Exemple d'analyse en composante principale

Cet exemple est tiré du livre « An R and S-Plus Companion to Multivariate Analysis » de B. Everitt paru chez Springer et adapté du document disponible sur le site de l'Ecole Polytechnique Fédérale de Lausanne à l'adresse suivante : http://stat.epfl.ch/webdav/site/stat/shared/StatMultivariee09/commandes_serie8.pdf.

Il porte sur des données de pollution atmosphérique recueillies dans 41 villes Américaines : SO2 (dioxyde de soufre), Temp (température moyenne annuelle en Fahrenheit), Manuf (nombre d'entreprises de plus de 20 employés), Pop (taille de la population en 1970, exprimée en milliers), Wind (vitesse moyenne du vent annuelle en miles par heure), Percip (moyenne annuelle des précipitations en pouces), Jours (nombre moyen de jours avec précipitation par an).

Pop et Manuf sont des variables d'écologie humaine.
Temp, Wind, Precip, Days sont des variables de climat. La température est donnée en valeurs négatives de sorte que pour toutes les 6 variables des valeurs élevées représentent une moins bonne attractivité environnementale.

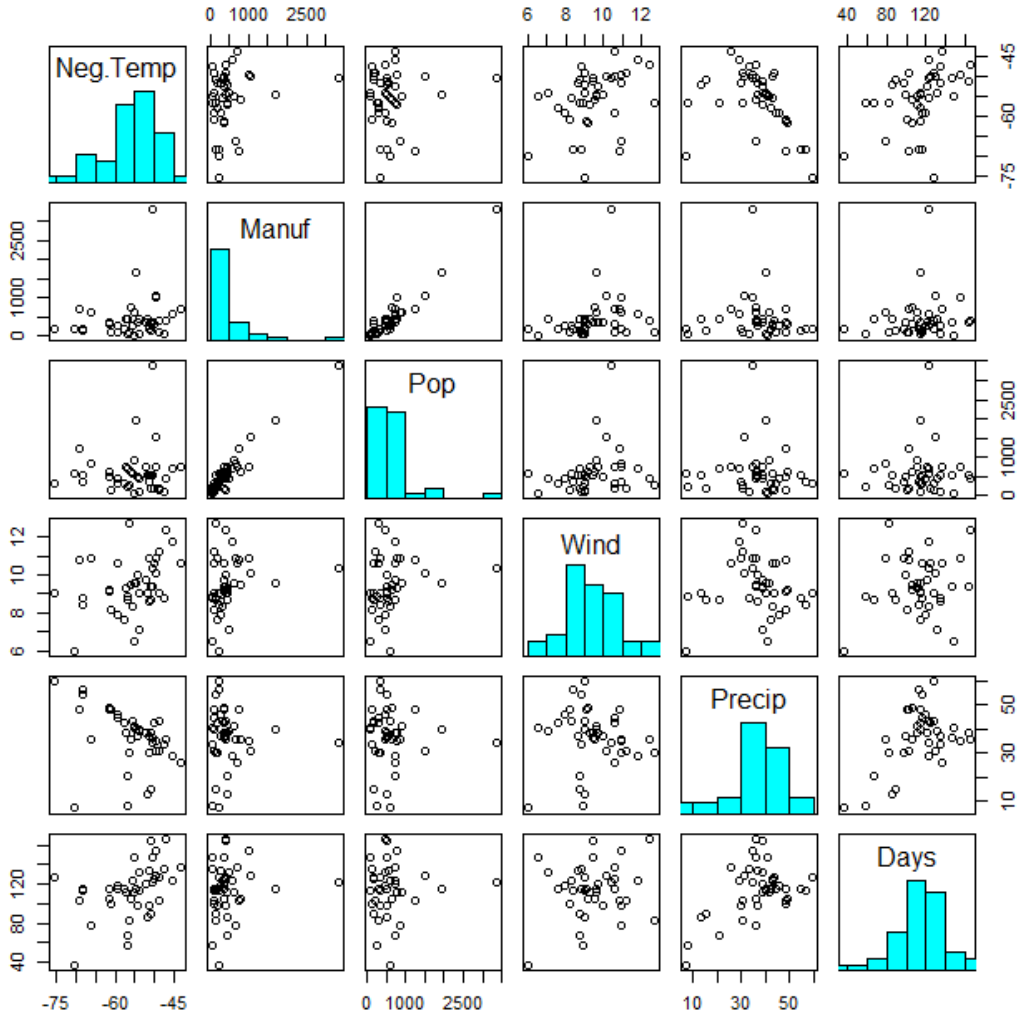
L'objectif de cette analyse est d'explorer la structure des données recueillies (sans prendre en compte SO2), puis d'étudier les déterminants de la pollution (c'est-à-dire, déterminer ce qui agit sur le niveau de SO2.

