

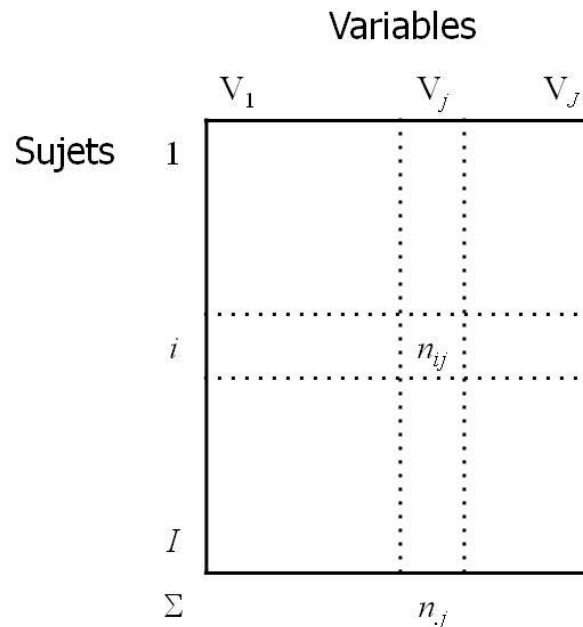
Analyse en Composantes Principales

Pr Roch Giorgi

 roch.giorgi@univ-amu.fr

Introduction (1)

- Étudier simultanément un nombre important de variables quantitatives
 - ✓ 2 variables quantitatives \Rightarrow nuage de points (espace de dimension 2)
 - ✓ J variables \Rightarrow espace de dimension J !



Introduction (2)

- Méthode pour obtenir un résumé « pertinent » des données initiales
 - ✓ Travail sur la matrice des corrélations (ou variances-covariances) entre les variables
 - On analyse la dispersion des données observées : la variance totale de toutes les variables à l'étude
 - ✓ Pour en extraire des dimensions plus globales en déformant le moins possible la réalité
 - Extraction, par un procédé mathématique adéquat, de composantes « en petit nombre »
 - Ces composantes étant orthogonales (indépendantes)
 - ✓ Et revenir à un espace de dimension réduite

Introduction (3)

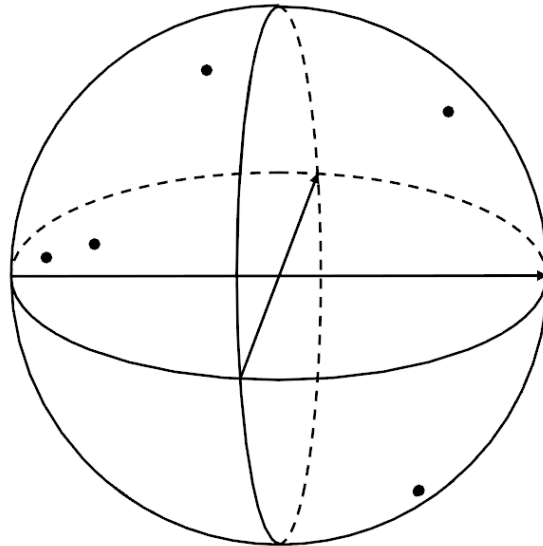
- Méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes
 - ✓ Possible également avec des variables qualitatives
- Construction du modèle statistique associé, estimation
- Représentations graphiques des individus, des variables et simultanée
- Qualité de représentation

Objectifs

- Typologie des individus
 - ✓ Ressemblances entre individus (lignes) du point de vue de l'ensemble des variables
 - ✓ En minimisant les déformations du nuage des points dans un sous-espace de dimension $q < J$
- Bilan des liaisons entre variables
 - ✓ Visualisation de la matrice des corrélations
 - ✓ Trouver des variables synthétiques explicitant au "mieux" les liaisons initiales entre ces variables
- Caractérisation des individus à partir des variables
- Réduction de l'espace dimension par un tableau de rang $q < J$

Approche Géométrique Intuitive (1)

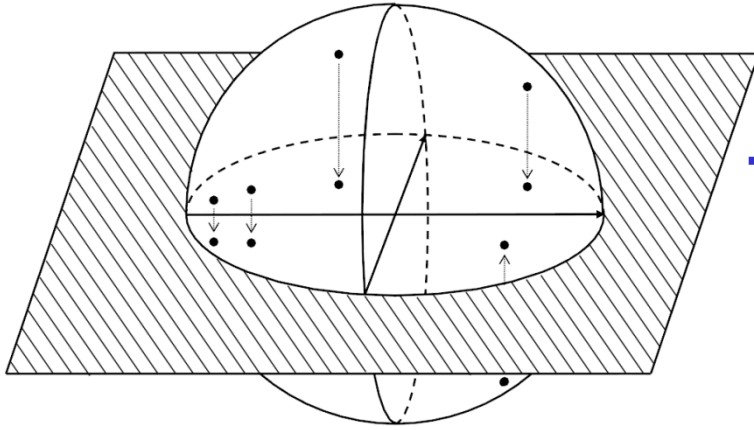
Exemple : 5 points sur une sphère de dimension de 3



Objectif : trouver une projection de ces points sur un plan en déformant le moins possible la structure originale

Source : *B. Falissard. The analyse of questionnaire data with R. CRC Press, 2012.*

Approche Géométrique Intuitive (2)

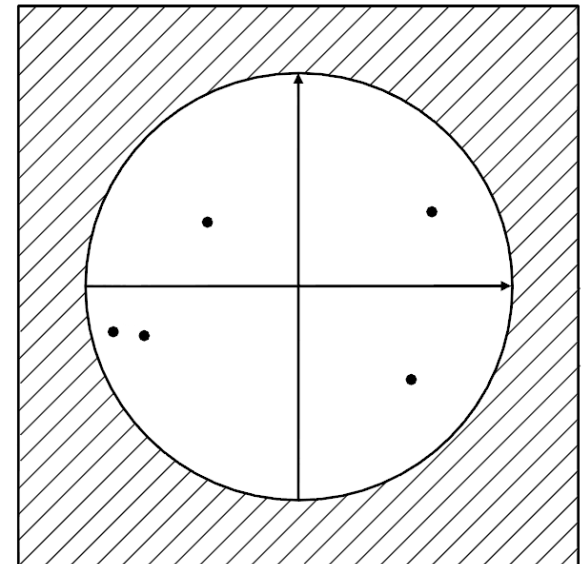


Pour comprendre globalement l'association entre les variables voici le plan sur lequel les points sont projetés

