# Analyse des données

# Chapitre 1 : Analyse en Composantes Principales (ACP)

Auteur : Sandrine Casanova

**Principe :**
Ce document constitue des notes de cours illustrées sur un jeu de données. Chaque concept est complété par un exemple qui contient des commentaires de sorties obtenues avec le logiciel R (package FactoMineR) données en annexe. Vous pouvez installer le package RcmdrPlugin.FactoMineR pour un menu FactoMineR dans Rcmdr. Les sorties R sont repérées par des numéros (ACP1, ACP2,...,ACP8). Les sorties SAS commentées, ainsi que le code SAS, sont également fournis à la fin du document.

## 1 Nature des données

Les données sont présentées sous la forme d'un tableau brut individus/variables noté $X$ et de taille $n \times p$ où $n$ est le nombre d'individus et $p$ est le nombre de variables quantitatives.

Notre exemple d'application concerne les 21 régions françaises sauf la Corse (les individus) caractérisées par différents indicateurs (les variables) de la démographie, de l'économie et de la géographie. Les variables considérées pour l'ACP sont les suivantes :

- 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)

## 2 Objectifs de l'ACP

L'analyse en composantes principales (ACP) est une méthode d'analyse multivariée descriptive appartenant à la famille des méthodes factorielles. L'ACP a deux objectifs principaux :

- *Résumer* le tableau $X$ par un petit nombre $k$ de nouvelles variables non corrélées entre elles et qui conservent au maximum l'information contenue dans les $p$ variables initiales.
  Intuitivement, on peut dire que ces nouvelles variables sont obtenues en “réunissant” les variables de départ qui sont bien corrélées entre elles. Le nombre $k$ de ces nouvelles variables est d'autant plus petit que les corrélations entre les $p$ variables initiales sont importantes. Comme sous-produit, l'ACP conduit à une visualisation des corrélations entre les variables initiales.
- *Interpréter* le tableau $X$ en utilisant les nouvelles variables et des représentations graphiques de type nuages de points.
  L'ACP permet notamment de repérer des individus atypiques ou des groupes d'individus ayant un comportement similaire par rapport aux caractères considérés.

# 3 Principe de l'ACP

## 3.1 Notion d'information

L'objectif de "conserver au maximum l'information contenue dans un tableau de données" suppose que l'on définisse mathématiquement la notion d'information. Cette information se fonde sur la variabilité des données et est mesurée par la variance.

*Définition :*
- l'information apportée par une variable quantitative $X^j$ est la variance de $X^j$.
- l'information apportée par un tableau de données $X = [X^1, X^2, \cdots, X^p]$ est la somme des variances des variables de $X$. On l'appelle *inertie* de $X$ et on la note $I_X$. On a :

$$I_X = \sum_{j=1}^{p} \mathrm{var}(X^j).$$

**Remarque 1** *Dans ce cours, l'ACP portera toujours sur des variables centrées réduites (de moyenne nulle et de variance égale à 1) $x^1, \cdots, x^p$ issues de $X^1, \cdots, X^p$ qui apportent chacune une information égale à 1. On note $x$ le tableau $X$ centré réduit.*
*En fait, il s'agit d'un cas particulier de l'ACP appelé parfois* ACP réduite.

## 3.2 Composantes principales

On définit la première composante principale $c^1$ comme une nouvelle variable combinaison linéaire des variables $x^1, x^2, \ldots, x^p$ s'exprimant sous la forme

$$c^1 = \alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_p x^p \text{ avec } \sum_{j=1}^{p} \alpha_j^2 = 1 \tag{1}$$

et telle que l'information apportée par $c^1$ est maximale.
Autrement dit, on cherche les coefficients $\alpha_1, \alpha_2, \ldots, \alpha_p$ tels que $c^1$ est de variance maximale.

