ÉCOLE NATIONALE D'INGÉNIEURS DE TUNIS

ANNÉE SCOLAIRE : 2021-2022

PROJET DE STATISTIQUES

ÉLABORÉ PAR :
Khalil Harrabi

ENCADRE PAR :
Anissa Rabhi

Table des matières

1
Introduction

1.1 Introduction

L'analyse en composantes principales consiste à transformer des variables liées entre elles en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées "composantes principales", ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Mathématiquement, l'analyse en composantes principales est un simple changement de base : passer d'une représentation dans la des facteurs définis par les vecteurs propres de la matrice des corrélations.

Jeux de données
Les données utilisées ici dans mon projet sont disponibles dans le lien suivant , elle contient les température pour 37 capital de pays. On possède 17 variables qui sont :
-Les 12 mois de l'année.
-5 variable quantitative (la moyenne, l'amplitude, le latitude, le longitude).
-une variable qualitative qui est la région.
Dans un premier lieu on va prendre les 30 premières pays et les 12 mois comme variables et individus actifs. Pour faire ça on implémenté le code suivant :

```
#importation et preparation de tableau de donn e

temperature <- read.table("D:/ACP/temperat.csv",header=TRUE, sep=";",
     dec=".", row.names=1)
head(temperature) # visualisation de la base


#Extraction des individus et variables actifs:

temperature.active <- temperature[1:30, 1:12]
```

2
Statistiques élémentaires

2.1 Statistiques descriptives

On essaie tous d'abord d'utiliser RStudio pour comprendre la répartition des données pour les différents individus et variables.

Figure 2.1 – Distribution des données

```
> summary(temperature.active)
    Janvier          Février            Mars            Avril             Mai              Juin            Juillet     
 Min.   :-9.300   Min.   :-7.600   Min.   :-2.70   Min.   : 2.900   Min.   : 6.50   Min.   : 9.30   Min.   :11.10  
 1st Qu.:-1.625   1st Qu.:-0.275   1st Qu.: 1.45   1st Qu.: 7.175   1st Qu.:12.03   1st Qu.:15.18   1st Qu.:17.20  
 Median : 0.150   Median : 1.850   Median : 5.25   Median : 8.850   Median :13.75   Median :16.75   Median :18.65  
 Mean   : 1.123   Mean   : 1.930   Mean   : 4.90   Mean   : 8.920   Mean   :13.54   Mean   :16.98   Mean   :19.18  
 3rd Qu.: 4.525   3rd Qu.: 4.800   3rd Qu.: 7.15   3rd Qu.:11.125   3rd Qu.:14.82   3rd Qu.:18.18   3rd Qu.:20.30  
 Max.   :10.500   Max.   :11.300   Max.   :12.80   Max.   :15.400   Max.   :20.10   Max.   :24.50   Max.   :27.40  
      Août         Septembre         Octobre          Novembre          Décembre     
 Min.   :10.60   Min.   : 7.90   Min.   : 4.500   Min.   :-1.100   Min.   :-6.000  
 1st Qu.:16.52   1st Qu.:12.85   1st Qu.: 8.625   1st Qu.: 2.900   1st Qu.: 0.225  
 Median :18.15   Median :14.75   Median :10.100   Median : 5.100   Median : 1.550  
 Mean   :18.61   Mean   :15.20   Mean   :10.670   Mean   : 5.750   Mean   : 2.667  
 3rd Qu.:19.88   3rd Qu.:16.70   3rd Qu.:12.250   3rd Qu.: 7.225   3rd Qu.: 5.350  
 Max.   :27.20   Max.   :23.80   Max.   :19.200   Max.   :14.600   Max.   :11.100  
> apply(temperature.active[,1:12],2,FUN=sd)# calcul des ecart-type
 Janvier  Février     Mars    Avril      Mai     Juin  Juillet     Août Septembre  Octobre Novembre Décembre 
5.123151 5.055086 4.428435 3.401663 2.936040 3.062161 3.389273 3.599687 3.800907 4.001823 4.192502 4.592147 
> |
```

On va se servir de ces données pour réaliser le tableau ci-dissous qui représente les données d'une manière plus simple.

Mois	Moyenne	Écart type
Janvier	1.123	5.123
Février	1.93	5.05
Mars	4.9	4.42
Avril	8.89	3.4
May	13.54	2.93
Juin	16.98	3.06
Juillet	19.18	3.38
Août	18.61	3.59
septembre	15.2	3.8
Octobre	10.67	4.0
Novembre	5.75	4.19
Décembre	2.66	4.59

Pour obtenir les données de ce tableau on a utilisé le code suivant :

```
head(temperature.active)
summary(temperature.active)
apply(temperature.active[,1:12],2,FUN=sd)# calcul des ecart-type
```

Ce tableau nous permet de mieux visualiser la répartition des données. On remarque que un changement normal et réalistes des températures qui sont élevées pendant les moins de l'été et faible pendant l'hiver. On constate de plus que les écarts types vaut sont élevés, donc les température des capitales sont largement distribués autour de la moyenne.

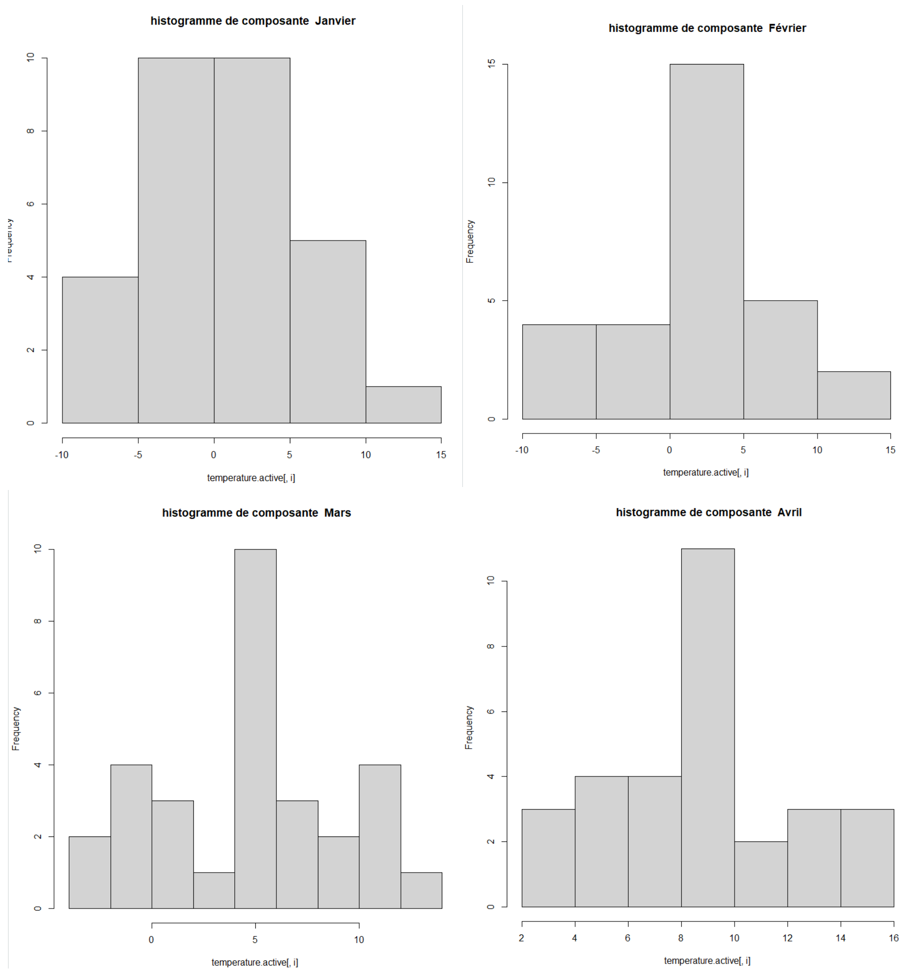
2.2 Description de toutes les variables : Histogramme.