Seuls certains points de l'analyse sont décrits dans ce document.

L'analyse a été effectuée avec le logiciel R. Le code utilisé est donné en fin de ce document.

Description des données

Avant l'ACP, on peut faire des graphiques (variables 2 à 2) pour étudier un peu le comportement général des données. Par exemple :



En regardant de manière approfondie ce graphique, on peut voir qu'il y a certainement une aberrance (Chicago, qui correspond aux points isolés et atypiques dans les figures) et peut-être même d'autres villes. Ce problème sera examiné plus tard.

Matrice des corrélations :

	Neg.Temp	Manuf	Pop	Wind	Precip	Days
Neg.Temp	1.00000000	0.19004216	0.06267813	0.34973963	-0.38625342	0.43024212
Manuf	0.19004216	1.00000000	**0.95526935**	0.23794683	-0.03241688	0.13182930
Pop	0.06267813	0.95526935	1.00000000	0.21264375	-0.02611873	0.04208319
Wind	0.34973963	0.23794683	0.21264375	1.00000000	-0.01299438	0.16410559
Precip	-0.38625342	-0.03241688	-0.02611873	-0.01299438	1.00000000	**0.49609671**
Days	0.43024212	0.13182930	0.04208319	0.16410559	0.49609671	1.00000000

ACP avec la matrice des corrélations pour mettre toutes les variables sur la même échelle

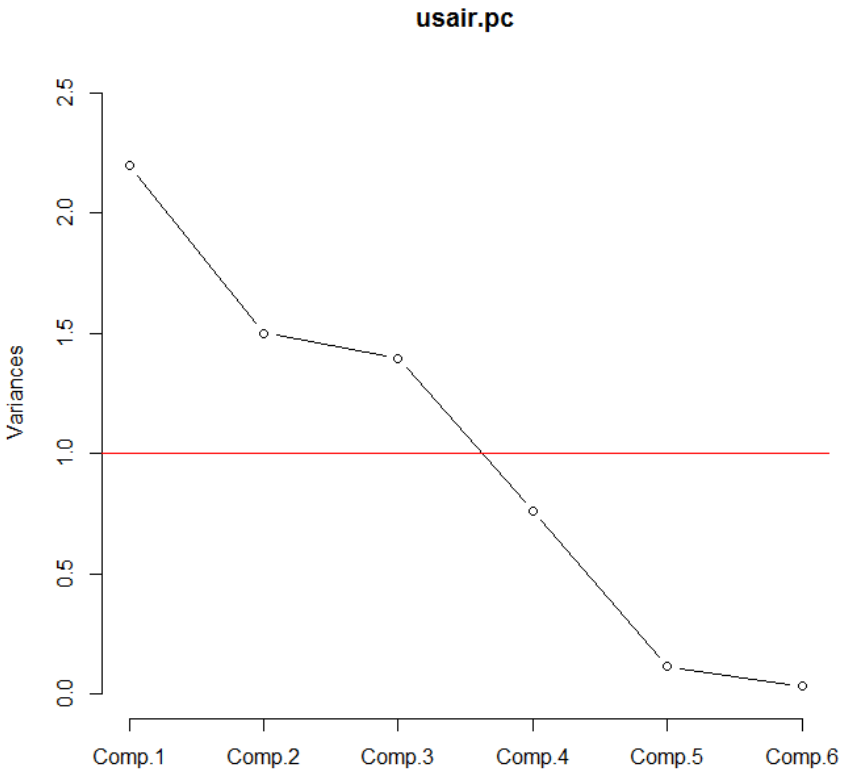
Importance des composantes :

```
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5      Comp.6
Standard deviation     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287 0.185599752
Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511 0.005741211
Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879 1.000000000
```

`Standard deviation` correspond aux écarts-types associes aux axes. Le carre correspond aux variances=valeur propres.

```
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
2.19616264 1.49994343 1.39464912 0.76022689 0.11457065 0.03444727
```

Les 3 premières composantes (valeurs propres > 1) expliquent 84,8 % de la variance initiale.