La deuxième composante principale $c^2$ est définie comme étant une nouvelle variable *non corrélée avec* $c^1$, combinaison linéaire des variables $x^j$, $j = 1, \ldots, p$ et de variance maximale.

La troisième composante principale $c^3$ est *non corrélée avec* $c^1$ *et* $c^2$, combinaison linéaire des $x^j$ et de variance maximum.

...

La $p$ème composante principale $c^p$ est *non corrélée avec* $c^1$, $c^2$, ..., $c^{p-1}$, combinaison linéaire des $x^j$ et de variance maximum.

On a ainsi défini $p$ composantes principales non corrélées entre elles et que l'on peut regrouper dans un tableau de composantes principales noté $C = [c^1, c^2, \cdots, c^p]$.

D'après la définition précédente, $\mathrm{var}(c^1) \geq \mathrm{var}(c^2) \geq \cdots \geq \mathrm{var}(c^p)$. En effet, chacune des composantes est définie à partir d'un critère de maximisation de variance mais avec une contrainte de plus pour $c^2$ que pour $c^1$ (coefficient de corrélation nul entre $c^2$ et $c^1$), pour $c^3$ que pour $c^2$ (coefficient de corrélation nul entre $c^3$ et $c^2$),..., pour $c^{p-1}$ que pour $c^p$ (coefficient de corrélation nul entre $c^p$ et $c^{p-1}$).

De plus, on montre que l'information apportée par le tableau $x$ se retrouve entièrement reconstitué dans le tableau $C$. Autrement dit, l'inertie de $C$ est égale à l'inertie de $x$ :

$$I_x = \sum_{j=1}^{p} \mathrm{var}(x^j) = I_C = \sum_{j=1}^{p} \mathrm{var}(c^j).$$

Mais, alors que chacune des colonnes de $x$ apporte la même information (égale à 1), les colonnes du tableau $C$ apporte une information qui décroît avec le numéro de la colonne.

On comprend dès lors que l'on peut atteindre le premier objectif de l'ACP, c'est-à-dire résumer le tableau $x$ par un tableau contenant moins de colonnes si les dernières composantes principales apportent peu d'information (i.e. sont de faible variance).

**Remarque :** on peut introduire l'ACP par d'autres critères que la maximisation de variance. L'approche géométrique notamment est souvent adoptée.

# 4 Détermination des composantes principales et propriétés

## 4.1 Centrer et réduire les variables initiales

$$x^j = \frac{X^j - \overline{X^j}}{\sigma_{X^j}}$$

(ramener leur moyenne à 0 et leur variance à 1).

## 4.2 Calcul de la matrice des corrélations

On peut démontrer que :

$$R = \frac{1}{n}\, x'\, x.$$

**Remarque 2** *$R$ est aussi la matrice des corrélations entre les variables initiales.*

**Exemple 1** *Voir ACP0*

*POPUL, NBBREV, NBENTR et TELEPH sont très corrélées positivement entre elles.*
*CHOM est corrélée négativement avec TACT.*
*SUPERF n'est pas corrélée aux autres variables.*

## 4.3 Calcul des composantes principales

On calcule les valeurs propres (eigenvalues) et les vecteurs propres (eigenvectors) de $R$, c.a.d. les $\lambda_j \in \mathbb{R}$ et les $v_j \in \mathbb{R}^p$ pour $j = 1, \ldots, p$ t.q.

$$R\, v_j = \lambda_j\, v_j.$$

On trie les $\lambda_j$ par ordre décroissant :

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

On admet que la composante $c^j$ (vérifiant les contraintes ci-dessus) se calcule par la formule suivante :

$$c^j = x\, v_j$$

**Exemple 2** *Voir ACP5 : tableau des 3 premières composantes principales*

**Proposition 1** *$c^j$ est une nouvelle variable combinaison linéaire des variables initiales.*

**Proposition 2**
- $\overline{c_j} = 0$,
- $var(c^j) = \lambda_j$,
- $r(c^j, c^k) = 0$ *pour* $j \neq k$,

**Remarque 3**

$$I_x = p = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} var(c^j)$$

**Exemple 3** *Voir ACP1*
$\lambda_1 = 4,329...$, $\lambda_2 = 1,429...$, $\lambda_3 = 1,0124...$ *etc...*

# 5 Choix des composantes principales

On doit choisir un nombre $k$ suffisant pour résumer l'information (inertie) de départ sans trop en perdre.