On présente 4 histogrammes des 4 premières variables de la base de données :

```

#histogrammes
for ( i in 1 : 4 )
```

```
{ hist ( temperature.active[,i] , main=paste ( " histogramme de
    composante ",names(
      temperature.active ) [i] ) ) }
```



Les histogrammes montrent que les variables n'ont pas des distributions classiques.

2.3 Résultats de corrélations

Pour visualiser la corrélation entre les variable on a implémenté le code qui nous permet d'obtenir la matrice de corrélation.

Les résultats qu'on a obtenu peuvent s'interpréter de la manière suivante :

C'est la matrice de variance covariance des variables centrées réduites.Elle possède p valeurs propres.

Figure 2.2 – Matrice de corrélation

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Janvier	1.0000000	0.9886928	0.9485078	0.7859004	0.5276209	0.4683026	0.5084391	0.6024576	0.7775550	0.8911703	0.9606176	0.9932455
Février	0.9886928	1.0000000	0.9735251	0.8424159	0.5910489	0.5331149	0.5613625	0.6532398	0.8182289	0.9113277	0.9686383	0.9805892
Mars	0.9485078	0.9735251	1.0000000	0.9276715	0.7269138	0.6527287	0.6729447	0.7587904	0.8935529	0.9576549	0.9726601	0.9520052
Avril	0.7859004	0.8424159	0.9276715	1.0000000	0.9221532	0.8655484	0.8559761	0.9104182	0.9688954	0.9553294	0.9020442	0.8117752
Mai	0.5276209	0.5910489	0.7269138	0.9221532	1.0000000	0.9715942	0.9495522	0.9613025	0.9259680	0.8389577	0.7200720	0.5785093
Juin	0.4683026	0.5331149	0.6527287	0.8655484	0.9715942	1.0000000	0.9868812	0.9780275	0.9113542	0.7950587	0.6713039	0.5198692
Juillet	0.5084391	0.5613625	0.6729447	0.8559761	0.9495522	0.9868812	1.0000000	0.9892559	0.9227032	0.8161291	0.6990450	0.5612406
Août	0.6024576	0.6532398	0.7587904	0.9104182	0.9613025	0.9780275	0.9892559	1.0000000	0.9654463	0.8813608	0.7778657	0.6490140
Septembre	0.7775550	0.8182289	0.8935529	0.9688954	0.9259680	0.9113542	0.9227032	0.9654463	1.0000000	0.9709000	0.9079598	0.8120702
Octobre	0.8911703	0.9113277	0.9576549	0.9553294	0.8389577	0.7950587	0.8161291	0.8813608	0.9709000	1.0000000	0.9765969	0.9201927
Novembre	0.9606176	0.9686383	0.9726601	0.9020442	0.7200720	0.6713039	0.6990450	0.7778657	0.9079598	0.9765969	1.0000000	0.9780147
Décembre	0.9932455	0.9805892	0.9520052	0.8117752	0.5785093	0.5198692	0.5612406	0.6490140	0.8120702	0.9201927	0.9780147	1.0000000

Le coefficient de corrélation nous donne deux informations que l'on doit interpréter :
- **Le sens de la relation entre les variables :** si le coefficient est négatif, plus la valeur de la première variable est élevé, plus la valeur de la deuxième diminue.
- **La force de la relation :** En examinant la valeur de chaque coefficient, nous pouvons dire que l'effet de la relation entre deux variables est de grande taille et que l'association est très forte, ou bien le contraire.

On remarque que le coefficient de corrélation entre toutes les variables est positive, donc ils sont positivement corrélées, elle varient dans le même sens, de plus le coefficient est toujours supérieure à 0.5 donc elles sont fortement corrélées. Un cas un peu particulier est la corrélation entre Juin et Janvier dont le coefficient est de 0.4 donc la corrélation est moins bonne.
On peut donc sortir par un conclusion : Lorsque la température augmente (ou bien diminue) pendant un mois de l'année elle augmente(ou bien diminue) pendant tous les autres mois et cette variation est proportionnel au coefficients de corrélation.

2.4 Contrôle de la linéarité des relations entre variables

Dans cette partie on présente le graphiques des relations entre les différentes variables de notre problème. On voit donc les relations entre les différentes variables ce qui nous facilitera l'étude :

```
#relation entre donn es
pairs(temperature.active)
```

On obtient le tracé ci-dessous : Ce tracé confirme bien les interprétations concernant les corrélations entre les variables.

FIGURE 2.3 – Valeur propres

On peut constater que les variables présentent des relations très compliquées à modéliser ce qui justifie q'on doit utiliser les techniques d'ACP afin de pouvoir trouver les relations inter-variables.

3

Statistiques élémentaires

3.1

On a implémenté les code suivant qui nous a permis d'effectuer l'analyse de composante principale a notre base de données.

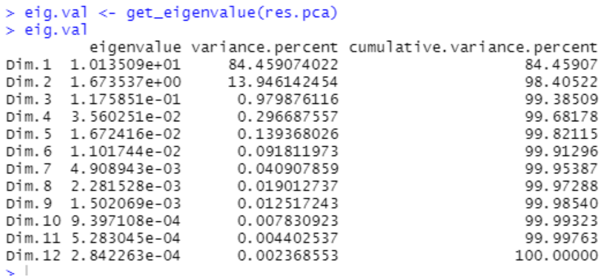
```
res.pca <- PCA(temperature.active, scale.unit = TRUE, graph = F)
```

3.2 Table des Valeurs propres et vecteurs propres

Les valeurs propres permettent d'effectuer un choix du nombre de composantes principales à retenir pour l'interprétation.
Le choix du nombre d'axes à interpréter se fait sur la base de règles.