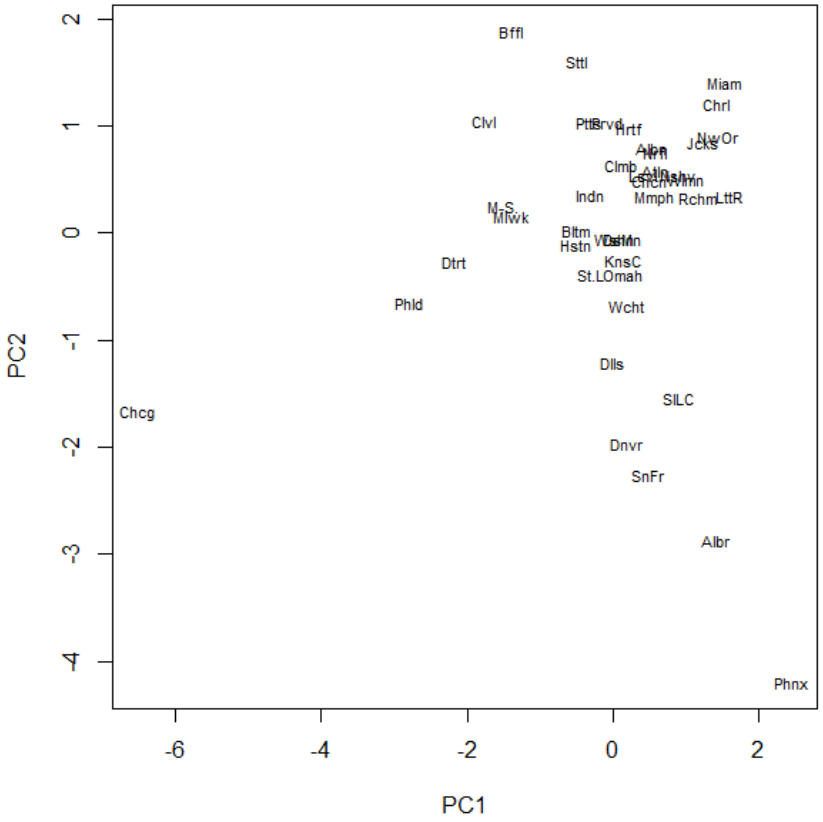


Matrice des composantes (les valeurs vides indiques des valeurs proches de 0) :

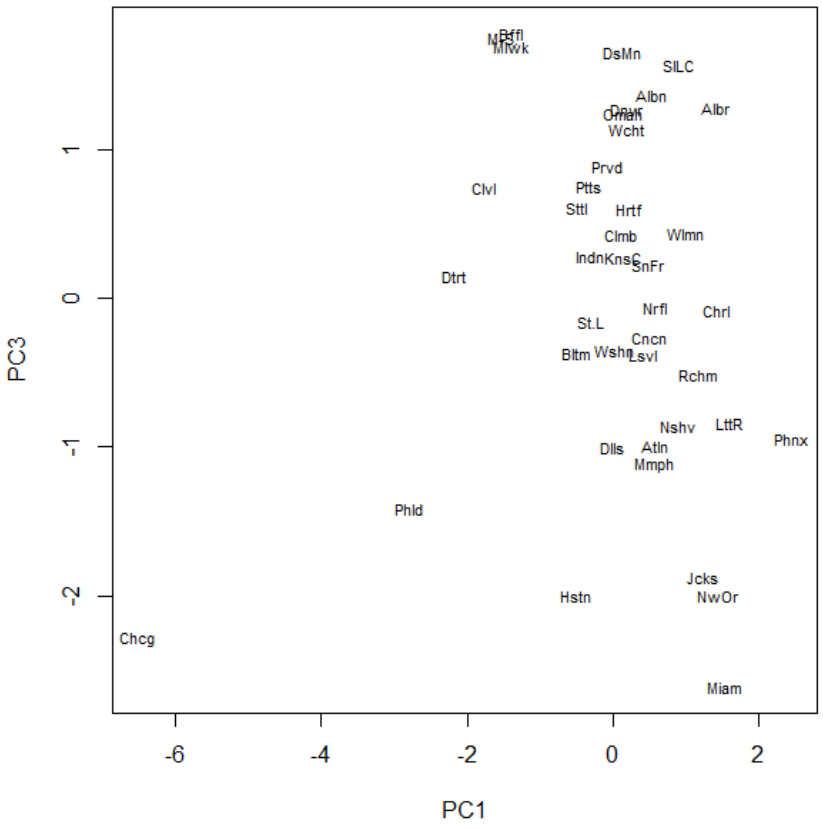
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Neg.Temp	-0.330	0.128	**0.672**	0.306	0.558	0.136
Manuf	**-0.612**	-0.168	-0.273	0.137	0.102	-0.703
Pop	**-0.578**	-0.222	-0.350			0.695
Wind	-0.354	0.131	0.297	-0.869	-0.113	
Precip		**0.623**	**-0.505**	-0.171	0.568	
Days	-0.238	**0.708**		0.311	-0.580	

