

Notes de cours et codes sur l'ACP

Zaineb Smida

Table des matières

1	Importation des données	1
2	Analyse Descriptive	2
2.1	Analyse Univariée	2
2.2	Analyse Bivariée	4
3	Analyse en Composantes Principales	7
3.1	ACP : Valeurs propres	8
3.2	ACP : Interprétation des variables	9
3.3	ACP : Interprétation des individus	15
3.4	Pour aller plus loin	20

Packages nécessaires :

```
install.packages(c("FactoMineR", "ggcorrplot", "kableExtra"))
```

1 Importation des données

Pour importer un fichier texte sous **R**, on utilise la fonction `read.table()`

```
regions <- read.table("data/regions.txt", header = TRUE)
```

Tableau 1 : Affichage des premières lignes du fichier

```
head(regions)
```

	NOM	REGION	POPUL	TACT	SUPERF	NBENTR	NBBREV	CHOM	TELEPH
1	A	Alsace	1624	39.14	8280	35976	241	5.2	700
2	Q	Aquitain	2795	36.62	41308	85531	256	10.2	1300
3	U	Auvergne	1320	37.48	26013	40494	129	9.3	600
4	N	Bas-Norm	1390	38.63	17589	35888	91	9.0	600
5	O	Bourgogn	1600	38.26	31582	40714	223	8.1	750
6	B	Bretagne	2795	36.62	27208	73763	296	9.5	1300

Pour chaque région de France, on observe :

- POPUL : population de la région (en milliers d'individus)
- TACT : taux d'activité (population active / population totale de la région) en pourcentage
- SUPERF : superficie de la région (en kilomètres carrés)
- NBENTR : nombre d'entreprises de la région
- NBBREV : nombre de brevets déposés au cours de l'année
- CHOM : taux de chômage (en pourcentage)
- TELEPH : nombre de lignes téléphoniques en place dans la région (en milliers)

On ajoute le nom des individus au `data.frame` :

```
row.names(regions) <- regions$REGION
```

2 Analyse Descriptive

2.1 Analyse Univariée

On considère dans un premier temps des outils statistiques utilisés pour décrire les variables quantitatives une par une.

- Résumé statistique des variables quantitatives (minimum, maximum, moyenne, médiane, quartiles) :

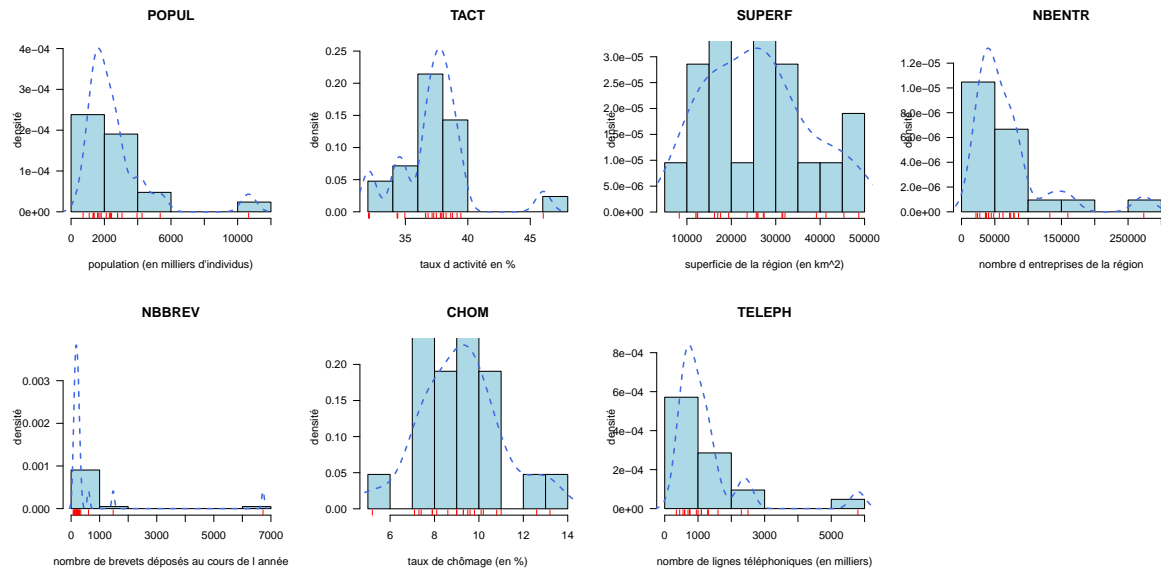
```
summary(regions[, -c(1, 2)])
```

POPUL	TACT	SUPERF	NBENTR
Min. : 720	Min. : 32.05	Min. : 8280	Min. : 21721
1st Qu.: 1590	1st Qu.: 36.62	1st Qu.: 16942	1st Qu.: 36285
Median : 2110	Median : 37.48	Median : 25809	Median : 48353
Mean : 2681	Mean : 37.23	Mean : 25728	Mean : 69827
3rd Qu.: 2795	3rd Qu.: 38.26	3rd Qu.: 31582	3rd Qu.: 78504
Max. : 10660	Max. : 46.04	Max. : 48698	Max. : 273604
NBBREV	CHOM	TELEPH	
Min. : 73.0	Min. : 5.200	Min. : 350	
1st Qu.: 155.0	1st Qu.: 7.900	1st Qu.: 700	
Median : 223.0	Median : 9.300	Median : 950	
Mean : 587.1	Mean : 9.186	Mean : 1262	
3rd Qu.: 278.0	3rd Qu.: 10.100	3rd Qu.: 1300	
Max. : 6722.0	Max. : 13.200	Max. : 5800	

- Affichage des distributions, en utilisant des histogrammes et densités non paramétriques:

Figure 1 : Histogrammes :

```
par(mfrow = c(2, 4), las = 1)
nom_var <- c("population (en milliers d'individus)",
  "taux d'activité en %",
  "superficie de la région (en km^2)",
  "nombre d'entreprises de la région",
  "nombre de brevets déposés au cours de 1 année",
  "taux de chômage (en %)",
  "nombre de lignes téléphoniques (en milliers)")
for(k in 3:9) {
  temp <- density(regions[, k])
  hist(regions[, k], main = names(regions)[k],
    xlab = nom_var[k-2], ylab = "densité", probability = T,
    col = "lightblue", ylim = range(temp$y))
  lines(temp, col = "royalblue", lty = 2, lwd = 1.5)
  rug(regions[, k], col = "red")
}
```



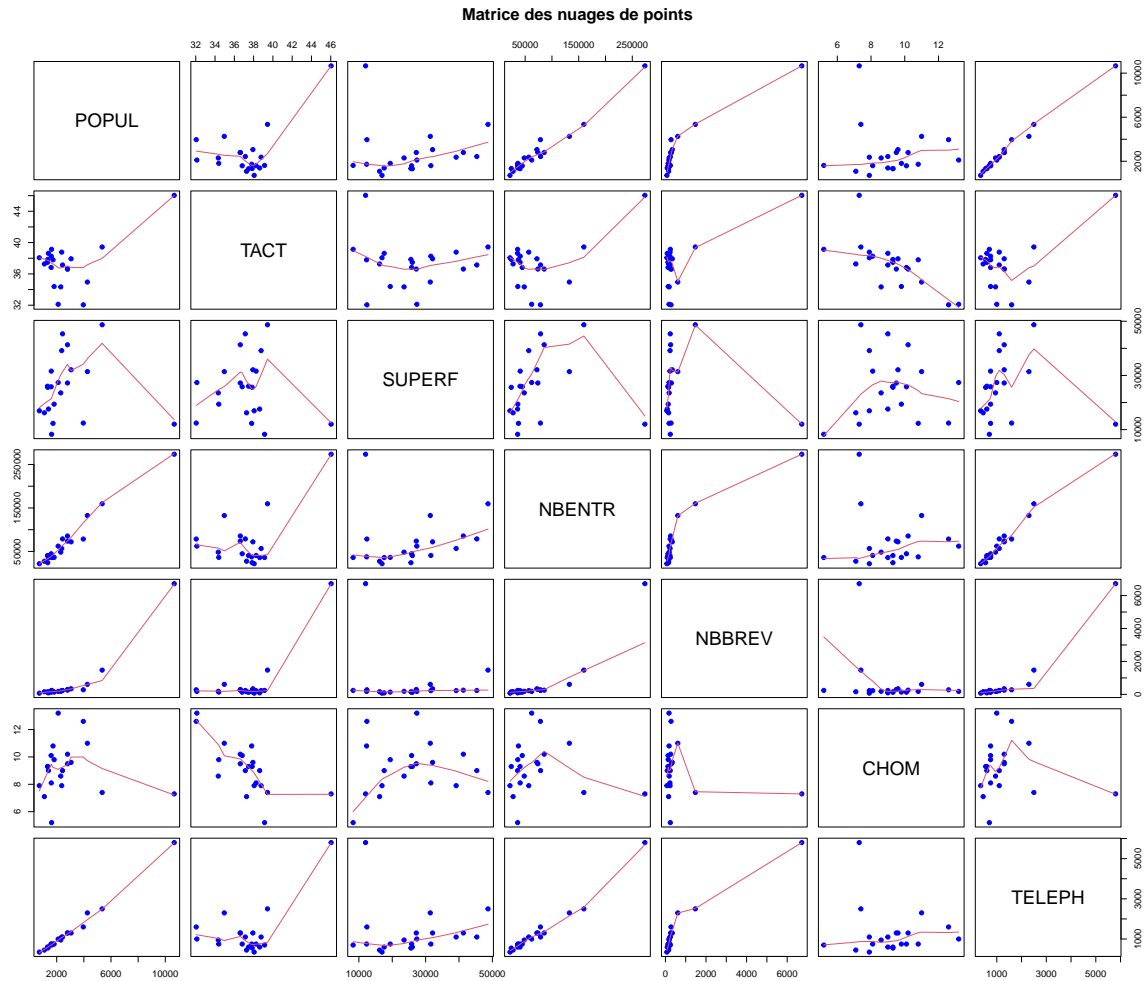
Remarque : les ordres de grandeurs sont parfois très différents d'une variable à une autre. En effet, certaines variables sont des comptages (POPUL, SUPERF), alors que d'autres sont des pourcentages (CHO); c'est pourquoi, il pourrait être intéressant de reproduire ce graphique sur les données centrées et réduites. Par ailleurs, on observe quelques valeurs atypiques (par exemple, NBBREV ou POPUL ont chacune une valeur extrêmement forte par rapport aux autres; il s'agit probablement de la région Ile-de-France)

2.2 Analyse Bivariée

- On trace les nuages de points entre chaque paire de variable, ce qui permet de détecter d'éventuels liens (linéaires ou non linéaires) entre les paires de variables ainsi que des valeurs atypiques. Par exemple, on constate que le lien entre certaines variables (comme SUPERF et NBENTR) n'est pas linéaire, à cause notamment d'une valeur (il s'agit de l'Ile-de-France dont la superficie est très faible comparativement au nombre d'entreprises qu'elle possède).