$$\text{Information } = I = p$$

## 5.1 Critère de la variance expliquée

Information totale : $I = p$,
Information apportée par la composante principale $c^j$ : $\text{var}(c^j) = \lambda_j$,

part d'inertie expliquée par $c^1$. : $\lambda_1/p$,
part d'inertie expliquée par $c^1$ et $c^2$ : $(\lambda_1 + \lambda_2)/p$,

$$\ldots$$

part de variance expliquée par les $k$ premières composantes principales : $\sum_{j=1}^{k} \lambda_j/p$,

$$\ldots$$

part d'inertie expliquée par les $p$ composantes principales :

$$\sum_{j=1}^{p} \lambda_j/p = p/p = 100\%.$$

**Critère :** $k$ est choisi tel que la part d'inertie expliquée soit suffisamment grande.

**Exemple 4** *Voir ACP1*
*En retenant les* $k = 3$ *premières composantes principales, on explique* $96,74\%$ *de l'inertie totale.*

### 5.2 Critère de Kaiser

Les variables initiales ont une variance $= 1$ (réduites).

**Critère :** retenir les composantes principales de variance $> 1$ car elles apportent plus d'information que les variables initiales,

$$k = \text{ nombre de } \lambda_j > 1.$$

**Exemple 5** *Voir ACP1*

$\lambda_1 > 1, \lambda_2 > 1, \lambda_3 > 1$ *et* $\lambda_4 < 1$. *On retient donc les 3 premières composantes principales.*

### 5.3 Critère de la différence

On regarde les différences entre valeurs propres :

$$\lambda_1 - \lambda_2,\ \lambda_2 - \lambda_3,\ \ldots$$

En général, ces différences diminuent.

**Critère :** retenir les $k$ composantes principales telles que la différence $\lambda_k - \lambda_{k+1}$ soit grande tandis que les différences $\lambda_j - \lambda_{j+1},\ j = k+1, \ldots, p-1$ soient petites.

**Exemple 6** *Voir ACP3 (Tableau SAS)*

*Avec ce critère, on retient* $k = 1$ *composante principale.*

## 6 Interprétation des composantes principales

On suppose, qu'après utilisation d'un des critères précédents, on a selectionné $k$ (petit) composantes principales (ou $k$ dimensions ou $k$ facteurs).

**Exemple 7** *Quel que soit le critère, on a vu que* $k \leq 3 \Rightarrow$ *on prend* $k = 3$.

Une des difficultés de l'ACP (et des analyses factorielles en général) est l'interprétation des composantes principales.

### 6.1 Présentation du problème

L'ACP conduit à une réduction du nombre de variables (de $p$ à $k$) mais si on connaît la signification des variables initiales, il n'en est pas de même des composantes principales.

### 6.2 Interprétation des coefficients des combinaisons linéaires

$$\begin{aligned} c^j &= x\, v_j,\ \forall j = 1, \ldots, p \\ c^j &= \sum_{k=1}^{p} v_j^k\, x^k \end{aligned}$$

**Exemple 8**

$$\begin{aligned} c^1 &= 0,46POPUL + 0,34TACT + 0,45NBENTR \\ &+ 0,46NBBREV - 0,14CHOM + 0,46TELEPH\ ? \end{aligned}$$

On connaît la "composition" de $c^j$ et les variables $x^k$ importantes sont associées aux grands coefficients $v_j^k$ (parce qu'elles ont la même variance).