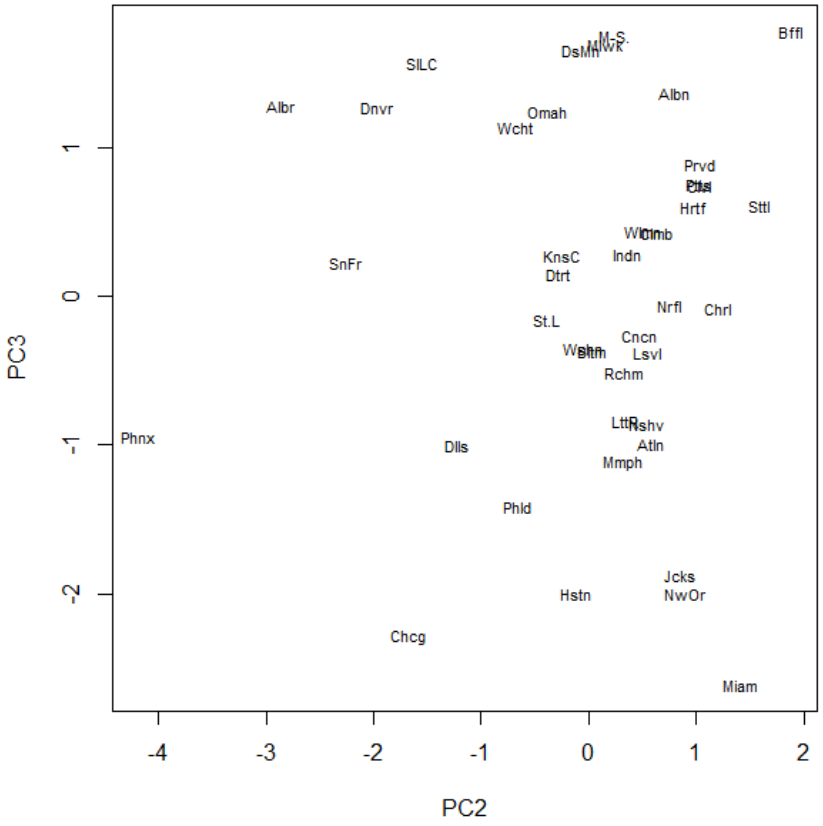
La première composante peut être considérée comme représentant un indice de « qualité de vie » avec des valeurs élevées indiquant un environnement relativement pauvres.
La seconde composante représente le « niveau d'humidité » des villes avec des valeurs élevées pour Precip et Days.
La troisième composante représente le contraste entre Precip et Neg.Temp et permet de séparer des villes ayant des températures élevées et de fortes précipitations de celles dont le climat est plus froids, plus secs. Elle correspond au « type de climat ».