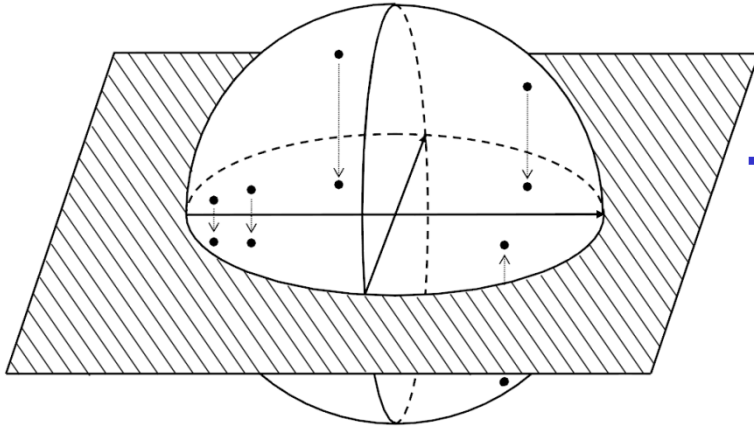


Si 2 points de l'hyper-sphère sont proches et s'ils sont à proximité du plan :

- Leurs projections seront également proches
- Les projections seront à proximité du cercle qui correspond à l'intersection de l'hyper-sphère avec le plan



Approche Géométrique Intuitive (3)

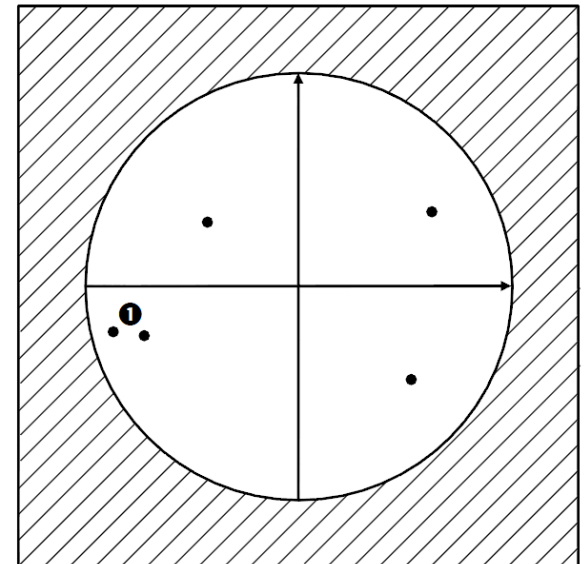


Pour comprendre globalement l'association entre les variables voici le plan sur lequel les points sont projetés

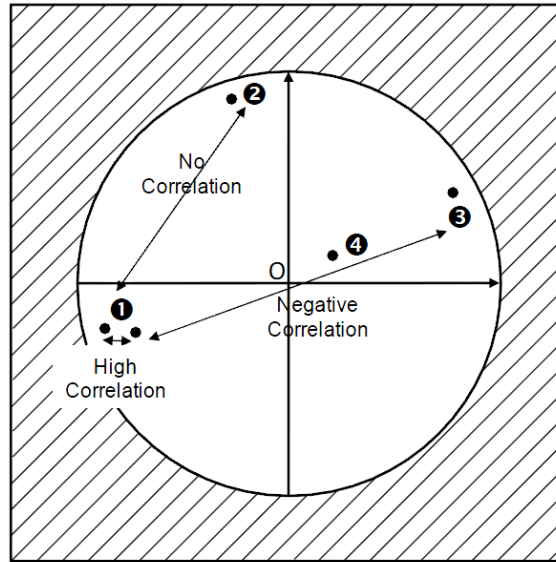


Les points en ❶ sont proches et à proximité du cercle

- Les points originels sur l'hyper-sphère sont également proches
- Les 2 variables correspondantes sont positivement et plutôt fortement corrélées



Approche Géométrique Intuitive (4)



Quand 2 points sont proches et à proximité du cercle, les variables correspondantes sont positivement corrélées ①

Quand 2 points sont à proximité du cercle et forment un angle droit avec l'origine (O), les variables correspondantes ne sont pas corrélées ②

Quand 2 points sont diamétralement opposés et à proximité du cercle, les variables correspondantes sont négativement corrélées ③

Quand un point est proche de (O), rien ne peut être dit ④

Principes (1)

- Analyse reposant sur la dispersion des données observées (variance totale observée)
- Utilisation de la matrice des variances-covariances (ou des corrélations)
- Procédure mathématique pour « extraire » des composantes (transformations mathématiques des variables initiales) expliquant la structure des données
 - ✓ Composante des observations est une combinaison linéaire des variables

$$\text{Composante}_c = \text{coef}_{c1} \cdot V_1 + \text{coef}_{c2} \cdot V_2 + \dots + \text{coef}_{cJ} \cdot V_J$$

où les coefficients sont à estimer

Principes (2)

Algorithme ayant 2 contraintes

1. La première composante doit maximiser la proportion de variance expliquée dans les variables initiales (V_1, \dots, V_J)

$$C_1 = a_{11}V_1 + a_{12}V_2 + \dots + a_{1J}V_J$$

- ✓ Les coefficients a_1 sont déterminés de manière à ce que la première composante maximise la proportion de variance expliquée dans les variables initiales (V_1, \dots, V_J)
- ✓ Une certaine contrainte est mise sur le vecteur a_1
- ✓ a_1 est le vecteur propre de la matrice de variance-covariance correspondant à sa plus grande valeur

Principes (3)

Algorithme ayant 2 contraintes (cont.)