Figure 2 : Nuage de points :

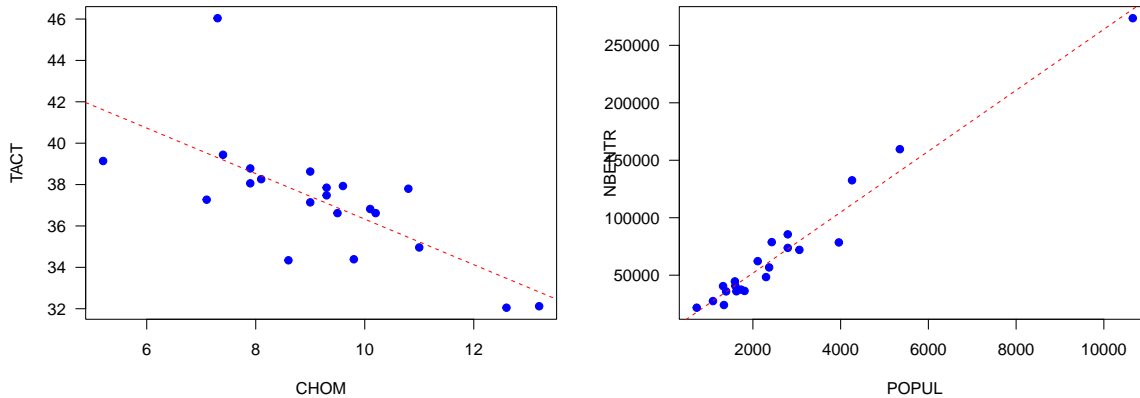
```
pairs(regions[, 3:9], main = "Matrice des nuages de points",
      pch = 19, col = "blue", panel = panel.smooth)
```



On peut tracer directement les nuages de points qui nous intéressent et représenter par la même occasion la droite de régression linéaire. On représente ici les liens les plus forts.

```
par(las = 1, mfrow = c(1, 2))
plot(TACT ~ CHOM, data = regions,
     pch = 19,
     col = "blue")
abline(lm(TACT ~ CHOM, data = regions), col = "red", lty = 2)

plot(NBENTR ~ POPUL, data = regions,
     pch = 19,
     col = "blue")
abline(lm(NBENTR ~ POPUL, data = regions), col = "red", lty = 2)
```



Enfin, on peut représenter la matrice de corrélations linéaires :

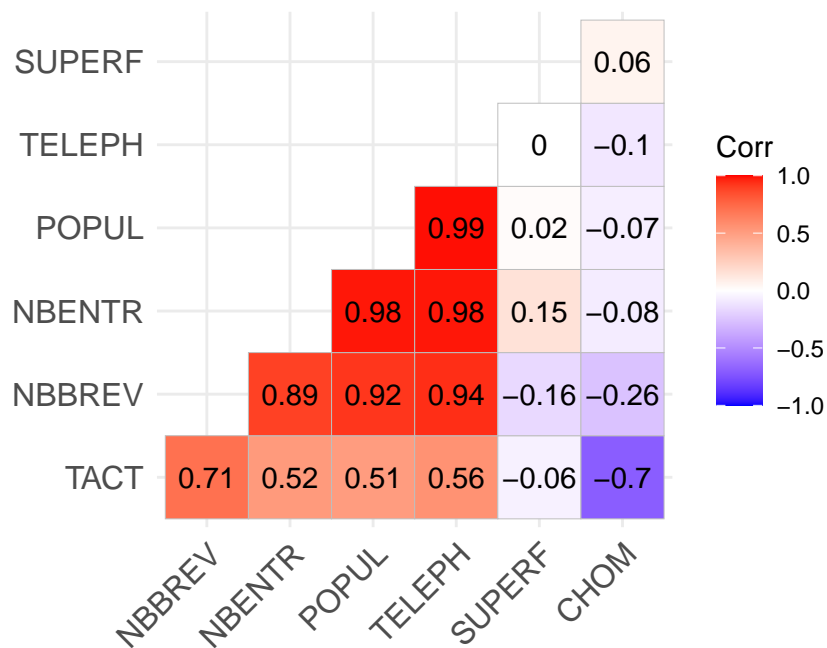
Figure 3 : Matrice des corrélations :

```
library("ggcorrplot")
cor <- cor(regions[3:9])
cor
```

	POPUL	TACT	SUPERF	NBENTR	NBBREV	CHOM
POPUL	1.00000000	0.51376438	0.024369703	0.98101936	0.9213741	-0.07313003
TACT	0.51376438	1.00000000	-0.059255061	0.51571338	0.7084501	-0.69854149
SUPERF	0.02436970	-0.05925506	1.000000000	0.14929185	-0.1639580	0.06205849
NBENTR	0.98101936	0.51571338	0.149291848	1.00000000	0.8916071	-0.07804957
NBBREV	0.92137414	0.70845007	-0.163957955	0.89160714	1.0000000	-0.25657627
CHOM	-0.07313003	-0.69854149	0.062058491	-0.07804957	-0.2565763	1.00000000
TELEPH	0.99391186	0.55526402	0.004764791	0.98290899	0.94444463	-0.09833108
TELEPH						
POPUL	0.993911864					
TACT	0.555264016					
SUPERF	0.004764791					
NBENTR	0.982908993					
NBBREV	0.9444446274					
CHOM	-0.098331084					
TELEPH	1.000000000					

ou bien, on peut représenter les résultats comme suit afin de mieux apprécier les corrélations positives/négatives (couleurs bleues/rouges):

```
ggcorrplot(cor, hc.order = TRUE, type = "lower", lab = TRUE)
```



Remarque : la variable POPUL est très positivement (linéairement) corrélée à TELEPH et NBENTR, ce qui est logique, puisque plus une région est peuplée, plus il doit y avoir de lignes téléphoniques et d'entreprises. On constate aussi que CHOM est négativement (linéairement) corrélée à TACT : plus il y a du chômage, moins le taux d'activité est élevé. L'ACP semble être une bonne méthode afin de bien appréhender l'ensemble de ces corrélations non nulles.