**Exemple 9** *Tableau des vecteurs propres de la matrice des corrélations sur le fichier des régions (avec la fonction* **princomp** *de R)*

`Loadings:`

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 |
|---|---|---|---|---|---|---|---|
| POPUL | 0.460 | -0.214 | | -0.208 | -0.233 | 0.704 | -0.384 |
| TACT | 0.349 | 0.496 | 0.148 | 0.720 | -0.301 | | |
| SUPERF | | -0.278 | 0.936 | | 0.185 | | |
| NBENTR | 0.456 | -0.234 | | -0.179 | -0.329 | -0.700 | -0.318 |
| NBBREV | 0.468 | | -0.157 | 0.145 | 0.839 | | -0.161 |
| CHOM | -0.144 | -0.737 | -0.256 | 0.605 | | | |
| TELEPH | 0.467 | -0.182 | | -0.116 | | | 0.851 |

*$c^1$ est surtout "constituée" des variables POPUL, NBBREV, TELEPH, NBENTR (dans le même sens : coeff. tous positifs).*

Mais cette méthode est rarement utilisée (grandeur des coeff. difficilement évaluable).

## 6.3 Étude des corrélations entre composantes principales et variables initiales

On peut directement interpréter ces corrélations.
Propriété :

$$r(c^j, x^k) = \sqrt{\lambda_j}\ v_j^k$$

**Exemple 10** *Voir ACP2 :*

- *$c^1$ est surtout très corrélée positivement avec POPUL, NBENTR, NBBREV et TELEPH.*
- *$c^2$ est fortement corrélée positivement avec CHOM et négativement avec TACT.*
- *$c^3$ est fortement corrélée positivement avec la variable SUPERF.*

*Question : pourquoi parler de $r(c^2,TACT)$=-0,59 et pas de $r(c^1,TACT)$=0,72 ? Il n'est pas possible de fixer des valeurs seuil aux coefficients de corrélation pour l'interprétation car cela dépend de l'ensemble des valeurs. On a 4 variables qui expliquent très bien $c^1$ et donc on choisit de ne pas intermpréter TACT sur $C^1$.Par contre, comme il est clair qu'il existe une corrélation négative entre $TACT$ et $CHOM$ et qu'il est intéressant de la remarquer dans l'interprétation de $c^2$, cela nous conduit à interpréter $TACT$ sur $C^2$.*

Mais en général, on préfère les **représenter** en considérant les composantes principales 2 par 2 et les interpréter graphiquement (possible car $k$ petit $\Rightarrow$ peu de possiblités).

**Exemple 11**

- *Figure 1 : plan $(c^1, c^2)$, on place les points de coordonnées : $(r(x^k, c^1), r(x^k, c^2))$ et pour repèrer les variables, on indique leur nom en abrégé sur le point.*

- *Figure 2 : plan* $(c^1, c^3)$*, on place les points de coordonnées :* $(r(x^k, c^1), r(x^k, c^3))$.

Les dessins s'inscrivent évidemment dans un carré $[-1, +1] \times [-1, +1]$ (coefficients de corrélation) et on peut montrer qu'en fait, les points sont toujours dans le cercle centré à l'origine et de rayon 1. On trace souvent ce cercle car il aide à l'interprétation.

En pratique,

- Pour chaque paire (ou plan) $((c^1, c^2), (c^1, c^3),$
  $(c^2, c^3), \dots (c^{k-1}, c^k))$, dessiner les corrélations.
- Tracer le cercle des corrélations.
- Repérer les corrélations fortes, c.a.d. les points proches du cercle. On ne doit pas s'intéresser aux variables trop éloignées du cercle car elles n'interviennent pas ou peu dans le calcul des composantes et donc ne servent pas à son interprétation.
- Interpéter chaque composante en fonction des corrélations fortes (positives et négatives).