2. Extraction des composantes suivantes indépendantes des précédentes

$$C_2 = a_{21}V_1 + a_{22}V_2 + \dots + a_{2J}V_J$$

⋮

$$C_J = a_{J1}V_1 + a_{J2}V_2 + \dots + a_{JJ}V_J$$

✓ Idem pour les coefficients a_j

Propriétés

- Si les valeurs propres de la matrice de variance-covariance sont notées $\lambda_1, \lambda_2, \dots, \lambda_J$, la variance de la $j^{\text{ème}}$ composante principale vaut λ_j
- La variance totale des J composantes principales = total de la variance des variables originales
 - ✓ La $j^{\text{ème}}$ composante principale compte pour une proportion P_j de la variation totale des données originales
 - ✓ Les m premières composantes principales, $m < J$, comptent pour une $P^{(m)}$ de la variation totale des données originales

Principes (4)

- La proportion de variance expliquée diminue avec le nombre de composantes
- La proportion de variance totale cumulée pourra atteindre 100 %
 - ✓ Nombre de composantes extraites = nombre de variables
- Critères statistiques pour déterminer le nombre de composantes pertinentes à extraire
- Réalisation de graphiques dans l'espace de dimension définie par le nombre de composantes extraites
 - ✓ Représentation des distances euclidiennes inter individus
 - ✓ Représentation des corrélations inter variables

Exemple

Pollution de l'air dans des villes des Etats-Unis

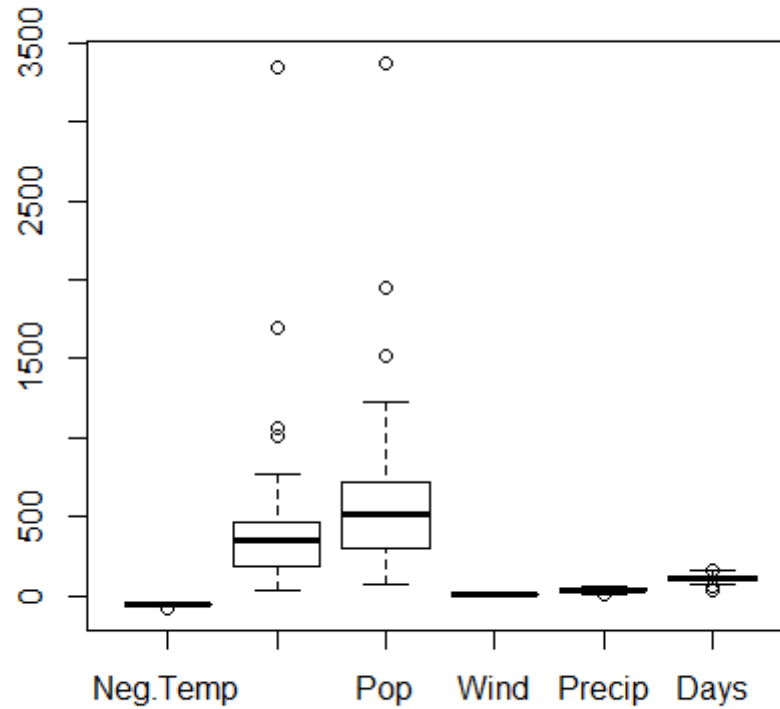
- Objectif initial de l'étude : Etudier les déterminants de la pollution atmosphérique par SO₂
- 41 villes
- 7 variables mesurées
 - ✓ SO₂, Neg.Temp, Manuf, Pop, Wind, Precip, Days
- Objectifs de l'analyse :
 1. Exploration des données par ACP
 2. Prédicteurs de SO₂

Descriptif

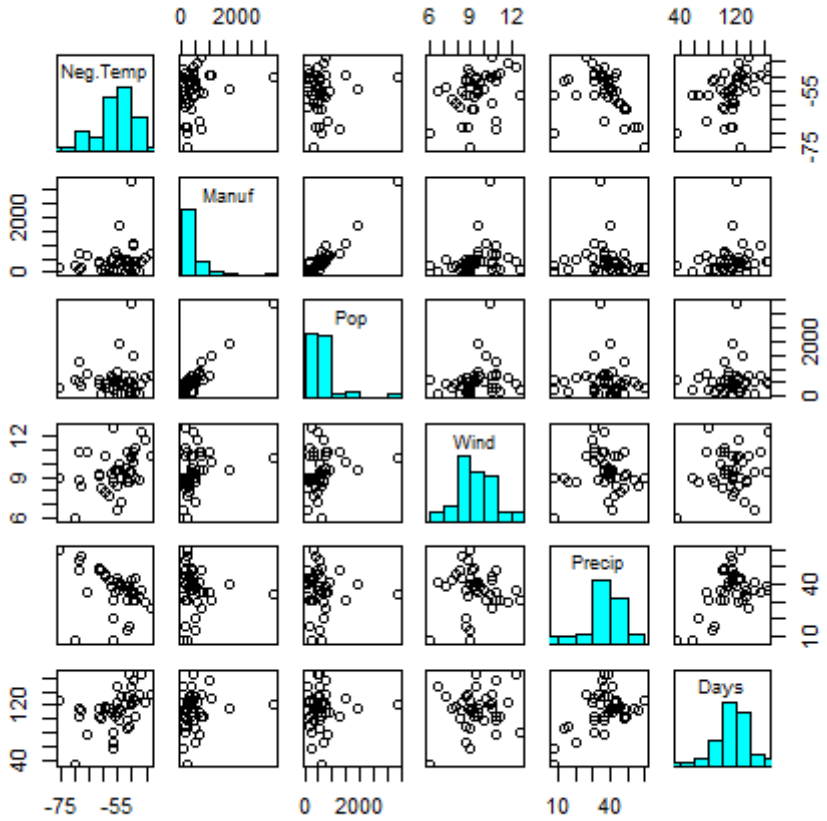
Neg. Temp	Manuf	Pop
Min. : -75.50	Min. : 35.0	Min. : 71.0
1st Qu.: -59.30	1st Qu.: 181.0	1st Qu.: 299.0
Median : -54.60	Median : 347.0	Median : 515.0
Mean : -55.76	Mean : 463.1	Mean : 608.6
3rd Qu.: -50.60	3rd Qu.: 462.0	3rd Qu.: 717.0
Max. : -43.50	Max. : 3344.0	Max. : 3369.0

Wind	Precip	Days
Min. : 6.000	Min. : 7.05	Min. : 36.0
1st Qu.: 8.700	1st Qu.: 30.96	1st Qu.: 103.0
Median : 9.300	Median : 38.74	Median : 115.0
Mean : 9.444	Mean : 36.77	Mean : 113.9
3rd Qu.: 10.600	3rd Qu.: 43.11	3rd Qu.: 128.0
Max. : 12.700	Max. : 59.80	Max. : 166.0

Box-Plot



Treillis



Matrice de Corrélations (1)

	Neg.Temp	Manuf	Pop	Wind	Precip	Days
Neg.Temp	1	0,19	0,06	0,35	-0,39	0,43
Manuf	0,19	1	0,96	0,24	-0,03	0,13
Pop	0,06	0,96	1	0,21	-0,03	0,04
Wind	0,35	0,24	0,21	1	-0,01	0,16
Precip	-0,39	-0,03	-0,03	-0,01	1	0,5
Days	0,43	0,13	0,04	0,16	0,5	1