3 Analyse en Composantes Principales

La fonction `PCA()` du package **FactoMineR** retourne par défaut le graphique des individus et le graphique des variables. En utilisant l'option `graph = FALSE`, on empêche la représentation graphique. En effet, avant de regarder le graphique des individus et des variables, il peut être intéressant de regarder un certain nombre d'informations.

```
library("FactoMineR")
res <- PCA(regions[, 3:9], graph = FALSE)
```

On récupère les informations qui nous intéressent, à savoir les valeurs propres, mais aussi les informations relatives aux individus et aux variables ainsi.

```
eigenvalues <- res$eig  
res_individus <- res$ind  
res_variabilites <- res$var
```

3.1 ACP : Valeurs propres

A partir des informations retournées, on affiche les valeurs propres et les parts d'inertie expliquées par chaque composante :

Tableau : ACP-Valeurs propres :

```
eigenvalues
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.329675886	61.85251266	61.85251
comp 2	1.429382161	20.41974516	82.27226
comp 3	1.012436783	14.46338261	96.73564
comp 4	0.182765737	2.61093910	99.34658
comp 5	0.032756318	0.46794741	99.81453
comp 6	0.010720602	0.15315145	99.96768
comp 7	0.002262513	0.03232161	100.00000

Commentaires :

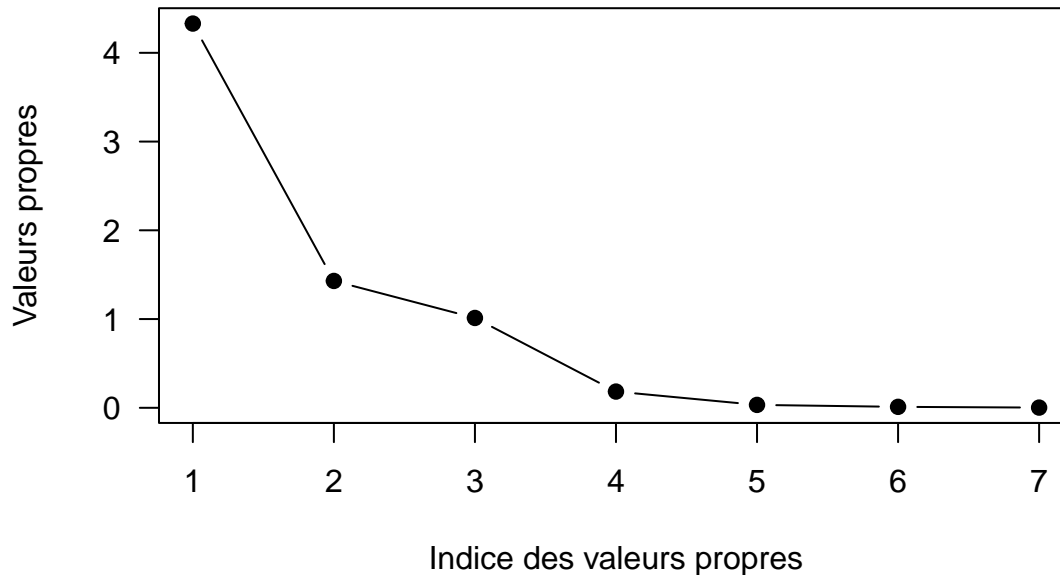
- les deux premiers axes expliquent à eux seuls 80% de la variance totale.
- les trois premières valeurs propres sont supérieures à 1.

On trace à présent l'éboulis des valeurs propres :

Figure : ACP-Eboulis des valeurs propres :

```
par(las = 1)  
plot(eigenvalues[,1], type = "b", pch = 19,  
      xlab = "Indice des valeurs propres",  
      ylab = "Valeurs propres",  
      main = "Eboulis des valeurs propres")
```


Eboulis des valeurs propres



Commentaires : d'après l'éboulis des valeurs propres, on voit un coude à $k = 4$. Donc, par ce critère, on retient uniquement les trois premières composantes principales.

3.2 ACP : Interprétation des variables

- On affiche les informations relatives aux variables :

Tableau : ACP-Variables :