Nous pouvons projeter les villes sur les différentes composantes (2 à 2)[1] :



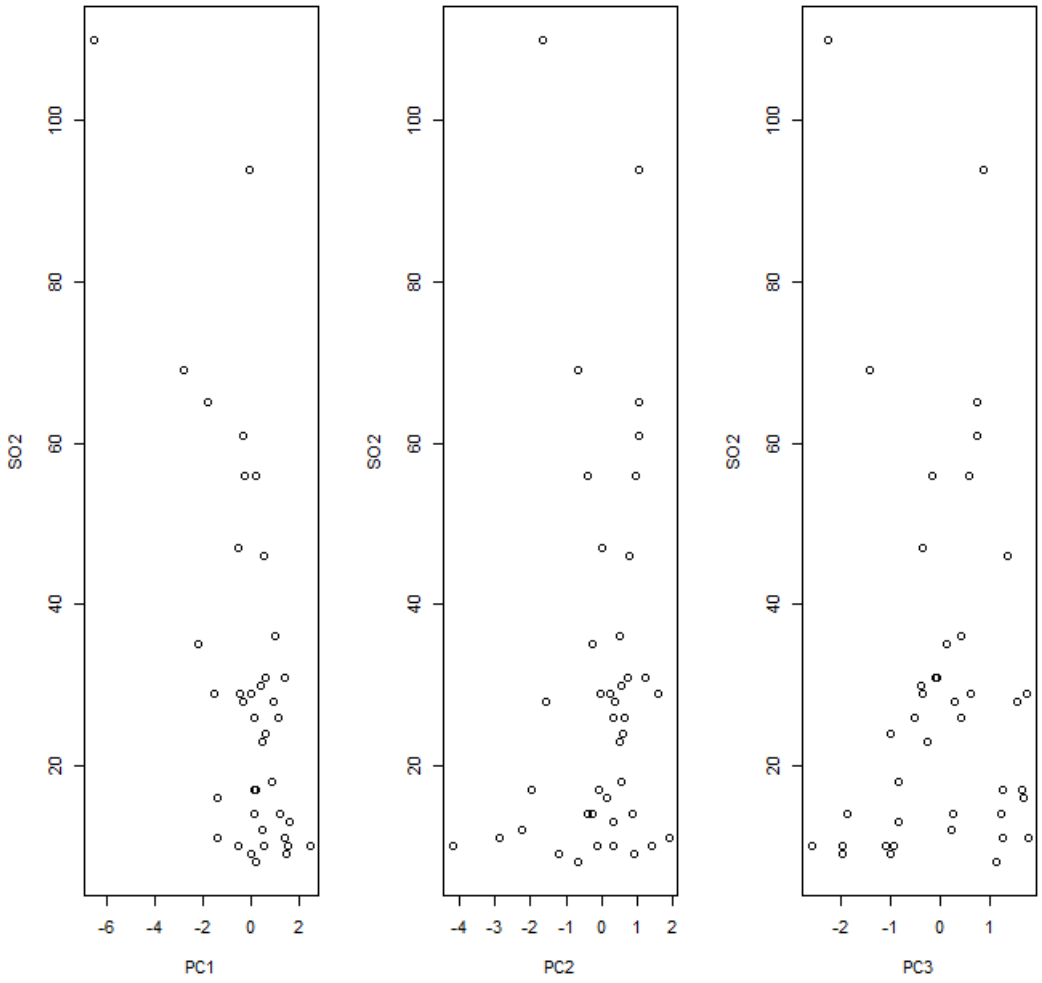
[1] Une représentation avec les axes allant de -1 à +1 est bien sur possible. Elle nécessite certains calculs qui ne sont pas détaillés ici.


PC3
PC1
-6
-4
-2
0
2
1
0
-1
-2
DsMn
SILC
Albn
Albr
Wcht
Prvd
Clvl
Ptts
Sttl
Hrtf
Clmb
Wlmn
Dtrt
SnFr
Nrfl
Chrl
St.L
Cncn
Lsvl
Rchm
Nshv
LttR
Phnx
Dlls
Atln
Mmph
Phld
Jcks
Hstn
NwOr
Chcg
Miam


PC3
PC2
-4
-3
-2
-1
0
1
2
1
0
-1
-2
Bffl
SILC
DsMn
Albn
Albr
Dnvr
Omah
Wcht
Prvd
Hrtf
Sttl
Clmb
KnsC
Indn
SnFr
Dtrt
Nrfl
Chrl
St.L
Cncn
Lsvl
Rchm
Phnx
Dlls
Atln
Mmph
Phld
Jcks
Hstn
NwOr
Chcg
Miam

On constate à nouveau que Chicago est atypique et que peut-être Phoenix et Philadelphia le sont aussi. Phoenix parait être la ville avec la meilleure qualité de vie par opposition à Philadelphia ou Chicago. Buffalo est préférable si vous souhaitez vivre dans un climat plutôt sec.