- Coefficients de corrélation linéaire des variables prises 2 à 2
 - ✓ Diagonale ne comporte que des 1
 - ✓ Matrice symétrique par rapport à la diagonale
- Coefficients
 - ✓ Positifs ou négatifs
 - ✓ Faibles ou forts
 - ✓ De -1 à $+1$
- Un certain degré d'intercorrélation est nécessaire pour extraire une composante correspondant à une fonction linéaire des variables initiales

Matrice de Corrélations (2)

- La matrice ne doit pas être singulière
 - ✓ Matrice singulière
 - Au moins une variable est parfaitement corrélée avec une autre variable ou avec une combinaison de plusieurs variables (ex : total des scores)
 - Déterminant = 0
 - ✓ Son déterminant doit être \neq de 0 ($> 0,00001$)
- La matrice ne doit pas être la matrice d'identité
 - ✓ Matrice d'identité
 - Matrice ne comportant que des 0 sauf des 1 sur la diagonale
 - Déterminant = 1
 - ✓ Son déterminant doit être \neq de 1
 - ✓ Testé par le test de sphéricité de Barlett

Matrice de Corrélations (3)

	Neg.Temp	Manuf	Pop	Wind	Precip	Days
Neg.Temp	1	0,19	0,06	0,35	-0,39	0,43
Manuf	0,19	1	0,96	0,24	-0,03	0,13
Pop	0,06	0,96	1	0,21	-0,03	0,04
Wind	0,35	0,24	0,21	1	-0,01	0,16
Precip	-0,39	-0,03	-0,03	-0,01	1	0,5
Days	0,43	0,13	0,04	0,16	0,5	1

```
> det(cor(usair[, -1]))  
[1] 0.01378401
```

```
> bartlett(usair[, -1])  
$pvalue  
[1] 6.878465e-31
```

Résultats de l'ACP (1)


Eigenvalues	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	①
Variance	2.196	1.500	1.395	0.760	0.115	0.034	②
% of var.	36.603	24.999	23.244	12.670	1.910	0.574	③
Cumulative % of var.	36.603	61.602	84.846	97.516	99.426	100.000	④

- ① Nombre de composantes extraites = 6 (\Leftrightarrow # variables)
- ② Valeurs propres de chaque composante (somme = 6)
 - ✓ Chaque variable a 1 unité de variance
- ③ % de variance totale expliquée par chaque composante
- ④ % cumulé de variance totale expliquée

Résultats de l'ACP (2)

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Phnx	5.069	-2.440	6.613	0.232	-4.191	28.563	0.684	0.942	1.550	0.035
LttR	1.939	-1.612	2.884	0.691	0.342	0.191	0.031	0.840	1.233	0.188
SnFr	2.337	-0.502	0.280	0.046	-2.255	8.271	0.932	-0.227	0.090	0.009
Dnvr	2.398	-0.207	0.048	0.007	-1.963	6.267	0.670	-1.266	2.804	0.279
Hrtf	1.474	-0.219	0.053	0.022	0.976	1.550	0.439	-0.595	0.618	0.163
Wlmn	1.256	-0.996	1.102	0.629	0.501	0.408	0.159	-0.433	0.328	0.119
Wshn	0.415	-0.023	0.001	0.003	-0.055	0.005	0.017	0.354	0.219	0.727
Jcks	2.456	-1.228	1.674	0.250	0.849	1.172	0.120	1.876	6.156	0.584
Miam	3.521	-1.533	2.611	0.190	1.405	3.209	0.159	2.607	11.882	0.548
Atln	1.328	-0.599	0.398	0.203	0.587	0.561	0.195	0.995	1.733	0.562



- Résultats sur les individus (les « villes » ici)
 - ✓ Position sur chaque composante



Résultats de l'ACP (3)

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Phnx	5.069	-2.440	6.613	0.232	-4.191	28.563	0.684	0.942	1.550	0.035
LttR	1.939	-1.612	2.884	0.691	0.342	0.191	0.031	0.840	1.233	0.188
SnFr	2.337	-0.502	0.280	0.046	-2.255	8.271	0.932	-0.227	0.090	0.009
Dnvr	2.398	-0.207	0.048	0.007	-1.963	6.267	0.670	-1.266	2.804	0.279
Hrtf	1.474	-0.219	0.053	0.022	0.976	1.550	0.439	-0.595	0.618	0.163
Wlmn	1.256	-0.996	1.102	0.629	0.501	0.408	0.159	-0.433	0.328	0.119
Wshn	0.415	-0.023	0.001	0.003	-0.055	0.005	0.017	0.354	0.219	0.727
Jcks	2.456	-1.228	1.674	0.250	0.849	1.172	0.120	1.876	6.156	0.584
Miam	3.521	-1.533	2.611	0.190	1.405	3.209	0.159	2.607	11.882	0.548
Atln	1.328	-0.599	0.398	0.203	0.587	0.561	0.195	0.995	1.733	0.562

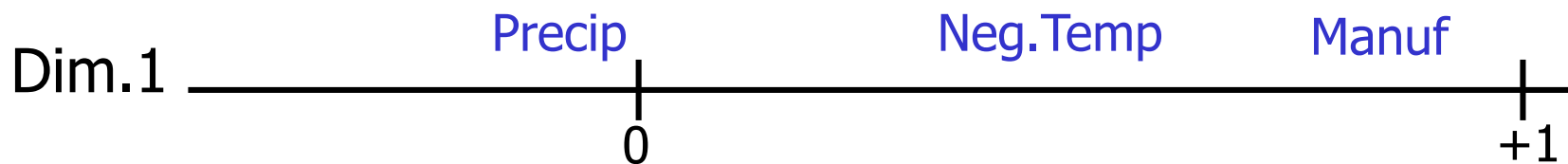


Contribution des individus sur la Dim.1 (leur somme=100)

Résultats de l'ACP (5)

Variables	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Neg.Temp	0.489	10.867	0.239	0.156	1.628	0.024	-0.793	45.116	0.629
Manuf	0.906	37.398	0.821	-0.206	2.824	0.042	0.322	7.447	0.104
Pop	0.856	33.388	0.733	-0.272	4.949	0.074	0.414	12.276	0.171
Wind	0.524	12.520	0.275	0.160	1.711	0.026	-0.351	8.836	0.123
Precip	-0.060	0.167	0.004	0.763	38.795	0.582	0.596	25.458	0.355
Days	0.353	5.660	0.124	0.867	50.093	0.751	-0.110	0.867	0.012