```
res_variables
```

```
$coord
      Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
POPUL  0.95783293  0.25547751 -0.04443395 -0.08882877 -0.04217485
TACT   0.72719947 -0.59263111  0.14930416  0.30773685 -0.05438998
SUPERF -0.01552123  0.33187207  0.94202679  0.03199371  0.03341067
NBENTR  0.94885018  0.27975109  0.08090534 -0.07645690 -0.05962944
NBBREV  0.97349367 -0.02238409 -0.15753009  0.06195308  0.15175965
CHOM   -0.29985584  0.88117115 -0.25791205  0.25882737 -0.01061849
TELEPH  0.97224065  0.21803299 -0.05362987 -0.04974105 -0.01427041
```

\$cor

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
POPUL	0.95783293	0.25547751	-0.04443395	-0.08882877	-0.04217485
TACT	0.72719947	-0.59263111	0.14930416	0.30773685	-0.05438998
SUPERF	-0.01552123	0.33187207	0.94202679	0.03199371	0.03341067
NBENTR	0.94885018	0.27975109	0.08090534	-0.07645690	-0.05962944
NBBREV	0.97349367	-0.02238409	-0.15753009	0.06195308	0.15175965
CHOM	-0.29985584	0.88117115	-0.25791205	0.25882737	-0.01061849
TELEPH	0.97224065	0.21803299	-0.05362987	-0.04974105	-0.01427041

\$cos2

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
POPUL	0.9174439224	0.0652687574	0.001974376	0.007890550	0.0017787178
TACT	0.5288190696	0.3512116276	0.022291733	0.094701967	0.0029582695
SUPERF	0.0002409085	0.1101390690	0.887414480	0.001023598	0.0011162730
NBENTR	0.9003166591	0.0782606742	0.006545675	0.005845657	0.0035556707
NBBREV	0.9476899200	0.0005010477	0.024815730	0.003838185	0.0230309906
CHOM	0.0899135264	0.7764626006	0.066518627	0.066991609	0.0001127524
TELEPH	0.9452518801	0.0475383848	0.002876163	0.002474172	0.0002036445

\$contrib

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
POPUL	21.189667460	4.56622163	0.1950122	4.3173024	5.4301516
TACT	12.213825781	24.57086965	2.2017901	51.8160400	9.0311415
SUPERF	0.005564123	7.70536194	87.6513473	0.5600599	3.4078098
NBENTR	20.794089045	5.47513998	0.6465268	3.1984427	10.8549154
NBBREV	21.888241635	0.03505345	2.4510893	2.1000570	70.3100704
CHOM	2.076680306	54.32155386	6.5701512	36.6543586	0.3442158
TELEPH	21.831931650	3.32579950	0.2840832	1.3537394	0.6216954

- On affiche les corrélations relatives aux variables :

```
res_variables$cor[,1:3]
```

	Dim.1	Dim.2	Dim.3
POPUL	0.95783293	0.25547751	-0.04443395
TACT	0.72719947	-0.59263111	0.14930416
SUPERF	-0.01552123	0.33187207	0.94202679
NBENTR	0.94885018	0.27975109	0.08090534
NBBREV	0.97349367	-0.02238409	-0.15753009
CHOM	-0.29985584	0.88117115	-0.25791205
TELEPH	0.97224065	0.21803299	-0.05362987

- On affiche les coordonnées des variables dans les trois premiers axes :

Tableau : ACP-Variables-Coordonnées :

```
res_variables$coord[, 1:3]
```

	Dim.1	Dim.2	Dim.3
POPUL	0.95783293	0.25547751	-0.04443395
TACT	0.72719947	-0.59263111	0.14930416
SUPERF	-0.01552123	0.33187207	0.94202679
NBENTR	0.94885018	0.27975109	0.08090534
NBBREV	0.97349367	-0.02238409	-0.15753009
CHOM	-0.29985584	0.88117115	-0.25791205
TELEPH	0.97224065	0.21803299	-0.05362987

Remarque : dans le cas d'une ACP réalisée sur la matrice de corrélation, les corrélations entre les variables et les axes correspondent aux coordonnées des variables sur les axes.

- On peut ajouter des couleurs afin de visualiser les variables les plus corrélées aux axes, en utilisant une petite fonction créée par nous-mêmes :

```
library(kableExtra)
```

Warning: le package 'kableExtra' a été compilé avec la version R 4.4.2

```
print_table <- function(my_tab, position = "H") {
  my_tab[, 1:ncol(my_tab)] <- lapply(my_tab[, 1:ncol(my_tab)],
    function(x) {
      cell_spec(round(x, 3), bold = T, color = spec_color(abs(x), end = 0.9),
        font_size = spec_font_size(abs(x)))
    })
  kbl(my_tab, escape = F, align = "c", position = position) %>%
    kable_styling(position = 'center', latex_options = 'HOLD_position') %>%
    kable_classic("striped", full_width = F)
}
```

- On affiche le tableau avec les couleurs pour visualiser correctement les corrélations :

```
print_table(as.data.frame(res_variables$coord)[, 1:3], position = "h")
```

	Dim.1	Dim.2	Dim.3
POPUL	0.958	0.255	-0.044
TACT	0.727	-0.593	0.149
SUPERF	-0.016	0.332	0.942
NBENTR	0.949	0.28	0.081
NBBREV	0.973	-0.022	-0.158
CHOM	-0.3	0.881	-0.258
TELEPH	0.972	0.218	-0.054

- on s'intéresse à présent aux contributions des variables pour déterminer quelles sont les variables qu'il faudra interpréter. Une forte contribution sera une contribution supérieure à $100/7 = 14.29$.

Tableau: ACP-Variables-Contributions :