**Exemple 12**
- *Figure 1 :*
  - $c^1$ *est très corrélée positivement avec POPUL, NBBREV, NBENTR, TELEPH donc* $c^1$ *peut s'interprèter comme une composante "Potentiel de développement économique" des régions (plan humain, économique).*
  - $c^2$ *est corrélée positivement avec CHOM et négativement avec TACT (elle oppose CHOM et TACT).* $c^2$ *est une mesure de l'activité de la région (+ cette variable est grande et - la région est active).*
- *Figure 2 :*
  - $c^1$ *déjà fait.*
  - $c^3$ *représente essentiellement la variable SUPERF.*

On remarque que plus le numéro de la composante est grand, moins on résume d'information. Le premier axe résume l'information apportée par 4 variables alors que le troisième axe n'intègre qu'une variable (plus trop d'intérêt compte tenu de l'objectif initial). Dans ce cas, on peut se contenter de 2 axes.

**Remarque 4** *La qualité de représentation des variables sur une axe est évaluée par le carré de la corrélation (appelé* $cos^2$ *dans R) de cette variable avec l'axe*

***Exemple 13*** *Voir tableau ACP3*

**Remarque 5** *Parfois, le premier axe incorpore toutes les variables. C'est le cas lorsque toutes les variables vont dans le même sens (fortement corrélées positivement). On parle alors de* **facteur de taille***.*

# 7 Interprétation des individus

On dispose de $k$ nouvelles variables dont on connaît la signification.

## 7.1 Graphique des individus

Pour interpréter le tableau de départ, on représente les individus sur les nouvelles variables choisies 2 à 2 (possible car $k$ petit). Lorsqu'on considère 2 composantes principales, on parle de **plan principal**.

**Exemple 14** *sur le plan* $(c^1, c^2)$ *(premier plan principal) : Figure 3, sur le plan* $(c^1, c^3)$ *: Figure 4.*

On interprète les graphiques obtenus comme n'importe quel dessin de type nuage de points en tenant compte de l'interprétation des comp. ppales. Mais, comme pour les corrélations, on ne doit pas interpréter des individus mal représentés.

**Exemple 15** *Figure 3 : Î de France se distingue des autres régions au niveau de la première composante (ou sur le premier axe).*

**Remarque 6** *On peut présenter l'ACP d'une facon* **géométrique** *très différente de celle proposée jusqu'ici. En particulier, on peut dire que :*
- *Un graphique permet une vision synthétique des données.*
- *Mais un graphique de type nuage de points est possible uniquement si il y a 2 variables (3 au plus). On est dans un espace (des individus)* $\mathbb{R}^p$ *impossible à représenter.*
- *Donc on* **projette** *les observations sur un espace* $\mathbb{R}^k$ *avec k petit tout en essayant de perdre le moins d'information possible.*
- *Mais problème : toute projection implique une déformation des distances (toujours plus courtes).*
- *Donc, pour interpréter les graphiques des individus, il faut que les distances soient bien conservée. En effet, des points, en apparence proches, peuvent être forts éloignés dans l'espace sur les autres dimensions laissées de côté par le graphique. D'où le paragraphe suivant.*

### 7.1.1 Mesure de la qualité de représentation des individus

On choisit, pour mesurer cette qualité de représentation, de regarder la **distance à l'origine** de chacun des individus.

- Au départ, un individu $x_i$ est à une certaine distance de l'origine :

$$d(x_i, O) = \sqrt{\sum_{j=1}^{p} \left(x_i^j\right)^2} = \sqrt{\sum_{j=1}^{p} \left(c_i^j\right)^2}$$

  (on parle aussi de norme de l'individu $x_i$ : $d(x_i, \mathrm{O}) = \|x_i\|$).
- Après projection, la norme devient :
  - sur un espace de dimension $k$,

$$\|Px_i\|_k = \sqrt{\sum_{j=1}^{k} \left(c_i^j\right)^2}.$$

  - sur un axe $c^j$ : $\|Px_i\|_1 = |c_i^j|$.

