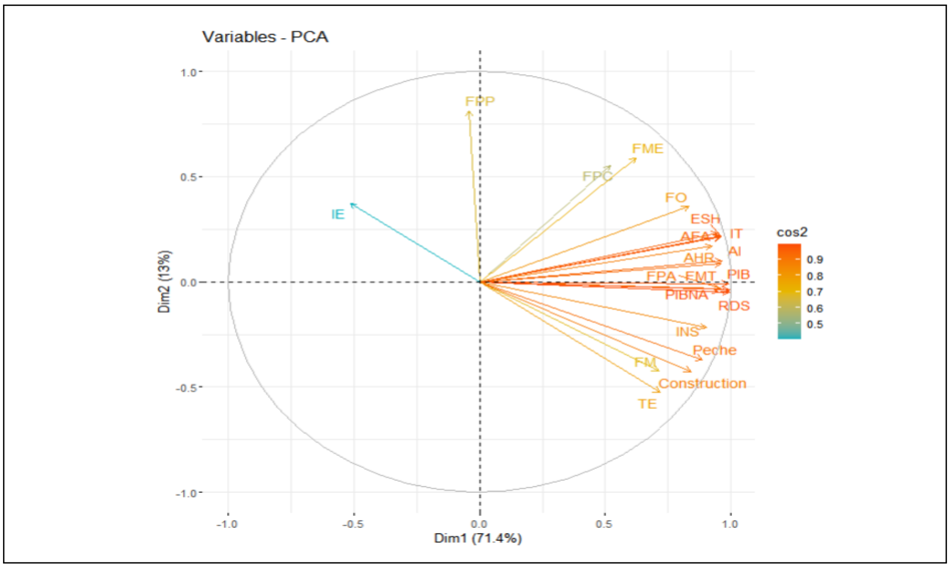


FIGURE 3.1 – Variables PCA

On remarque que la plupart des variables sont proche de l'axe 1 c qui est normal car on déjà dit que l'axe 1 résume 71% de l'information. Les variables qui sont corrélé positivement avec l'axe 1 ont une qualité de représentation très importantes.

Le domaine IE (industrie d'Extraction) il corrélé négativement avec l'axe 1 ce qui explique pourquoi il a une qualité de représentation très faibles 0.264.

La majorité des autre domaine PIB 0.97, RDS 0.98 et AI 0.93 ont une valeur de cos2 très importante(c-à-d proche de 1) .

Finalement on constate que les variables qui sont corrélé avec l'axe 2 ont une qualité de représentation moyen

3.3 Interprétation des individus

QUELS SONT LES POINTS QUI NOUS INTÉRESSENT ? ?

Les points les plus intéressants sont généralement ceux qui sont assez proches d'un des axes, et assez loin de l'origine. Ces points sont bien corrélés avec cet axe et sont les points explicatifs pour l'ax.