```
print_table(as.data.frame(res_variables$contrib)[, 1:3], position = "h")
```

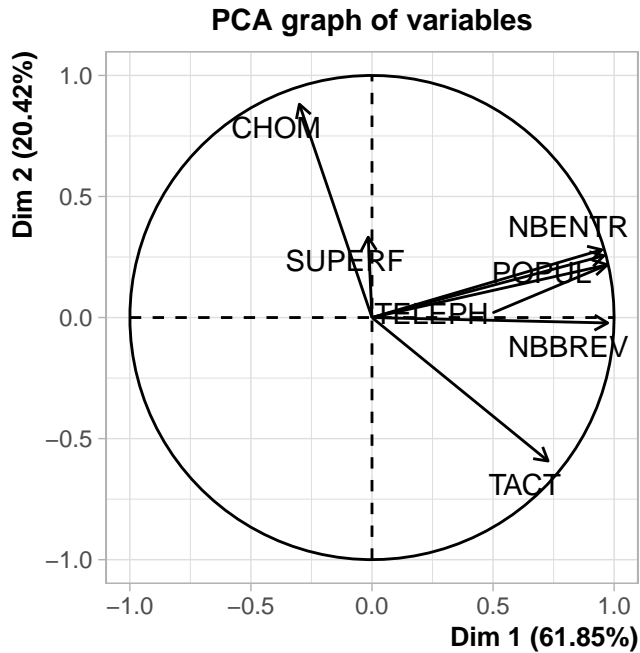
	Dim.1	Dim.2	Dim.3
POPUL	21.19	4.566	0.195
TACT	12.214	24.571	2.202
SUPERF	0.006	7.705	87.651
NBENTR	20.794	5.475	0.647
NBBREV	21.888	0.035	2.451
CHOM	2.077	54.322	6.57
TELEPH	21.832	3.326	0.284

Sur l'axe 1, les variables NBREV, TELEPH, POPUL et NBENTR ont des contributions fortes. La variable CHOM a une contribution importante sur l'axe 2. On constate que la contribution de TACT est plus importante sur l'axe 2 que sur l'axe 1. On interprétera donc cette variable sur l'axe 2 uniquement (bien que cette variable soit aussi corrélée à l'axe 1). Enfin, la variable SUPERF a une forte contribution sur l'axe 3.

- on représente le graphique des variables sur les deux premiers axes de l'ACP :

Figure : ACP-Graphique des variables (dim 1 - dim 2) :

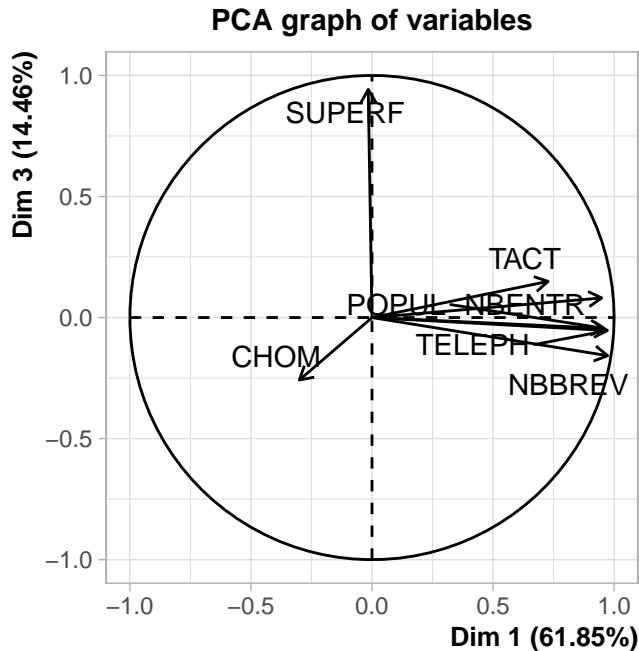
```
plot(res, axes = c(1, 2), choix = "var")
```



- on représente le graphique des variables sur les axes 1 et 3 de l'ACP :

Figure : ACP-Graphique des variables (dim 1 - dim 3) :

```
plot(res, axes = c(1, 3), choix = "var")
```

**Commentaires :**

- on voit que l'axe 1 représente près de 62% de l'inertie totale et le deuxième axe représente 20%. Autrement dit, ces deux variables expliquent à elles seules plus de 80% de l'inertie totale. L'axe 3 représente environ 15% de l'inertie totale.
- L'axe 1 est fortement corrélé positivement aux variables NBENTR, POPUL, TELEPH et NBBREV. Il s'agit de toutes les variables qui représentent à la base des comptages, à l'exception de SUPERF. On peut supposer que cette dernière variable se comporte différemment des autres. Autrement dit, on observera des coordonnées fortes positives sur l'axe 1 pour les régions qui ont les plus grandes valeurs de NBENTR, POPUL, TELEPH et NBBREV et au contraire, on observera des coordonnées fortes négatives pour les régions ayant des plus petites valeurs pour NBENTR, POPUL, TELEPH et NBBREV.
- L'axe 2 est quant à lui fortement corrélé positivement à la variable CHOM et négativement à la variable TACT. Autrement dit, cet axe va opposer des régions avec un taux de chômage élevé et un faible taux d'activité (coordonnées fortes positives sur l'axe 2) avec des régions avec un taux de chômage peu élevé et un fort taux d'activité (coordonnées fortes négatives sur l'axe 2).
- L'axe 3 est corrélé à la variable SUPERF. Autrement dit, cet axe va opposer les régions de grande taille (coordonnées fortes positive sur l'axe 3) avec les régions de petites tailles.

3.3 ACP : Interprétation des individus

- Dans un premier temps, on affiche les coordonnées des individus dans les trois premières composantes principales :

Tableau : ACP-individus-Coordonnées :

```
df <- as.data.frame(res_individus$coord)[, 1:3]
print_table(df)
```

	Dim.1	Dim.2	Dim.3
Alsace	-0.281	-2.735	-0.768
Aquitain	-0.11	0.998	1.196
Auvergne	-0.934	-0.349	0.09
Bas-Norm	-0.797	-0.895	-0.523
Bourgogn	-0.59	-0.785	0.749
Bretagne	-0.126	0.309	0.083
Centre	-0.073	-0.568	1.435
Champ-Ar	-1.027	-0.498	0.051
Fr-Comte	-0.983	-1.56	-0.447
Hte-Norm	-0.861	-0.082	-1.29
Ile-de-F	8.524	-0.366	-1.204
Lang-Rou	-1.381	2.451	-0.64
Limousin	-1.147	-1.428	-0.443
Lorraine	-0.824	0.038	-0.215
Midi-Pyr	-0.175	0.421	1.745
Nord-PdC	-0.518	2.192	-1.878
Pays-Loi	0.083	0.263	0.536
Picardie	-1.212	0.286	-0.726
Poit-Cha	-0.925	0.157	-0.083
Pr-Cte-A	0.851	1.864	0.108
Rh-Alpes	2.506	0.284	2.226

- on vérifie la qualité de représentation des individus. Il existe plusieurs outils :

Tableau : ACP-individus-distance :