La figure suivante montre les valeurs de SO2 en fonction des composantes.



Il semble que la pollution ne soit en relation qu'avec la première composante principale. Cela est maintenant étudié dans un modèle de régression.

Etude des déterminants de la pollution

Une régression linéaire de SO2 avec les 3 composantes principales est ainsi effectuée :

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            30.049      2.907  10.336 1.85e-12 ***
usair.pc$scores[, 1]   -9.942      1.962  -5.068 1.14e-05 ***
usair.pc$scores[, 2]    2.240      2.374   0.943    0.352
usair.pc$scores[, 3]   -0.375      2.462  -0.152    0.880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.62 on 37 degrees of freedom
Multiple R-squared: 0.4182,     Adjusted R-squared: 0.371
F-statistic: 8.866 on 3 and 37 DF,  p-value: 0.0001473
```

Clairement la pollution est prédite par la première composante principale uniquement (la seule pour laquelle le degré de signification, `Pr(>|t|)` soit $\leq 0{,}05$). Quand la qualité de vie diminue (augmentation de la première composante principale) on a une augmentation de la pollution.

Code R utilisé

Remarque : Il est nécessaire d'avoir téléchargé et installé les packages suivants : scatterplot3d, rgl, bpca

```
# Charger le fichier USAIR dans R.
usair <- read.csv2("c:\\USAIR.csv", sep=""), row.names=1)

# On enlève la variable SO2 car on va s'intéressé à l'expliquer par l'ensemble
# des autres variables.
SO2 <- usair[, c("SO2")]
usair <- usair[, c(2:7)]

# Graphiques (variables 2 à 2) pour étudier un peu le comportement général des données.
# Par exemple :
panel.hist <- function(x, ...){
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
}
pairs(usair, diag.panel=panel.hist)

# ACP avec la matrice des corrélations pour mettre toutes les variables
# sur la même échelle.
usair.pc <- princomp(usair, cor=T)
summary(usair.pc)

# Calcul des valeurs propres :
summary(usair.pc)$sdev^2

# Scree plot des valeurs propres
screeplot(usair.pc, type=c("lines"), ylim=c(0, 2.5))
abline(h=1, col="red")

# Examen des vecteurs propres :
usair.pc$loadings

# Graphes des villes projetées sur les 3 premières composantes principales
# (représentations 2 à 2) :
old.par <- par(no.readonly=TRUE)
par(pty="s")
# Composante 1 vs 2 :
plot(usair.pc$scores[,1],usair.pc$scores[,2],
xlab="PC1",ylab="PC2",type="n",lwd=2)
text(usair.pc$scores[,1],usair.pc$scores[,2],
labels=abbreviate(row.names(usair)),cex=0.7,lwd=2)
# Composante 1 vs 3 :
plot(usair.pc$scores[,1],usair.pc$scores[,3],
xlab="PC1",ylab="PC3",type="n",lwd=2)
text(usair.pc$scores[,1],usair.pc$scores[,3],
labels=abbreviate(row.names(usair)),cex=0.7,lwd=2)
# Composante 2 vs 3 :
plot(usair.pc$scores[,2],usair.pc$scores[,3],
xlab="PC2",ylab="PC3",type="n",lwd=2)
```

```
text(usair.pc$scores[,2],usair.pc$scores[,3],
labels=abbreviate(row.names(usair)),cex=0.7,lwd=2)
par(old.par)

#Graphes des villes (en 3 dimensions sur les 3 premières composantes principales).
library(bpca)
bp <- bpca(usair, lambda.end=3)
plot(bp, rgl.use=T, var.factor=2)

#Graphes de la variables SO2 avec chacune des composantes principales :
old.par <- par(no.readonly=TRUE)
par(mfrow=c(1,3))
attach(usair)
plot(usair.pc$scores[,1],SO2,xlab="PC1")
plot(usair.pc$scores[,2],SO2,xlab="PC2")
plot(usair.pc$scores[,3],SO2,xlab="PC3")
par(old.par)

# Régression de SO2 avec les 3 composantes principales :
summary(lm(SO2~usair.pc$scores[,1]+usair.pc$scores[,2]+usair.pc$scores[,3]))
```