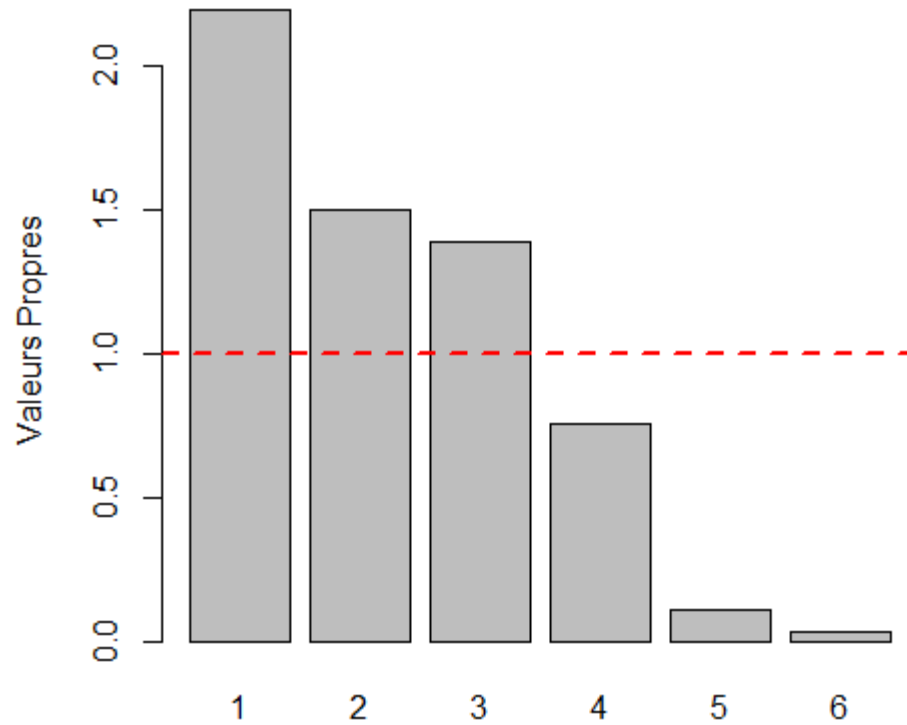
- Résultats sur les variables
 - ✓ Position sur chaque composante



Détermination du Nombre de Composantes

- Règle de Kaiser
 - ✓ Si tous les éléments sont indépendants, les composantes sont toutes de variance=1
 - ✓ Conservation des composantes dont valeur propre > 1
- Eboulis des valeurs propres (screeplot)
 - ✓ Critère graphique
 - ✓ Arrêt de l'extraction des composantes déterminé par le changement de pente

Nombre de Composantes



Résultats de l'ACP sur 3 Composantes (1)

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Variance	2.196	1.500	1.395	0.760	0.115	0.034
% of var.	36.603	24.999	23.244	12.670	1.910	0.574
Cumulative % of var.	36.603	61.602	84.846	97.516	99.426	100.000

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Phnx	5.069	-2.440	6.613	0.232	-4.191	28.563	0.684	0.942	1.550	0.035
LttR	1.939	-1.612	2.884	0.691	0.342	0.191	0.031	0.840	1.233	0.188
SnFr	2.337	-0.502	0.280	0.046	-2.255	8.271	0.932	-0.227	0.090	0.009
Dnvr	2.398	-0.207	0.048	0.007	-1.963	6.267	0.670	-1.266	2.804	0.279
Hrtf	1.474	-0.219	0.053	0.022	0.976	1.550	0.439	-0.595	0.618	0.163
Wlmn	1.256	-0.996	1.102	0.629	0.501	0.408	0.159	-0.433	0.328	0.119
Wshn	0.415	-0.023	0.001	0.003	-0.055	0.005	0.017	0.354	0.219	0.727
Jcks	2.456	-1.228	1.674	0.250	0.849	1.172	0.120	1.876	6.156	0.584
Miam	3.521	-1.533	2.611	0.190	1.405	3.209	0.159	2.607	11.882	0.548
AtlN	1.328	-0.599	0.398	0.203	0.587	0.561	0.195	0.995	1.733	0.562

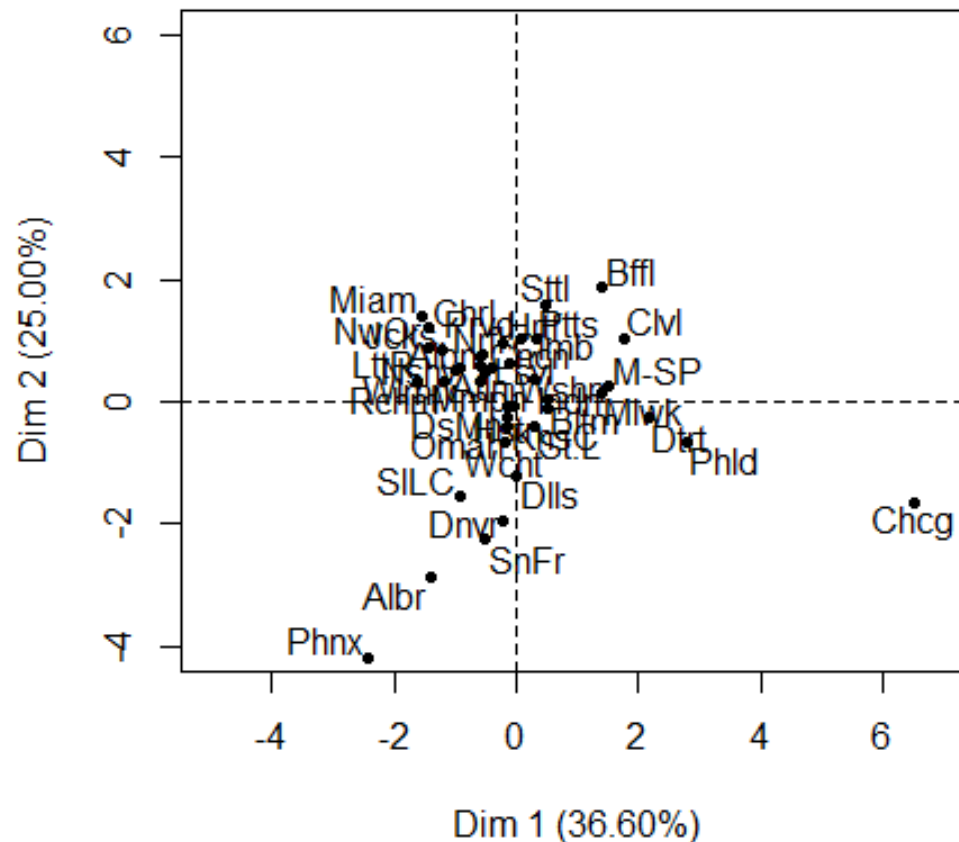
Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Neg.Temp	0.489	10.867	0.239	0.156	1.628	0.024	-0.793	45.116	0.629
Manuf	0.906	37.398	0.821	-0.206	2.824	0.042	0.322	7.447	0.104
Pop	0.856	33.388	0.733	-0.272	4.949	0.074	0.414	12.276	0.171
Wind	0.524	12.520	0.275	0.160	1.711	0.026	-0.351	8.836	0.123
Precip	-0.060	0.167	0.004	0.763	38.795	0.582	0.596	25.458	0.355
Days	0.353	5.660	0.124	0.867	50.093	0.751	-0.110	0.867	0.012