```
dist <- res_individus$dist
dist
```

Alsace Aquitain Auvergne Bas-Norm Bourgogn Bretagne Centre Champ-Ar

```

2.9573426 1.5742264 1.0668530 1.4099552 1.2529130 0.3819612 1.5617667 1.2605640
Fr-Comte Hte-Norm Ile-de-F Lang-Rou Limousin Lorraine Midi-Pyr Nord-PdC
1.9504813 1.7673062 8.6221446 2.9088008 1.8920299 1.2294339 1.8086227 2.9832118
Pays-Loi Picardie Poit-Cha Pr-Cte-A Rh-Alpes
0.7143063 1.5046773 1.0242483 2.1254096 3.4019149

```

Tableau : ACP-individus-Cos2 : on voit les individus qu'il faudrait commenter (sur l'axe 1, les valeurs doivent être supérieures à 0.5, sur l'axe 2, les valeurs doivent être supérieures à 0.25 et sur l'axe 3, les valeurs doivent être supérieures à 0.15).

```

df <- as.data.frame(res_individus$cos2)[, 1:3]
print_table(df)

```

	Dim.1	Dim.2	Dim.3
Alsace	0.009	0.855	0.068
Aquitain	0.005	0.402	0.577
Auvergne	0.766	0.107	0.007
Bas-Norm	0.319	0.403	0.138
Bourgogn	0.222	0.392	0.358
Bretagne	0.109	0.656	0.047
Centre	0.002	0.132	0.844
Champ-Ar	0.664	0.156	0.002
Fr-Comte	0.254	0.639	0.053
Hte-Norm	0.238	0.002	0.533
Ile-de-F	0.977	0.002	0.02
Lang-Rou	0.225	0.71	0.048
Limousin	0.367	0.57	0.055
Lorraine	0.449	0.001	0.031
Midi-Pyr	0.009	0.054	0.931
Nord-PdC	0.03	0.54	0.396
Pays-Loi	0.013	0.135	0.563
Picardie	0.649	0.036	0.233
Poit-Cha	0.815	0.023	0.007
Pr-Cte-A	0.16	0.769	0.003
Rh-Alpes	0.543	0.007	0.428

Tableau : ACP-individus-contributions : une forte contribution est une contribution supérieure à $100 \times 1/n$.


```
res_individus$contrib[,1:3]
```

	Dim.1	Dim.2	Dim.3
Alsace	0.086809057	24.918240556	2.77697466
Aquitain	0.013354231	3.320148006	6.72792386
Auvergne	0.959263358	0.405348182	0.03814559
Bas-Norm	0.698403995	2.668985940	1.28614398
Bourgogn	0.382913962	2.051885229	2.64152920
Bretagne	0.017424029	0.318786476	0.03210965
Centre	0.005890698	1.074512863	9.67915870
Champ-Ar	1.160771928	0.825102637	0.01209954
Fr-Comte	1.062778731	8.104999281	0.94013550
Hte-Norm	0.816244227	0.022235584	7.82883659
Ile-de-F	79.909714674	0.446550507	6.82366070
Lang-Rou	2.096178253	20.014217475	1.92825951
Limousin	1.446776818	6.792794994	0.92397560
Lorraine	0.747130490	0.004932781	0.21734802
Midi-Pyr	0.033581859	0.590516971	14.32255754
Nord-PdC	0.294728061	16.007515980	16.58803857
Pays-Loi	0.007533380	0.229981867	1.35172124
Picardie	1.616360083	0.272965763	2.48038661
Poit-Cha	0.940375100	0.082116609	0.03229326
Pr-Cte-A	0.797096286	11.579145499	0.05452279
Rh-Alpes	6.906670782	0.269016800	23.31417890

- Enfin, on va représenter le graphique des individus sur les deux premiers axes de l'ACP, ainsi que sur les axes 1 et 3 :

Figure : ACP-Graphique des individus (dim 1 - dim 2) :