Pour chaque individu, on compare sa norme de départ avec sa norme après projection en calculant le rapport des 2, soit sur un axe :

$$\mathrm{RAP}_j = \frac{|c_i^j|}{\sqrt{\sum_{l=1}^{p} \left(x_i^l\right)^2}} = \frac{|c_i^j|}{\sqrt{\sum_{l=1}^{p} \left(c_i^l\right)^2}}.$$

**Exemple 16** *ACP 6 : colonnes RAP1, RAP2 et RAP3 (appelés* $cos^2$ *sous R) représentant les rapports de normes au carré.*
*On a interprété l'individu I sur la première composante mais est-il bien représenté ? oui car* 98% **de sa norme au carré est conservée ou reconstituée sur cet axe**.

**Remarque 7**

$$\sum_{j=1}^{p} RAP_j^2 = \sum_{j=1}^{p} \frac{\left(c_i^j\right)^2}{\sum_{l=1}^{p} \left(c_i^l\right)^2} = 100\%.$$

Règle (qui peut être modifiée selon les cas) : on dira qu'un individu est bien représenté
- sur l'axe 1 si $RAP_1 > 0.5$ pour cet individu
- sur l'axe 2 si $RAP_2 > 0.25$ pour cet individu
- sur l'axe 3 si $RAP_3 > 0.15$ pour cet individu

Pour un axe donné, on **interprète seulement les individus bien représentés sur cet axe**.

### 7.1.2 Commentaire de l'ACP

**Exemple 17** – *Figure 3 :*

- *Une région est bien représentée sur C1 si cos2>0.5. Donc les régions bien représentées sur C1 sont : Auvergne, Champagne-Ardennes,Ile de France, Picardie, Poitou-Charentes et Rhône-Alpes.*
  *U, E, I, D, T, et R peuvent être interprétés. On repère sur le graphique que l'Ile de France (I) est seule à droite sur l'axe 1. Ce qui signifie qu'elle a C1 élevée et donc (voir question 3) que Idf se caractérise par : popul, nbbrev, nbentr et teleph : élevées (car toutes les corr. sont positives). La région Rhône-Alpes (R) est aussi relativement à droite sur le graphique par rapport aux autres régions. Donc l'idf et Rhône-Alpes s'opposent à l'Auvergne, la Bourgogne, la Picardie et Poitou-Charentes qui ont un potentiel de développement démographique et économique peu élevé.*
  *Difficulté de l'interprétation car régions agglomérés à cause de I (refaire l'analyse sans I).*
- *Deuxième axe : une région est bien représentée sur C2 si cos2>0.25 Donc les régions bien représentées sur C2 sont : Alsace, Aquitaine, Basse- Normandie, Bourgogne, Bretagne, Franche-Comté, Languedoc-Roussillon, Limousin, Nord-Pas-de-Calais et Provence-Côte d'Azur.*
  *A, S, F sont opposées à G, P, Z et Q.*
  *A, S et F se caractérisent par un tact élevé et un chômage faible alors que G, P, Z et Q se caractérisent par un chômage élevé et un taux d'activité faible. Enfin, N, O et P sont des régions moyennement économiquement dynamiques.*

- *Figure 4, troisième axe uniquement : une région est bien représentée sur C3 si cos2>0.15 Donc les régions bien représentées sur C3 sont : Aquitaine, Bourgogne, Centre, Haute-Normandie, Midi-Pyrénées, Nord-pas de Calais, Pays de Loire, Picardie et Rhône-Alpes.*
  *P, H et D sont opposées aux régions Q, M, R et C P, H et D sont de superficie faible alors que C, Q, M et R ont une superficie relativement élevée. O et Y ont une superficie moyenne.*
  *Mais limite de la méthode : on ne commente pas A et I car mal représentées alors que ce sont les plus petites régions (Cf. tableau initial). Raison : RAP1(I)=0,98 et RAP2(A)=0,92 (plus rien sur RAP3). Dans ce cas, représenter SUPERF directement et conserver uniquement 2 composantes principales.*

# 8 Compléments

## 8.1 Effet "taille"

- Lorsque toutes les variables initiales sont corrélées positivement entre elles, la première composante principale définit un facteur de taille.
- Une matrice symétrique (dans notre cas la matrice des corrélations) ayant tous ses termes positifs admet un premier vecteur propre dont toutes les composantes sont de même signe.
- Si on les choisit positives, la première composante principale est alors corrélée positivement avec toutes les variables.