Résultats de l'ACP sur 3 Composantes (2)

Projection des individus sur les composantes

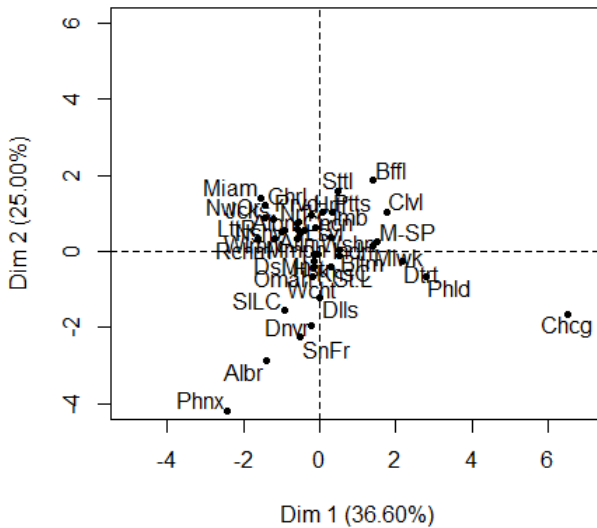
Individuals factor map (PCA)



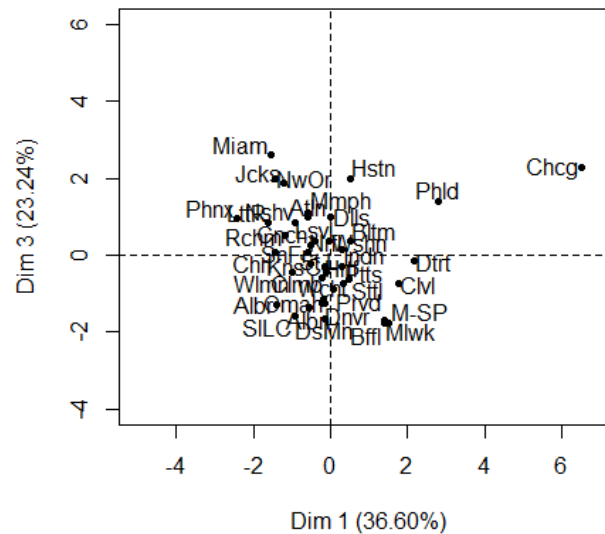
Résultats de l'ACP sur 3 Composantes (3)

Projection des individus sur les composantes

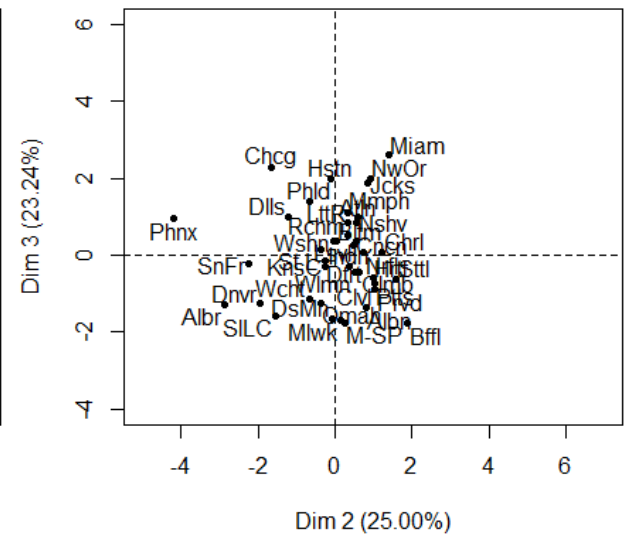
Individuals factor map (PCA)



Individuals factor map (PCA)