- **La règle de Kaiser :** Elle consiste à retenir les axes pour lequels les valeurs propres sont supérieures à 1 (1 étant la moyenne de l'ensemble des valeurs propres). Il est à noter qu'on peut aussi avoir des résultats d'ACP dont la somme des valeurs propres n'est pas égale à p (nombre de variable) (cas de l'ACP non réduite). Dans ce cas, il faut adapter cette règle de Kaiser et retenir les valeurs propres supérieures à la moyenne des valeurs propres, et non plus à 1.

```
res.pca <- PCA(temperature.active, scale.unit = TRUE, graph = F)
```



```
> eig.val <- get_eigenvalue(res.pca)
> eig.val
        eigenvalue variance.percent cumulative.variance.percent
Dim.1  1.013509e+01     84.459074022                    84.45907
Dim.2  1.673537e+00     13.946142454                    98.40522
Dim.3  1.175851e-01      0.979876116                    99.38509
Dim.4  3.560251e-02      0.296687557                    99.68178
Dim.5  1.672416e-02      0.139368026                    99.82115
Dim.6  1.101744e-02      0.091811973                    99.91296
Dim.7  4.908943e-03      0.040907859                    99.95387
Dim.8  2.281528e-03      0.019012737                    99.97288
Dim.9  1.502069e-03      0.012517243                    99.98540
Dim.10 9.397108e-04      0.007830923                    99.99323
Dim.11 5.283045e-04      0.004402537                    99.99763
Dim.12 2.842263e-04      0.002368553                   100.00000
>
```

FIGURE 3.1 – Valeur propres

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

Visualisation des valeurs propres :

```
eig.val <- get_eigenvalue(res.pca)
eig.val
print(eig.val[ , 1 ] )
plot ( 1 : 12 , eig.val[ ,1] , xlb=" num ro des axes " , ylab="
    valeur
propre " , type="b" )
```

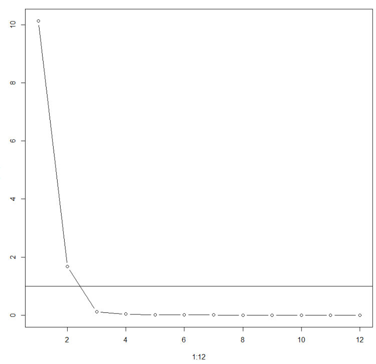


FIGURE 3.2 – Le graphique des valeurs propres

- **La règle de l'éboulis combinée avec celle de Kaiser** est une des meilleurs. En effet, on commence par regarder combien de valeurs propres sont supérieures à

la moyenne. Puis on regarde si la dernière valeur propre retenue (supérieure à la moyenne) est suffisamment éloignée de celle qui la suit (inférieure à la moyenne).Si oui, on reste sur la décision de la règle de Kaiser, si non, on coupera au saut plus important le plus près.
On peut présenter aussi d'une autre manière les valeurs propres et ceci avec leurs pourcentages.

```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))
```

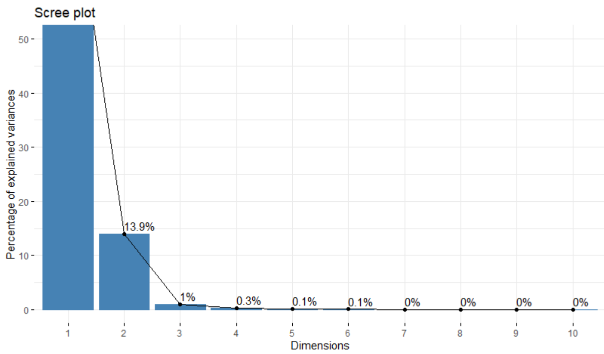


Figure 3.3 – Le graphique des valeurs propres

Dans notre exemple,le nombre de variables est 12, c'est bien la somme des valeurs propres. On retiendra donc 2 axes pour l'interprétation. le premier et le deuxième axe comportent respectivement 54 et 13.9 de l'inertie totale du nuage, et le plan (1,2) totalise 67.9 % de cette variance totale.

4

Étude des individus : Résultats sous R

4.1 Coordonnées des individus, contribution et qualité de la représentation d'un individu

Nous stockons le résultat dans une variable, ainsi nous pourrons avoir les coordonnées des individus mais ainsi la qualité de contribution sur chacun des axes et ensuite tracer le graphe des individus.

Résultats des individus :

```
> res.ind=res.pca$ind
> res.ind$coord
                Dim.1       Dim.2       Dim.3        Dim.4       Dim.5
Amsterdam  -0.27037180 -1.23130886 -0.02511286 -0.355423533 -0.17537417
Athènes     7.13751025  1.10111666  0.51851730 -0.209593685  0.24789727
Berlin     -0.78059176  0.16483673 -0.24365038 -0.144603860 -0.18654455
Bruxelles   0.13819360 -1.03799959 -0.11322011 -0.047518038 -0.14647001
Budapest    1.19066400  1.87518068 -0.52677595  0.072644043  0.08020867
Copenhague -1.96271269 -0.34603086  0.50770511 -0.135617021  0.08937668
Dublin     -1.00273320 -2.54607701 -0.14066517  0.031747700  0.16727621
Helsinki   -4.53954452  0.61268442  0.67192361 -0.201498840  0.07460563
Kiev       -2.20346337  2.17051626 -0.13510311 -0.133286196  0.05496064
Cracovie   -1.75269351  1.02712432 -0.24153971  0.001993114 -0.02145905
Lisbonne    5.12319158 -1.40970070 -0.27967104 -0.159832046  0.10166498
Londres    -0.43578779 -1.43701147 -0.05988663  0.058199989  0.14190121
Madrid      3.58798241  0.84726097  0.39609529  0.654520407 -0.20481243
Minsk      -3.73590396  1.54900957 -0.01230652 -0.188900855  0.09170347
Moscou     -3.95765094  2.34240635 -0.26618053 -0.034491785  0.04014797
Oslo       -3.80722206  0.45583849  0.33414843  0.217620444  0.03649739
Paris       0.92869133 -0.75475624 -0.06934294 -0.147275043 -0.18880972
Prague     -0.59673838  0.85142146 -0.23642741  0.080371362  0.07270224
Reykjavik  -5.21777205 -2.83941013  0.02291357  0.178616993  0.02690947
Rome        4.91025757  0.45228351  0.15463540  0.010274667  0.10977190
Sarajevo   -0.32915763  0.46717264 -0.33167688 -0.008422393 -0.20320869
Sofia      -0.07311342  0.94596830 -0.21205484 -0.020655010 -0.18816251
Stockholm  -3.65551977  0.14907564  0.94503069  0.110372800 -0.07825936
Anvers      0.09581888 -1.02112716 -0.07419215 -0.214583078 -0.11432083
Barcelone   5.51565252 -0.23085551  0.22215878  0.035021362  0.02983277
Bordeaux    2.36837258 -0.58731739 -0.28604841  0.161945049  0.05399257
Edimbourg  -1.78939439 -2.21757598 -0.13943205  0.112395168  0.09072664
Francfort  -0.09593470  0.30317224 -0.45448536  0.142707770  0.14083078
Genève     -0.16239800  0.31740487 -0.26494995  0.242454894  0.04279454
Gènes       5.37236921  0.02669779  0.33959383 -0.109184377 -0.18637973
```

On remarque que Athéne, Rome, Barcelone et Génes sont les plus représentatifs

positivement avec le première axe et Stockholm Reykjavik Helsinki sont les plus représentatifs négativement.
On remarque que Moscou, Minsk, Kiev sont les plus représentatifs positivement avec le deuxième axe et que Lisbonne, Londres, Edimbourg sont les plus représentatifs négativement.