```
plot(res, axes = c(1, 2), choix = "ind")
```

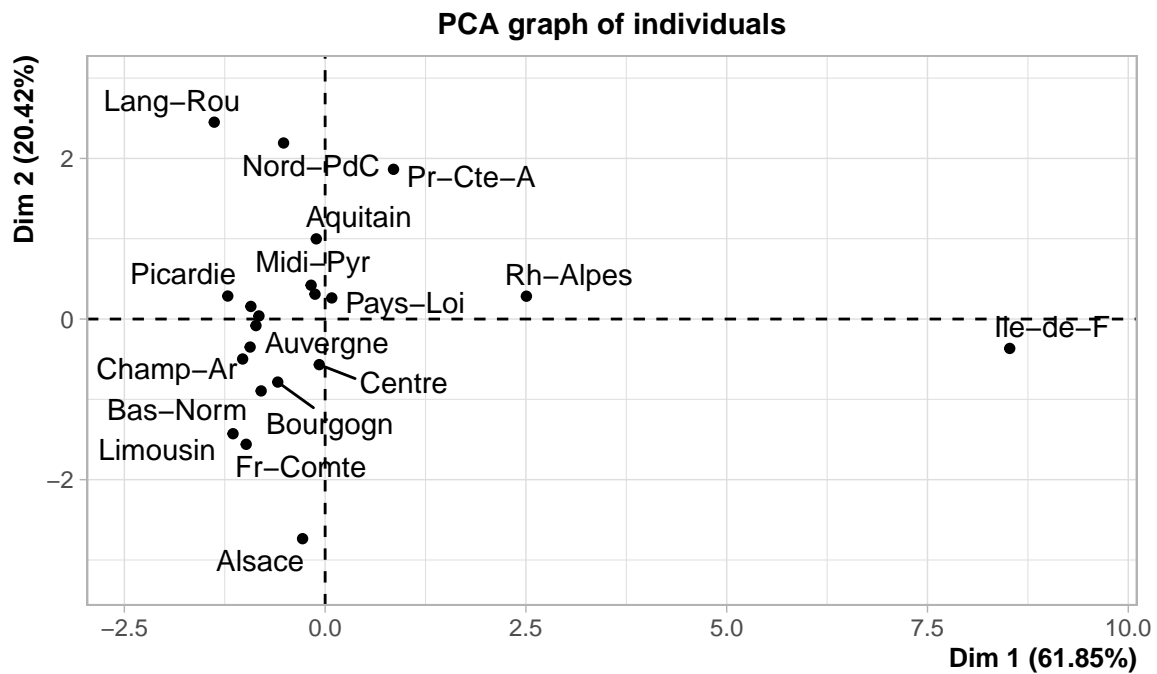
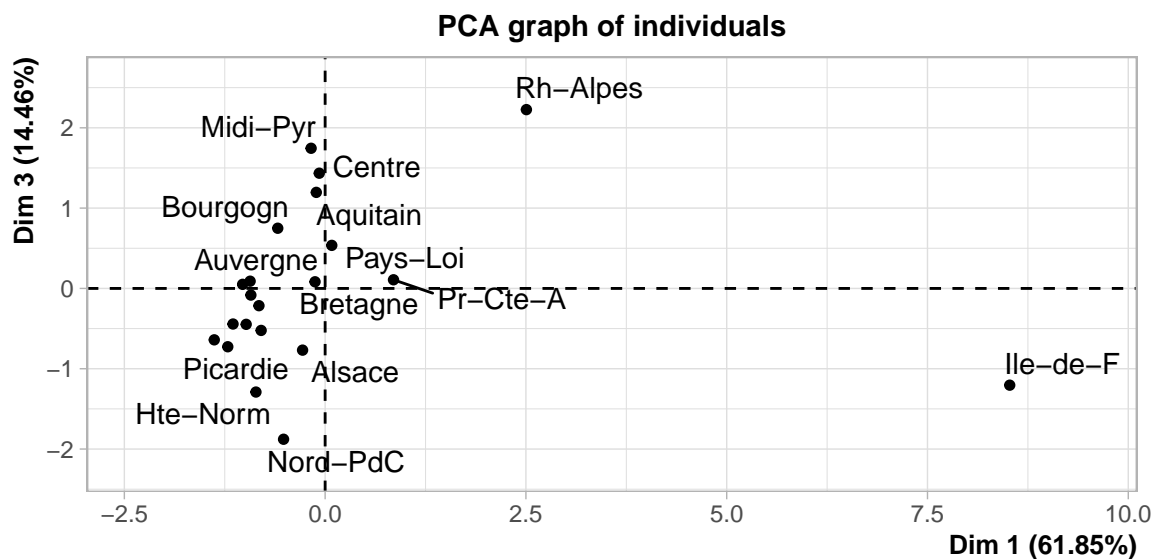


Figure : ACP-Graphique des individus (dim 1 - dim 3) :

```
plot(res, axes = c(1, 3), choix = "ind")
```

**Commentaires :**

- Les régions d'Ile-de-France et Rhône-Alpes sont celles qui ont les coordonnées les plus fortes sur l'axe 1. Autrement dit, il s'agit des régions qui ont les plus fortes valeurs de NBENTR, POPUL, TELEPH et NBBREV. Au contraire, les régions Picardie, Poitou-Charrentes, Auvergne, Champagne-Ardenne ont des valeurs plutôt faibles sur l'axe 2, ce qui en fait des régions avec des valeurs plus faibles de NBENTR, POPUL, TELEPH et NBBREV.
- Sur l'axe 2, on constate que les régions Langudoc-Roussillon, Nord-Pas-de-Calais, Aquitaine et PACA ont des valeurs élevées, ce qui impliquent des valeurs fortes de CHOM

et faible de TACT. Au contraire, l'Alsace, Bourgogne, Franche-Comté, Limousin ont des valeurs élevées négatives, ce qui indique que la variable TACT a une valeur élevée alors que CHOM a une valeur faible.

- La région Provence-Côte d'Azur a des valeurs plutôt fortes sur les deux axes, ce qui semble indiquer que c'est une grosse région en terme de NBENTR, POPUL, TELEPH et NBBREV et avec du chômage.
- Sur l'axe 3, on voit que les régions Rhône-Alpes, Midi-Pyrénées et Centre ont les plus grandes valeurs et correspondent aux régions avec les plus grandes superficies. Au contraire, les régions Ile-de-France et Pas-de-Calais ont des valeurs négatives fortes sur l'axe 3 et correspondent à des petites régions en terme de superficie

3.4 Pour aller plus loin

On a vu que l'Ile-de-France semblait fortement influencer les résultats de l'ACP. Il s'agirait donc de refaire la même analyse, en ajoutant cette observation comme "individu supplémentaire".

```
res <- PCA(regions[, 3:9], ind.sup = 11, graph = FALSE)
```

Commentaires :

En refaisant l'analyse, on devrait se rendre compte que deux axes sont suffisants, l'axe 1 opposant les régions avec les plus grandes valeurs pour NBENTR, POPUL, TELEPH et NBBREV, le deuxième axe opposant les régions avec une forte proportion de travailleurs v.s. un fort taux de chômage.