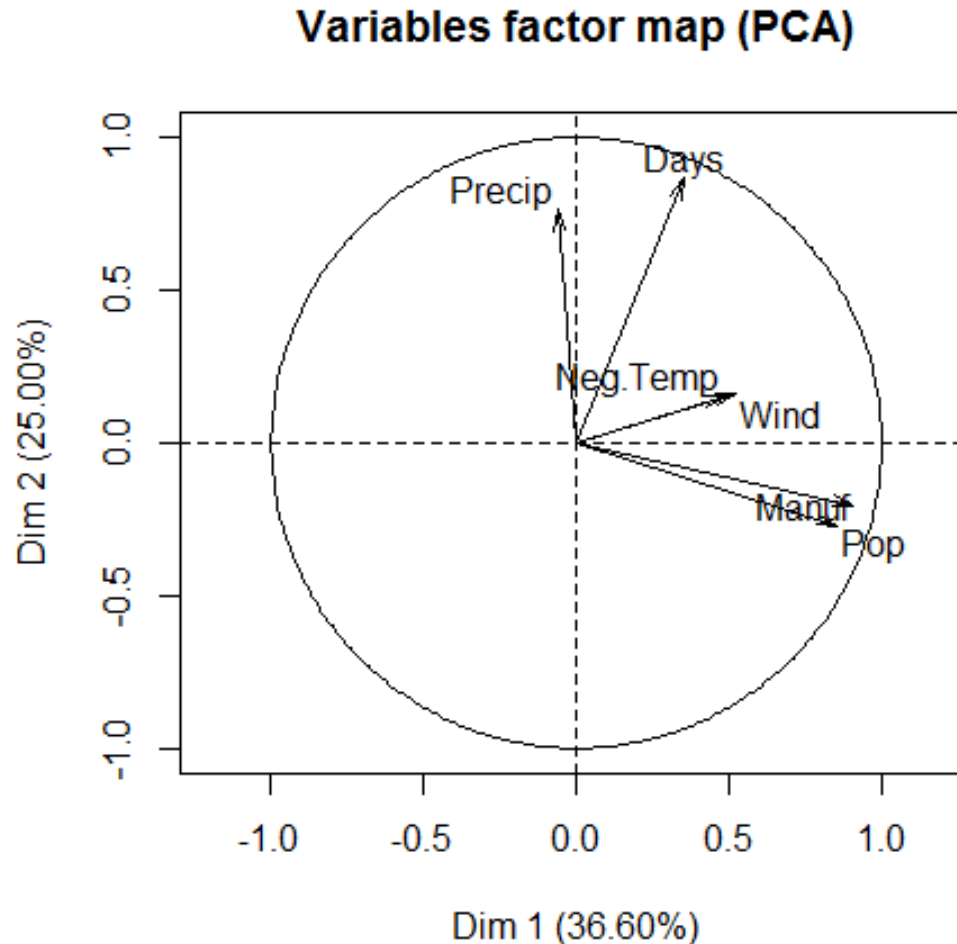


Individuals factor map (PCA)



Résultats de l'ACP sur 3 Composantes (4)

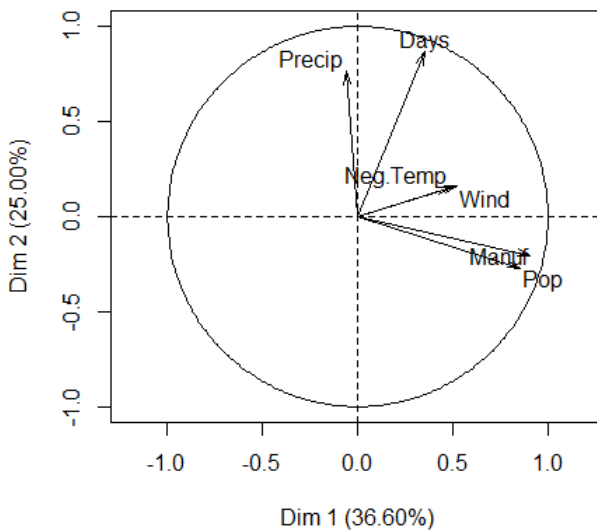
Projection des variables sur les composantes



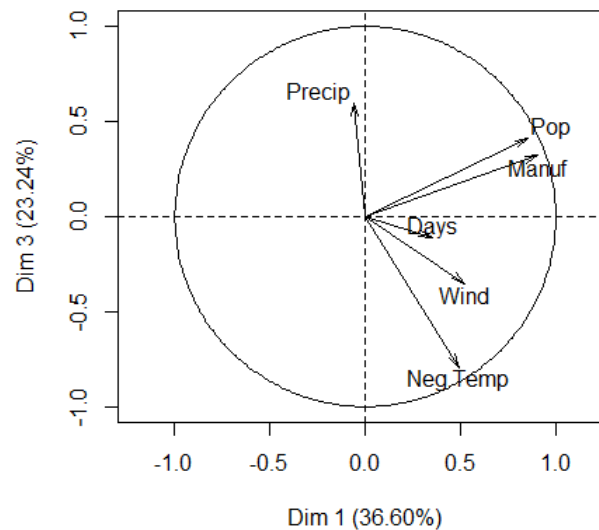
Résultats de l'ACP sur 3 Composantes (5)

Projection des variables sur les composantes

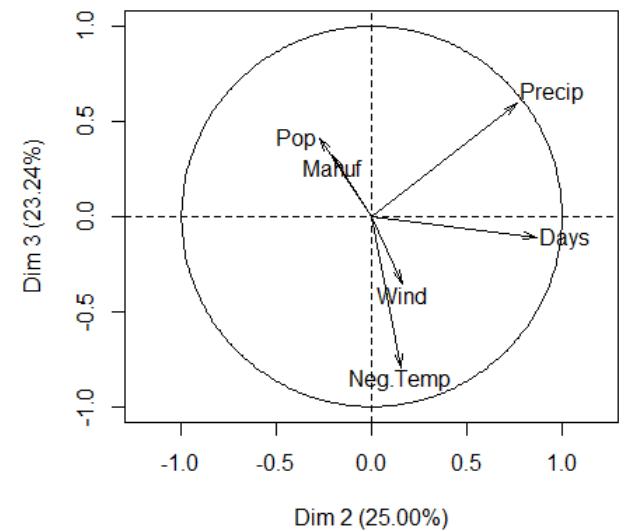
Variables factor map (PCA)



Variables factor map (PCA)



Variables factor map (PCA)



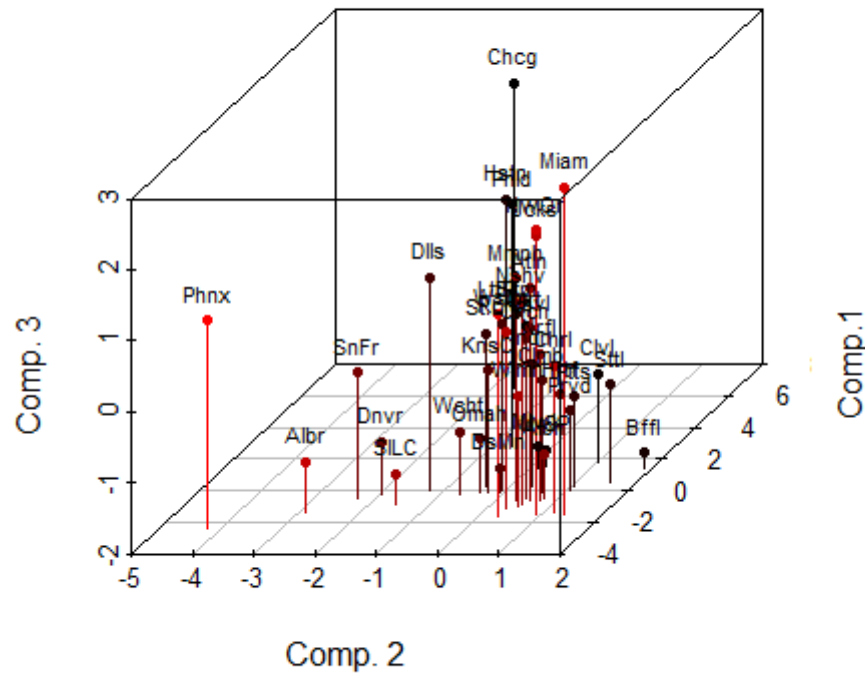
Description des Dimensions et Test

```
$Dim.1
$Dim.1$quanti
      correlation      p.value
Manuf      0.9062726 4.440892e-16
Pop        0.8563007 9.467982e-13
Wind       0.5243698 4.329391e-04
Neg.Temp   0.4885176 1.192814e-03
Days       0.3525785 2.376305e-02
```