Contribution :

```
> res.pca$ind$contrib
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Amsterdam	0.024042187	3.019794665	0.017877990	1.182743e+01	6.13007242
Athènes	16.755009327	2.414960692	7.621717267	4.112961e+00	12.24835686
Berlin	0.200400642	0.054119200	1.682908514	1.957753e+00	6.93584647
Bruxelles	0.006280975	2.146039779	0.363390418	2.114050e-01	4.27594149
Budapest	0.466261575	7.003733966	7.866439563	4.940810e-01	1.28226277
Copenhague	1.266965078	0.238491590	7.307172982	1.721974e+00	1.59214563
Dublin	0.330690693	12.911790570	0.560918164	9.436756e-02	5.57702663
Helsinki	6.777597019	0.747682260	12.798707942	3.801397e+00	1.10937295
Kiev	1.596845332	9.383600098	0.517436457	1.663292e+00	0.60205746
Cracovie	1.010329750	2.101306616	1.653877799	3.719309e-04	0.09178155
Lisbonne	8.632416324	3.958195417	2.217283942	2.391806e+00	2.06004680
Londres	0.062459903	4.113048896	0.101668437	3.171348e-01	4.01334563
Madrid	4.234009168	1.429808270	4.447599624	4.010925e+01	8.36078299
Minsk	4.590316055	4.779160395	0.004293357	3.340920e+00	1.67612297
Moscou	5.151410483	10.928684222	2.008532694	1.113857e-01	0.32126359
Oslo	4.767246408	0.413871368	3.165229382	4.434019e+00	0.26549597
Paris	0.283657302	1.134636702	0.136310989	2.030750e+00	7.10531021
Prague	0.117116782	1.443884953	1.584608548	6.047847e-01	1.05348889
Reykjavik	8.954088617	16.058303333	0.014883733	2.987058e+00	0.14432601
Rome	7.929754315	0.407441179	0.677866534	9.884021e-03	2.40168849
Sarajevo	0.035633546	0.434708563	3.118578840	6.641545e-03	8.23036106
Sofia	0.001758107	1.782364676	1.274743181	3.994376e-02	7.05668195
Stockholm	4.394904646	0.044264621	25.317344740	1.140571e+00	1.22069431
Anvers	0.003019627	2.076840095	0.156042258	4.311110e+00	2.60486431
Barcelone	10.005642468	0.106150947	1.399114630	1.148323e-01	0.17738689
Bordeaux	1.844808257	0.687051223	2.319558932	2.455464e+00	0.58103512
Edimbourg	1.053084760	9.794909357	0.551126858	1.182751e+00	1.64060481
Francfort	0.003026932	0.183072141	5.855528778	1.906748e+00	3.95302487
Genève	0.008673863	0.200664516	1.990004224	5.503767e+00	0.36501531
Gènes	9.492549859	0.001419692	3.269233224	1.116141e+00	6.92359560

```
>
```

Pour l'axe 1 : Athènes, Barcelone, Gènes participent le plus à la création de l'axe du côté positif. En effet les variables contribuent toutes dans le même sens à la formation de l'axe.
Pour l'axe 2 : Dublin, Moscou, Reykjavik participent le plus à la création de l'axe du coté positif. les variables sont toutes du même côté de l'axe. **Qualité de représentation des individus :**

```
> res.pca$ind$cos2
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Amsterdam	0.041718429	8.652452e-01	3.599130e-04	7.209373e-02	1.755240e-02
Athènes	0.969682882	2.307825e-02	5.117558e-03	8.361662e-04	1.169714e-03
Berlin	0.761730141	3.396731e-02	7.421427e-02	2.614045e-02	4.350291e-02
Bruxelles	0.016628008	9.381207e-01	1.116122e-02	1.965990e-03	1.867935e-02
Budapest	0.271039859	6.722657e-01	5.305259e-02	1.008915e-03	1.229978e-03
Copenhague	0.895343466	2.782953e-02	5.990999e-02	4.274690e-03	1.856627e-03
Dublin	0.133329495	8.596042e-01	2.623787e-03	1.336534e-04	3.710426e-03
Helsinki	0.954163662	1.738089e-02	2.090442e-02	1.879938e-03	2.577161e-04
Kiev	0.504395333	4.894242e-01	1.896226e-03	1.845567e-03	3.138073e-04
Cracovie	0.728239920	2.500970e-01	1.383056e-02	9.417304e-07	1.091650e-04
Lisbonne	0.925497396	7.007254e-02	2.757967e-03	9.007862e-04	3.644490e-04
Londres	0.083115597	9.037595e-01	1.569612e-03	1.482444e-03	8.812611e-03
Madrid	0.904248534	5.042218e-02	1.102013e-02	3.009079e-02	2.946455e-03
Minsk	0.850620225	1.462354e-01	9.230281e-06	2.174764e-03	5.125259e-04
Moscou	0.737394468	2.583149e-01	3.335627e-03	5.600880e-05	7.588432e-05
Oslo	0.974178827	1.396511e-02	7.504144e-03	3.182890e-03	8.952531e-05
Paris	0.571387202	3.773994e-01	3.185608e-03	1.436965e-02	2.361764e-02
Prague	0.305275719	6.214605e-01	4.792035e-02	5.537669e-03	4.531270e-03
Reykjavik	0.770466747	2.281602e-01	1.485830e-05	9.028784e-04	2.049245e-05
Rome	0.989971699	8.399158e-03	9.818205e-04	4.334607e-06	4.947623e-04
Sarajevo	0.204714138	4.123773e-01	2.078597e-01	1.340328e-04	7.802334e-02
Sofia	0.005201916	8.708077e-01	4.375881e-02	4.151642e-04	3.445366e-02
Stockholm	0.932572125	1.550949e-03	6.232697e-02	8.501749e-04	4.274219e-04
Anvers	0.008116390	9.217654e-01	4.866051e-03	4.070537e-02	1.155344e-02
Barcelone	0.996184796	1.745125e-03	1.616118e-03	4.016171e-05	2.914293e-05
Bordeaux	0.922336882	5.671987e-02	1.345453e-02	4.312458e-03	4.793553e-04
Edimbourg	0.392099696	6.022009e-01	2.380727e-03	1.546962e-03	1.007985e-03
Francfort	0.025488114	2.545453e-01	5.720395e-01	5.640031e-02	5.492644e-02
Genève	0.100657396	3.845134e-01	2.679242e-01	2.243604e-01	6.989740e-03
Gènes	0.993768080	2.454166e-05	3.970753e-03	4.104633e-04	1.196054e-03

Pour la qualité, il faut s'assurer que le cosinus carré supérieure à 0.5.