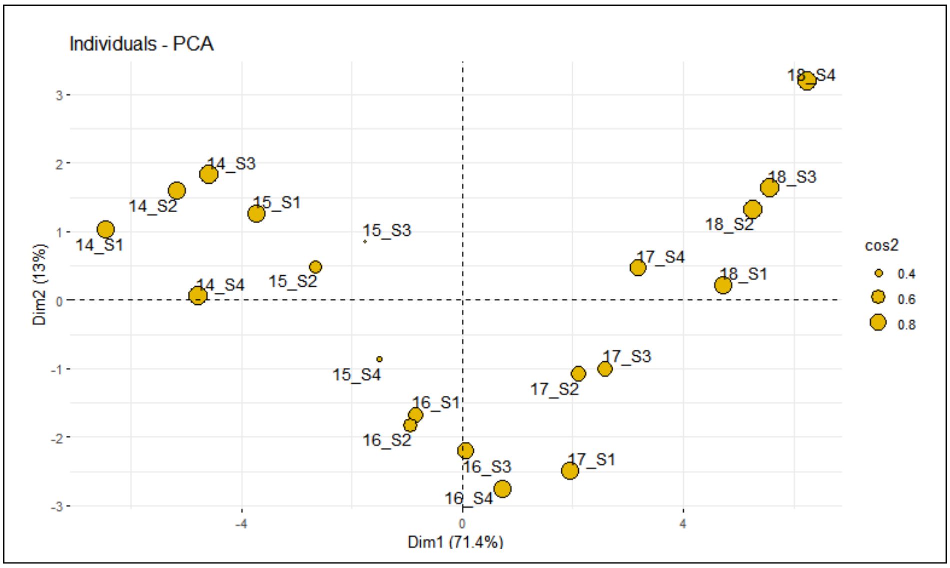


FIGURE 3.2 – Individus PCA

- Les points grands représentent une qualité de représentation très importante .
- Les points moyens représentent une qualité de représentation importante.
- Les petits points représentent une faible représentation

Alors on remarque que les années qui sont corrélé avec l'axe 1 et 2 et loin du centre de diagramme Ont une qualité de représentation très important, Ex : les années 2018 et 2014.

Par contre que les années situées près du centre sont donc généralement mal représentées par le plan factoriel. Leur interprétation ne peut donc pas être effectuée avec confiance.

3.4 Interprétation Des données réels

On s'intéresse maintenant aux points de proximité (c.-à-d situés loin du centre) les deux points qui sont proche l'une de l'autre dans le diagramme, Alors les individus qu'ils représentent soient très similaires.

Il se peut que sur un axe ils sont très proche, alors que sur un autre ils sont très loin l'un de l'autre. Il faut donc les regarder par rapport à tous les axes qui ont été retenus pour l'analyse. S'ils sont bien corrélés avec l'axe qui les montre proche, alors, on peut conclure qu'ils sont vraiment proches.

3.5 Conclusions

La majorité des années subit une augmentation de valeurs Ajouter dans les 3 premiers sessions et subit une démunissions dans la session 4 de chaque année. Et la première session de l'année suivant départ du prix de la 4 -ème session de l'année précèdent.

Si on observe le diagramme des variables on constate que la majorité des domaines sont corrélé positivement avec soit l'axe 1 soit 2 sauf le domaine IE (Industrie d'Extraction), alors cela montre que tous les domaines subissent une augmentation du Valeur Ajouter chaque année, contrairement au IE subit une progression fluctuante.

CONCLUSION GÉNÉRALE

L'analyse des données joue un rôle clé dans le processus d'évaluation de la qualité des données en indiquant les problèmes liés à la qualité des données dans une enquête particulière. Ainsi, l'analyse peut influer sur les améliorations futures au processus d'enquête.

L'analyse en composantes principales (ACP) est un outil extrêmement puissant de synthèse de l'information, très utile lorsque l'on est en présence d'une somme importante de données quantitatives à traiter et interpréter.

L'apparition au cours des dernières années de logiciels chaque fois plus performants et faciles à utiliser rend aujourd'hui accessible ce type d'analyses des données à tous les chercheurs en sciences sociales, et non plus aux seuls spécialistes. C'est pourquoi nous proposons ici de présenter le principe et l'intérêt de l'ACP à partir d'un exemple simple, celui d'une analyse portant sur les trente-deux entités fédérales du Mexique fondée sur 12 variables démographiques socio-économiques et culturelles.

BIBLIOGRAPHIE

[1] **WIKIPEDIA**. ST Company Information **[en ligne]**. Mis à jour en December 2022. Disponible sur : https ://www.wikipedia.org/

[2] **XLSTATE**. Pour les notions due l'ACP **[en ligne]**. Disponible sur :
https ://www.xlstat.com/fr/solutions/fonctionnalites/analyse-en-composantes-principales-acp

[3] **StackOverFlow**. pour resoudres les problemes du code **[en ligne]**. Disponible sur :
https ://stackoverflow.com/

[4] **HCP**. pour trouver Nos donnees. Disponible sur :
https ://www.hcp.ma/

[5] **STHDA**. pour nous aider a trouver le code sources et comprendre les etapes de l'ACP. Disponible sur : http ://www.sthda.com/french/