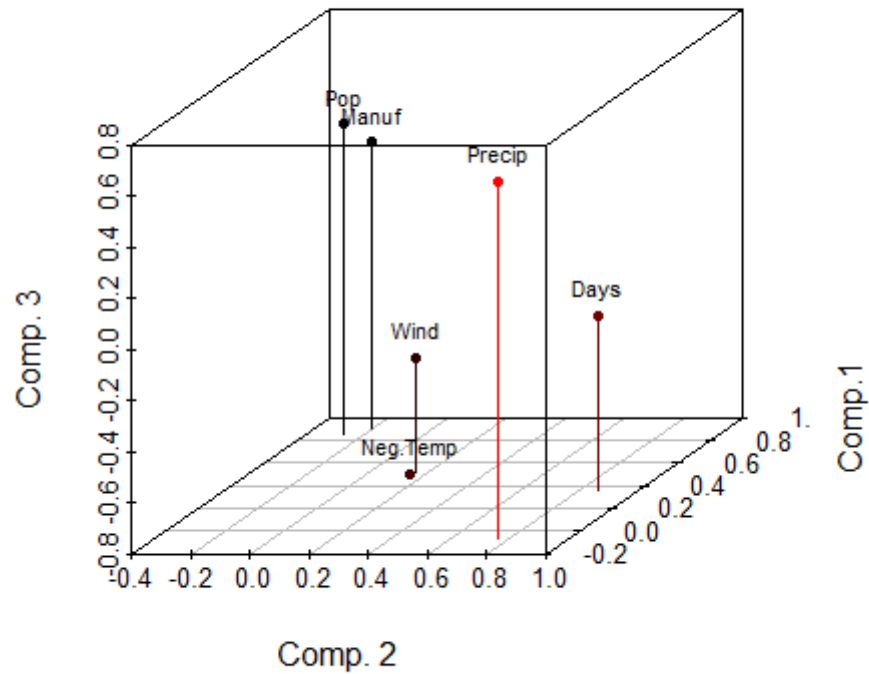
```
$Dim.2
$Dim.2$quanti
      correlation      p.value
Days   0.8668156 2.373657e-13
Precip 0.7628275 6.741009e-09
```

```
$Dim.3
$Dim.3$quanti
      correlation      p.value
Precip 0.5958649 3.950252e-05
Pop     0.4137753 7.161809e-03
Manuf   0.3222658 3.989110e-02
Wind    -0.3510421 2.442315e-02
Neg.Temp -0.7932295 6.253312e-10
```

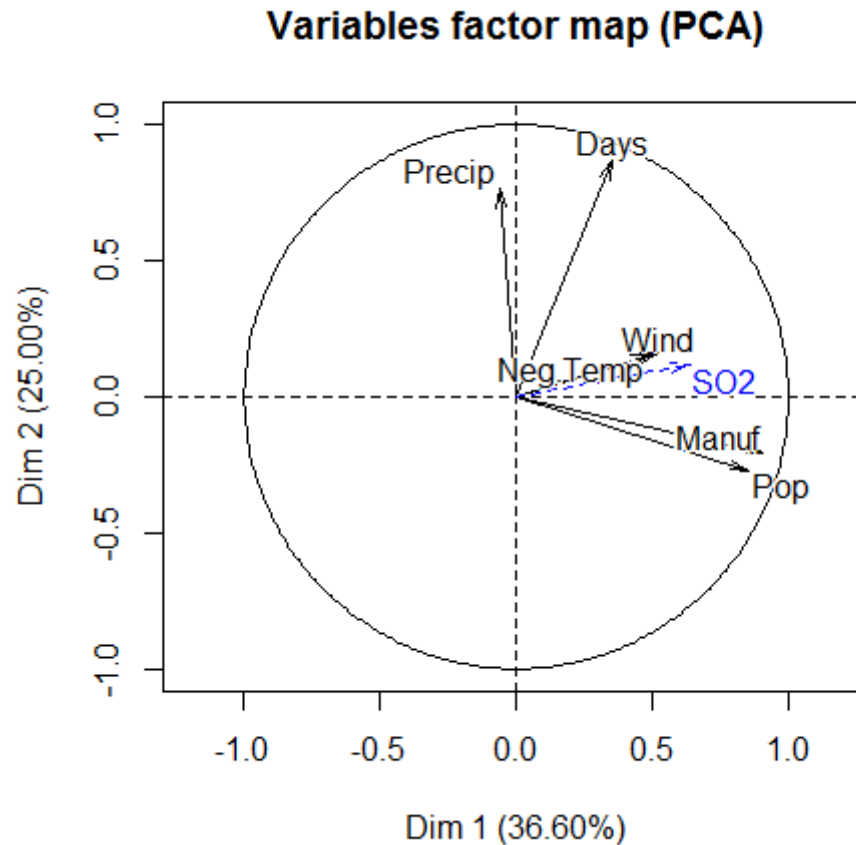
Représentation en 3D – Individus



Représentation en 3D – Variables



Ajout de Variable Supplémentaire



Effets sur la Pollution : SO2

- Régression linéaire de SO2 sur les composantes

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	30.049	2.907	10.336	1.85e-12
usair.pca3\$ind\$coord[, 1]	9.942	1.962	5.068	1.14e-05
usair.pca3\$ind\$coord[, 2]	2.240	2.374	0.943	0.352
usair.pca3\$ind\$coord[, 3]	0.375	2.462	0.152	0.880

...

Sources

- Brian S. Everitt. An R and S-PLUS Companion to Multivariate Analysis. Springer 2005.
- B. Falissard. The analyse of questionnaire data with R. CRC Press, 2012.
- Besse P., Baccini A. Data mining I - Exploration statistique, 2005.
- FactoMineR : <http://factominer.free.fr/>