4.2 Plan des individus :

Ce graphe permet de resumé tous les tableux précedents :

```
fviz_pca_ind(res.pca,
             col.ind = "cos2", # Colorer par le cos2
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
)
```

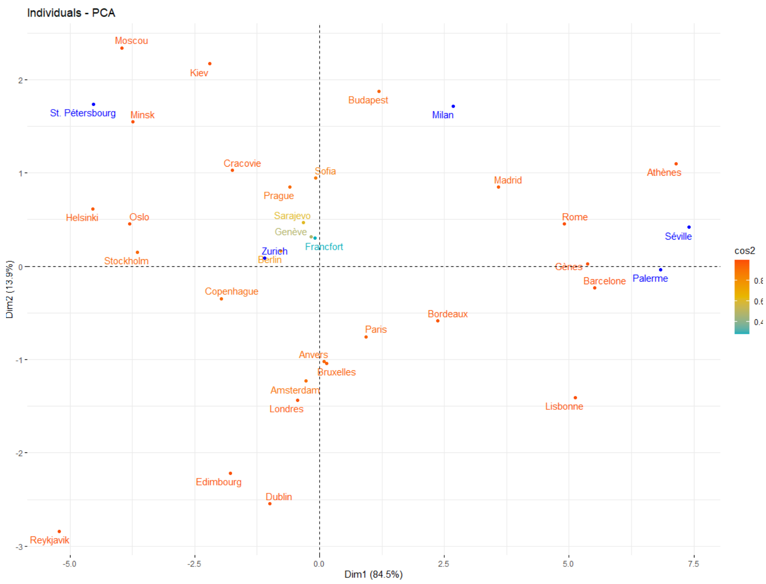


FIGURE 4.1 – Le graphique des individus

L'analyse de ce graphe se fait de la manière suivante :
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 correlés avec cet axe et sont les points explicatifs pour l'axe : Ce sont les points les plus "parlants"; leur "vraie distance" de l'origine est bien représentée sur le plan factoriel.

Dans le mapping ci-dessus, on voit clairement que Génes est extrêmement corrélé avec l'axe horizontal. De même, Dublin et Edimbourg notamment sont très bien correlés à l'axe vertical.

La corrélation de chaque point sur un axe exprime la qualité de représentation du point sur l'axe. Elle prend des valeurs entre 0 (pas corrélé du tout) et 1 (fortement corrélé). Si cette valeur est proche de 1, alors le point est bien représenté sur l'axe.

Les points situés près du centre sont donc généralement mal représentés par le plan factoriel. Leur interprétation ne peut donc pas être effectuée avec confiance.

On s'intéresse donc essentiellement aux points bien représentés (i.e. situés loin du centre). Si deux points sont proche l'un de l'autre, il est probable que les réponses des individus qu'ils représentent soient très similaires. Il faut cependant se méfier : il se peut que sur un axe ils soient 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.

5

Études des variables : Résultats sous R

5.1 Détermination des variables expliquant le mieux un axe donnée

Lorsque l'on a beaucoup de variables, une description automatique des axes par les variables est possible à l'aide de cette commande pour le plan (1,2)

```
> dimdesc(res.pca, axes=c(1,2))
$Dim.1
$quanti
           correlation      p.value
Octobre      0.9916022 2.018455e-26
Septembre    0.9870659 8.299040e-24
Avril        0.9761329 4.119872e-20
Novembre     0.9576545 1.126113e-16
Mars         0.9490605 1.418833e-15
Août         0.9181556 8.935781e-13
Février      0.8935001 3.050753e-11
Décembre     0.8874718 6.345882e-11
Mai          0.8807790 1.365181e-10
Juillet      0.8627772 8.708853e-10
Janvier      0.8600572 1.126340e-09
Juin         0.8473776 3.493998e-09

attr(,"class")
[1] "condes" "list"

$Dim.2
$quanti
          correlation     p.value
Juin        0.5237662 0.002973073
Juillet     0.4813665 0.007078694
Mai         0.4410026 0.014714031
Août        0.3816750 0.037413859
Février    -0.4362516 0.015950215
Décembre   -0.4466105 0.013358569
Janvier    -0.5037736 0.004536847
```

La détermination des variables expliquant chacun des axes est réalisée en examinant leurs coordonnées (table des valeurs propres) qui sont elle-même reliées à leur contribution. -Les variables les plus corrélées à la première dimension sont dans l'ordre : **Octobre, Septembre, Avril, Novembre**. -Les variables les plus corrélées à la deuxième dimension sont dans l'ordre : **Juin, Juillet, Mai**

5.1.1 Plan des variables :

C'est une représentation où, pour deux composantes principales, par exemple c1 et c2, on représente chaque variable z j par un point d'abscisse cor(z j , c1) et d'ordonnée cor(z j , c2).

Les deux premières dimensions contiennent 50% de l'inertie totale (l'inertie est la variance totale du tableau de données, i.e. la trace de la matrice de corrélation).

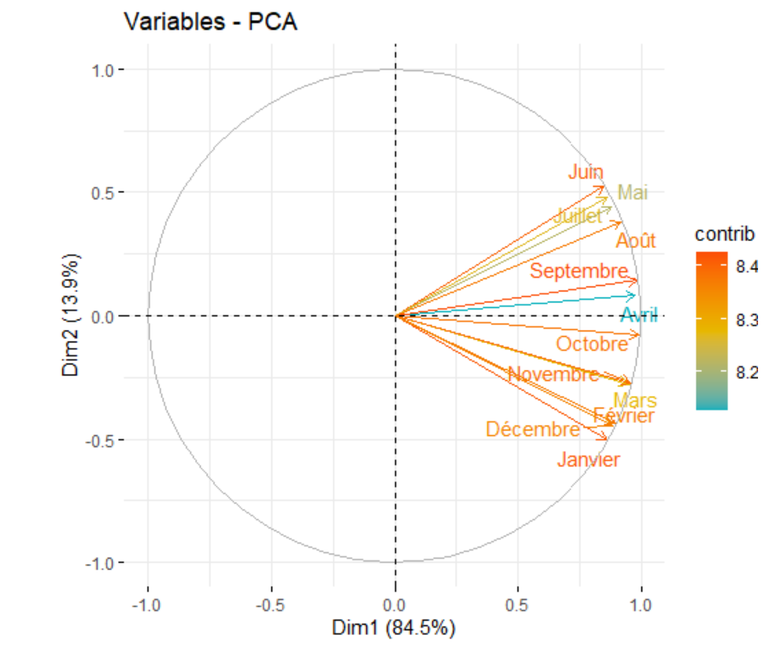


FIGURE 5.1 – Le graphique des variables

On constate que la totalité de nos variables sont corrélées positivement avec la première dimension qui représente 84,58% des données.
Avril, **Octobre** et **Septembre** sont les plus corrélés avec cet axe. **Juin**, **Décembre** et **Janvier** ont une assez bonne corrélation avec le deuxième axe qui représente 13,9% des données ?