- Si de plus les corrélations entre variables sont toutes de même ordre, la première composante principale est proportionnelle à la moyenne des variables initiales.
- La deuxième composante principale différencie alors des individus de taille semblable. On l'appelle facteur de forme.

## 8.2 Rotation (Varimax)

- Une des difficultés de l'ACP est l'interprétation des axes.
- Mais lorsqu'il y a de nombreuses variables avec des corrélations moyennes, l'interprétation est difficile.
- Le rôle des méthodes de rotation est de rendre ces corrélations plus tranchées en faisant pivoter les axes, d'où une lecture facilitée.
- La rotation VARIMAX fait tourner les axes en préservant leur orthogonalité mais le premier facteur n'est plus l'axe de plus grande variance.
- La rotation Varimax cherche à maximiser la variance des corrélations dans chaque colonne du tableau des corrélations entre les composantes principales et les variables initiales.
- La quantité d'information traduite sur les 3 axes reste la même mais la répartition entre les 3 axes a été fortement modifiée.

## 8.3 Variables et individus supplémentaires (ou illustratifs)

- Il n'est pas étonnant de trouver de fortes corrélations entre la première composante et certaines variables initiales car $c^1$ maximise (admis) :

$$\sum_{j=1}^{p} r^2(c, x^j).$$

- Par contre, si on trouve une forte corrélation entre une composante principale et une variable qui n'a pas servi à l'analyse, le caractère probant de ce phénomène est plus élevé.

D'où la pratique courante de partager l'ensemble des variables en 2 groupes :

- d'une part les variables "actives" qui servent à déterminer les axes principaux,
- d'autre part les variables passives (ou supplémentaires ou illustratives) que l'on relie *a posteriori* aux composantes principales (avec le calcul des corrélations).
- On peut également ne pas faire participer à l'analyse une partie des individus (on calcule les corrélations sans eux), ce qui permettra de vérifier sur cet échantillon test des hypothèses formulées après une ACP sur des individus actifs.
- Il est de plus immédiat de positionner de nouveaux individus sur les axes principaux.
- On peut aussi choisir de placer en individus supplémentaires (illustratifs) certains individus atypiques.

## 8.4 Contribution d'un individu à un axe

- La contribution (en pourcentage) de l'individu $i$ à la composante $c^j$ est définie par

$$\frac{\frac{1}{n} {c_i^j}^2}{\lambda_j} \times 100.$$

- Une forte contribution est une contribution $> \dfrac{1}{n} \times 100$.

– Un individu ayant une forte contribution modifie l'analyse. On a donc intérêt à le porter en individu supplémentaire.

**Exemple 18** *Voir Tableau ACP7*

### 8.5 Contribution d'une variable à un axe

– Comme $\lambda_j = \sum_{k=1}^{p} r^2(c^j, x^k)$, la contribution (en pourcentage) de la variable $x^k$ à la composante $c^j$ est définie par

$$\frac{r^2(c^j, x^k)}{\lambda_j} \times 100.$$

– Mais cette quantité a peu d'intérêt en ACP car elle n'apporte rien de plus que le coefficient de corrélation.