6

Conclusion ACP

6.1 Plan Principal ;synthèse

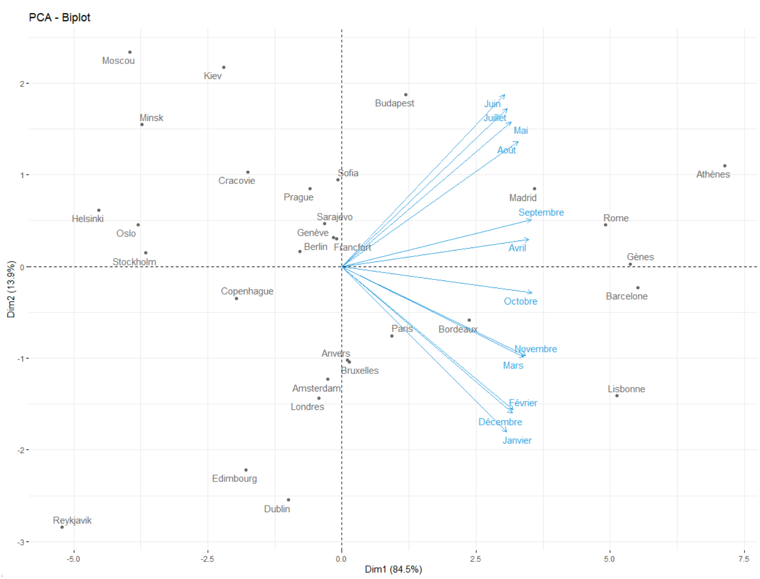


FIGURE 6.1 – Le graphe Biplot

L'interprétation des nouvelles variables (des axes factoriel) se fera à l'aide des individus et variables contribuant le plus à l'axe avec la règle suivante :

Si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés par une valeur élevée de la variable.

On donne un sens à un axe à partir des coordonnées des variables et des individus. Les résultats obtenus dans les chapitres précédents montrent que les capitales qui sont proches géographiquement sont corrélée ou contribuent au même axe. Dans la partie qui suit on va vérifier ceci.

6.2 Classification ascendante hiérarchique

On va tracer le Dendrogramme.

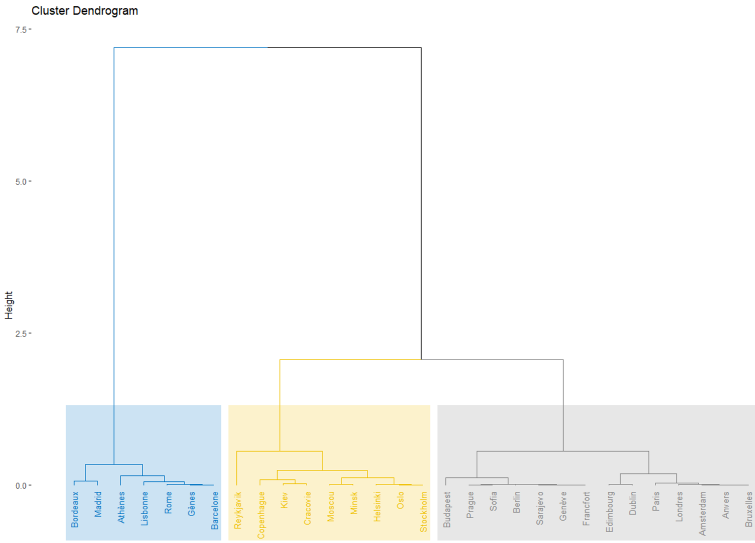


FIGURE 6.2 – Dendrogramme

Le dendrogramme suggère une solution à 3 groupes. On peut constater que les groupes représentent les capitaux des villes de l'ouest, du centre et du l'east de l'Europe ce qui est très pertinent.

Les classes sur les plans factoriels

FIGURE 6.3 – Le graphe des clusters

Les variables qui décrivent le plus les classes

```
> res$desc.var$quanti
$`1`
             v.test Mean in category Overall mean sd in category Overall sd      p.value
Juillet   -2.134593        16.785714    19.180000       2.4549824   3.332307 3.279425e-02
Juin      -2.221652        14.728571    16.980000       2.5172061   3.010692 2.630687e-02
Août      -2.625396        15.485714    18.613333       2.2585890   3.539184 8.654823e-03
Mai       -2.772366        10.842857    13.536667       2.4289075   2.886692 5.565041e-03
Septembre -3.350290        10.985714    15.200000       1.6685568   3.737022 8.072694e-04
Décembre  -3.672301        -2.914286     2.666667       1.8302448   4.514963 2.403765e-04
Janvier   -3.695831        -5.142857     1.123333       2.6250753   5.037042 2.191684e-04
Octobre   -3.709557         5.757143    10.670000       0.9194053   3.934561 2.076225e-04
Novembre  -3.711759         0.600000     5.750000       0.9396048   4.122034 2.058242e-04
Avril     -3.773961         4.671429     8.920000       1.5489298   3.344488 1.606760e-04
Février   -3.903286        -4.600000     1.930000       2.3366643   4.970121 9.489556e-05
Mars      -4.123232        -1.142857     4.900000       1.1043513   4.354002 3.735926e-05

$`2`
NULL

$`3`
            v.test Mean in category Overall mean sd in category Overall sd      p.value
Septembre 4.467196        21.400000    15.200000       1.2793227   3.737022 7.925139e-06
Août      4.377089        24.366667    18.613333       1.5445244   3.539184 1.202746e-05
Octobre   4.366098        17.050000    10.670000       1.6194135   3.934561 1.264859e-05
Novembre  4.300345        12.333333     5.750000       1.8803073   4.122034 1.705328e-05
Juillet   4.258285        24.450000    19.180000       1.7095321   3.332307 2.060016e-05
Avril     4.049555        13.950000     8.920000       0.9639329   3.344488 5.131503e-05
Juin      3.982794        21.433333    16.980000       1.5412837   3.010692 6.810972e-05
Décembre  3.955912         9.300000     2.666667       1.9714631   4.514963 7.624314e-05
Mars      3.937245        11.266667     4.900000       1.0734161   4.354002 8.242228e-05
Février   3.902434         9.133333     1.930000       1.5195028   4.970121 9.523017e-05
Janvier   3.809605         8.250000     1.123333       1.7604450   5.037042 1.391892e-04
Mai       3.774557        17.583333    13.536667       1.2720281   2.886692 1.602920e-04
```

De le résultat ci-dessus, on constate que :

les variables Juillet, Juin et Août sont les plus significativement associées au cluster 1. Par exemple, la valeur moyenne de la variable Août dans le cluster 1 est de 15, ce qui est inférieure à la moyenne globale (18) dans tous les clusters. Par conséquent, on peut conclure que le cluster 1 se caractérise par un faible taux de la variable Août par rapport à tous les autres clusters.

-Aucune des variables n'est pas significatif associées au cluster 2.

. . .etc . . .

Les composantes qui sont le plus associées aux classes

```
> res$desc.axes$quanti
$`1`
         v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.1 -3.615057        -3.873868 9.251859e-16       0.8550377   3.183565 0.0003002816

$`2`
         v.test Mean in category Overall mean sd in category Overall sd     p.value
Dim.3 -3.075966       -0.1712797  1.70905e-15       0.2156072  0.3429069 0.002098215

$`3`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.1 4.461039         5.274494 9.251859e-16        1.043907   3.183565 8.156325e-06
```

Les résultats ci-dessus indiquent que les individus dans les groupes 1 et 3 ont des coordonnées élevées sur l'axes 1. Les individus du groupe 2 ont des coordonnées élevées sur le deuxième axe. Les individus appartenant au troisième groupe ont des coordonnées élevées sur les axes 1, 2 et 3. **Les individus qui représente le plus les classes**

```
> res$desc.ind$para
Cluster: 1
     Oslo  Helsinki Stockholm     Minsk    Moscou
0.3150942 0.8273069 0.9117616 0.9732407 1.7789363
---------------------------------------------------
Cluster: 2
   Genève Francfort    Berlin  Sarajevo    Anvers
0.6912069 0.7289928 0.7337622 0.8143666 0.8370759
---------------------------------------------------
Cluster: 3
    Gènes Barcelone      Rome  Lisbonne    Madrid
0.3098218 0.4351883 0.5002264 1.6425446 1.9536518
```

Pour chaque groupe, les 5 meilleurs individus les plus proches du centre du cluster sont affichés. Ces individus sont appelés paragones. La distance entre chaque individu et le centre du groupe est fournie. Par exemple, les individus représentatifs pour le groupe 1 inclus : Oslo, Hilsinki,Stokholm, Minsk, Moscou.

7
Ajout des individus et variables supplémentaires

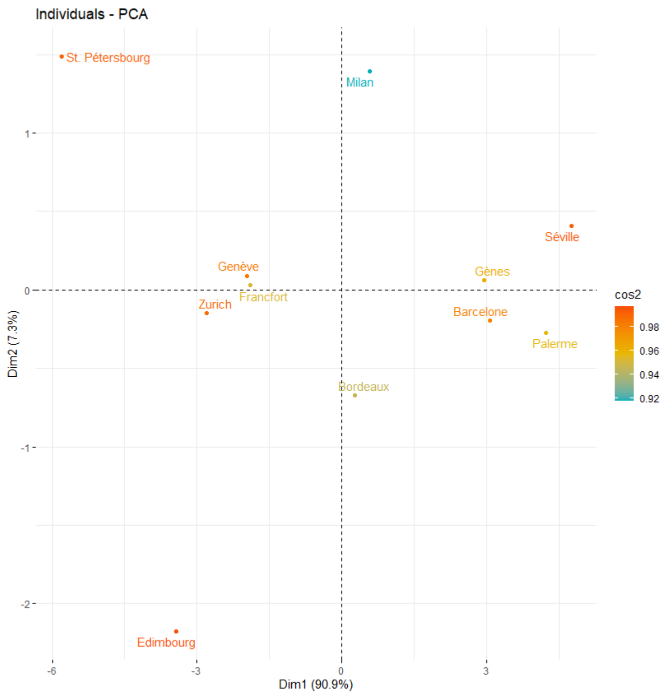
7.1 Implémentations

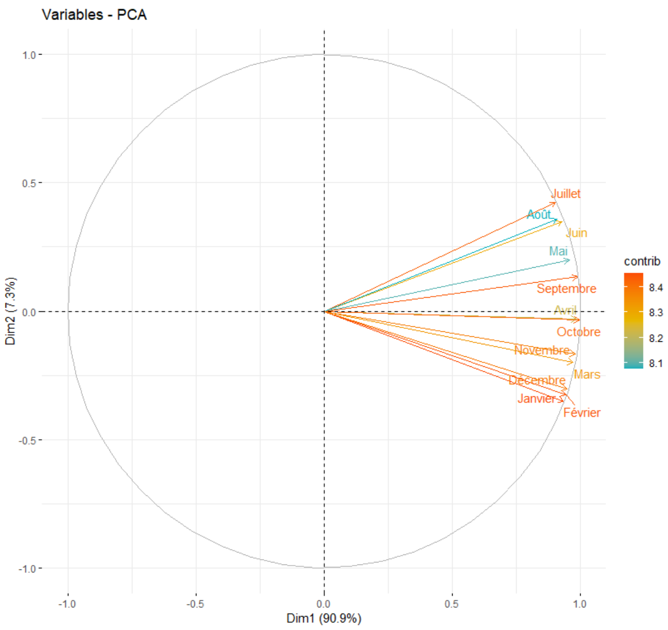
Maintenant on va tester la fiabilité de notre acp en changeant en utilisant d'autre individues. Le code pour cette question est le suivant :

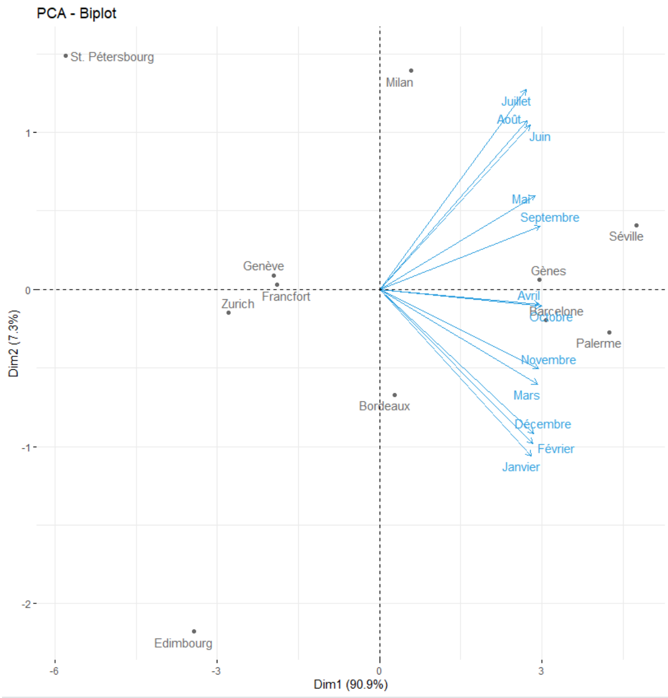
```
#Donn es suplemantaires;
temperature.passive<-temperature[25:35,1:12]
res.pca1 <- PCA(temperature.passive, graph=F)

#graphique des individus
fviz_pca_ind(res.pca1,
             col.ind = "cos2", # Colorer par le cos2
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
)
#graphique des variables
fviz_pca_var(res.pca1,
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
   repel = TRUE      )
#Biplot des individus et des variables (principal)
fviz_pca_biplot(res.pca1, repel = TRUE,
                col.var = "#2E9FDF",  col.ind = "#696969"  )
```

On obtient les nouveaux graphes des individus et des variables suivants :







7.2 Interprétations

Dans un premier lieu on constate que les pourcentages des données portées sur chaque ont changées. La pourcentage du premier axe a augmenté et devenu 90.9% alors que celle de deuxième axe a diminué et devenu 7.3%.
Le graphes des variables des données supplémentaires montre que les corrélations entre les variables n'a pas globalement changer.
En comparant les résultats obtenu pour les premiers individus on remarques que les individus supplémentaires on presque le même comportement avec les axes du plan principale.

8

Régression linéaire :

8.1 Introduction :

Il est impossible de faire une régression linéaire multiple sur le jeu de données qu'on a utilisé pour l'ACP. Donc on utilise le jeu de données suivant :

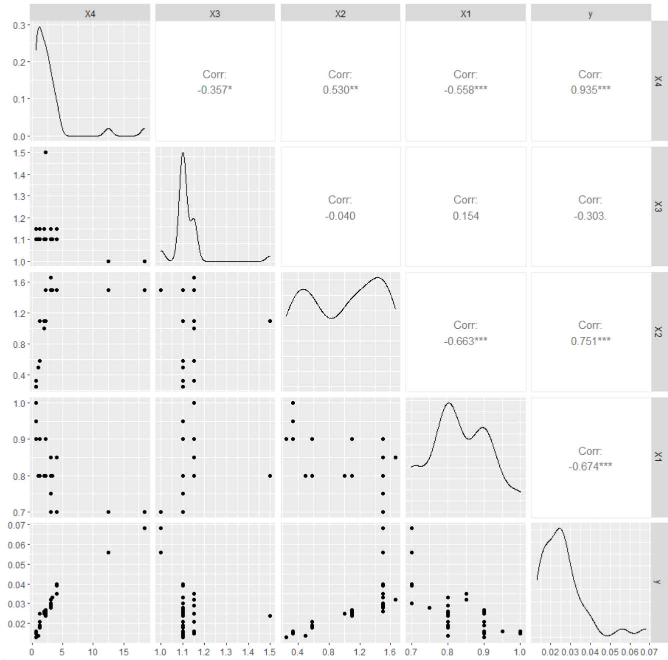
```
data <- read.csv2("D:/ACP/trempe.csv")
head(data)
tail(data)
str(data)
summary(data)
```

```
> head(data)
    X4   X3   X2   X1     y
1 0.58 1.10 0.25 0.90 0.013
2 0.66 1.10 0.33 0.90 0.016
3 0.66 1.10 0.33 0.90 0.015
4 0.66 1.10 0.33 0.95 0.016
5 0.66 1.15 0.33 1.00 0.015
6 0.66 1.15 0.33 1.00 0.016
> tail(data)
      X4   X3  X2   X1     y
27  3.33 1.10 1.5 0.80 0.033
28  4.00 1.10 1.5 0.70 0.039
29  4.00 1.10 1.5 0.70 0.040
30  4.00 1.15 1.5 0.85 0.035
31 12.50 1.00 1.5 0.70 0.056
32 18.50 1.00 1.5 0.70 0.068
```

```
> str(data)
'data.frame':   32 obs. of  5 variables:
 $ X4: num  0.58 0.66 0.66 0.66 0.66 0.66 1 1.17 1.17 1.17 ...
 $ X3: num  1.1 1.1 1.1 1.1 1.15 1.15 1.1 1.1 1.1 1.1 ...
 $ X2: num  0.25 0.33 0.33 0.33 0.33 0.33 0.5 0.58 0.58 0.58 ...
 $ X1: num  0.9 0.9 0.9 0.95 1 1 0.8 0.8 0.8 0.8 ...
 $ y : num  0.013 0.016 0.015 0.016 0.015 0.016 0.014 0.021 0.018 0.019 ...
> summary(data)
       X4               X3               X2               X1                y
 Min.   : 0.58   Min.   :1.000   Min.   :0.2500   Min.   :0.7000   Min.   :0.01300
 1st Qu.: 1.17   1st Qu.:1.100   1st Qu.:0.5800   1st Qu.:0.8000   1st Qu.:0.01875
 Median : 2.10   Median :1.100   Median :1.1000   Median :0.8000   Median :0.02500
 Mean   : 2.77   Mean   :1.119   Mean   :0.9762   Mean   :0.8313   Mean   :0.02628
 3rd Qu.: 3.00   3rd Qu.:1.150   3rd Qu.:1.5000   3rd Qu.:0.9000   3rd Qu.:0.02925
 Max.   :18.50   Max.   :1.500   Max.   :1.6600   Max.   :1.0000   Max.   :0.06800
> nrow(data)
[1] 32
> length(data)
[1] 5
```

8.2 Interprétation des données :

On trace le ggpairs des données pour voir les corrélation entre nos données :



On constate que y est bien corrolée positivement avec X2 et X4 et bien corrélée négativement avec X1 avec une corrélation moins bonne avec X3. Ce graphe nous donne une bonne idée sur nos données avant d'utilsé la regression linéaire multiple.

8.3 Implémentation de la régression linéaire :

On va utiliser la régression linéaire multiple pour chercher à expliquer y en fonction de quatre variables X1,X2, X3, X4 :

```
res=lm(y~X1+X2+X3+X4,data=df)
summary(res)
pred=predict(model,test)
S=0
for( i in 1:20)
{
   print(i)
   print(pred[i])
   S=S+(pred[i]-test$y[i])^2}
mse=S/19
mse
```

On obtient les resultats suivants :
On trace quelques graphes qui sont utiles pour visualiser les résultats de la régression :

```
Call:
lm(formula = y ~ X1 + X2 + X3 + X4)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0037679 -0.0008603  0.0000849  0.0013507  0.0022679

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.012093   0.007323   1.651  0.11429
X1           0.001096   0.005889   0.186  0.85420
X2           0.008879   0.002706   3.281  0.00373 **
X3          -0.001032   0.004302  -0.240  0.81295
X4           0.001494   0.001483   1.007  0.32589
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001679 on 20 degrees of freedom
Multiple R-squared:  0.9108,    Adjusted R-squared:  0.8929
F-statistic: 51.04 on 4 and 20 DF,  p-value: 3.232e-10
```

8.4 Interprétations :

8.4.1 Significativité

Avant d'interpréter un coefficient (sens, magnitude de l'effet), il convient de s'assurer que celui ci est significatif, autrement dit, qu'il est significativement différent de zéro (H0, soit une absence d'effet). Pour cela on utilise un test de Student. On calcule la statistique t pour chaque variable.
On utilise souvent un niveau de 5% (soit un intervalle de confiance de 95%). Si la t-value calculée est supérieure (en valeur absolue) à la valeur théorique déterminée, alors on rejette H0 : Le coefficient est bien significativement différent de zéro, et on peut l'interpréter (signe, magnitude,..). Un autre moyen de réaliser le test est de regarder la p-value associée au coefficient, soit la probabilité pour que la valeur t-calculée si t supérieur en valeur absolue à la valeur théorique. Si cette probabilité est inférieur au seuil utilisé (ici 5%), alors le coefficient est significatif.

8.4.2 Signe du coefficient

Avec un modèle linéaire, le signe du coefficient associé à une variable indique le sens de l'effet de cette variable sur la variable à expliquer. Par exemple, ici les coefficients de X1,X2 et X4 sont positifs ; cela signifie que si on augmente X1 ça va avoir tendance à faire augmenter y.

8.4.3 Qualité du modèle

Enfin, on peut regarder la qualité de la régression (au regard des données), mesurée par le coefficient de détermination (R-Squared ou R2), qui se définit comme la part de variation dans la variable y qui est expliquée par des variations dans les variables explicatives (souvent exprimé en

Plus sa valeur est proche de 1, et plus l'adéquation entre le modèle et les données observées va être forte. Cependant, cette valeur est fortement influencée, entre autres, par le nombre de variables explicatives incluses dans la regression. Le R2 ajusté (Adjusted R-Squared) va alors tenir compte de ce nombre et sera donc plus correct.

8.4.4 Conclusion

Finalement on peut donc écrire notre modèle comme suit sous cette forme :

$$y = \alpha * X_1$$

ce modèle ceci en regardant de RMSE qui égale à 26 qui assez elevée dans notre cas.

9
Conclusion

Dans ce projet on vient de faire plusieurs études statistiques.
La première étant l'étude des analyse de composantes principales sur une base de données de la variation de la température dans les capitales européennes.
La deuxième est une analyse de régression linéaire multiple.