**Exemple 19** *Voir Tableau ACP4*

## 9 Conclusion

L'ACP est une méthode statistique applicable :
– à un tableau individus / variables
– pour $p$ variables quantitatives
– $p > 3$
– certaines variables bien corrélées entre elles

**Remarque 8** *Si R=Id, alors l'ACP ne sert à rien car toutes les valeurs propres sont égales à 1 (R=Id).*

## Annexe 1 : sorties R (package FactoMineR, fonction PCA)

```
Tableau ACP0
        CHOM NBBREV NBENTR POPUL SUPERF  TACT TELEPH
CHOM    1.00  -0.26  -0.08 -0.07   0.06 -0.70  -0.10
NBBREV -0.26   1.00   0.89  0.92  -0.16  0.71   0.94
NBENTR -0.08   0.89   1.00  0.98   0.15  0.52   0.98
POPUL  -0.07   0.92   0.98  1.00   0.02  0.51   0.99
SUPERF  0.06  -0.16   0.15  0.02   1.00 -0.06   0.00
TACT   -0.70   0.71   0.52  0.51  -0.06  1.00   0.56
TELEPH -0.10   0.94   0.98  0.99   0.00  0.56   1.00
```

```
Tableau ACP1
> res$eig
        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
```

```
> res$var

$cor
             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             Tableau ACP2
NBBREV  0.97349367 -0.02238409 -0.15753009
CHOM   -0.29985584  0.88117115 -0.25791205
TELEPH  0.97224065  0.21803299 -0.05362987

$cos2
              Dim.1        Dim.2       Dim.3
POPUL  0.9174439224 0.0652687574 0.001974376
TACT   0.5288190696 0.3512116276 0.022291733
SUPERF 0.0002409085 0.1101390690 0.887414480
NBENTR 0.9003166591 0.0782606742 0.006545675          Tableau ACP3
NBBREV 0.9476899200 0.0005010477 0.024815730
CHOM   0.0899135264 0.7764626006 0.066518627
```

```
TELEPH 0.9452518801 0.0475383848 0.002876163

$contrib
             Dim.1       Dim.2      Dim.3
POPUL  21.189667460  4.56622163  0.1950122
TACT   12.213825781 24.57086965  2.2017901
SUPERF  0.005564123  7.70536194 87.6513473
NBENTR 20.794089045  5.47513998  0.6465268
NBBREV 21.888241635  0.03505345  2.4510893
CHOM    2.076680306 54.32155386  6.5701512
TELEPH 21.831931650  3.32579950  0.2840832
```

Tableau ACP4

```
> resu$ind
$coord
        Dim.1       Dim.2       Dim.3
A -0.28094406 -2.73490669 -0.76838621
Q -0.11019117  0.99830340  1.19600815
U -0.93391267 -0.34881724  0.09005665
N -0.79687591 -0.89506993 -0.52292379
O -0.59004881 -0.78480247  0.74941316
B -0.12586693  0.30933835  0.08262498
C -0.07318477 -0.56792323  1.43453917
E -1.02733194 -0.49766582  0.05071986
F -0.98301188 -1.55976911 -0.44708369
H -0.86148437 -0.08169737 -1.29015598
I  8.52387615 -0.36611635 -1.20448755
G -1.38054779  2.45105543 -0.64028945
S -1.14693317 -1.42793382 -0.44322460
L -0.82420562  0.03847953 -0.21496683
M -0.17473894  0.42101737  1.74503400
P -0.51766414  2.19202649 -1.87798069
Y  0.08276225  0.26274268  0.53608934
D -1.21228965  0.28624500 -0.72619507
T -0.92467241  0.15699988 -0.08286088
Z  0.85131980  1.86432697  0.10766701
R  2.50594606  0.28416692  2.22640243
```

Tableau ACP5

```
$cos2
        Dim.1        Dim.2       Dim.3
A 0.009024776 0.8552276735 0.067508094
Q 0.004899584 0.4021524538 0.577210076
U 0.766308069 0.1069022247 0.007125616
N 0.319426364 0.4029983800 0.137551763
O 0.221785947 0.3923546991 0.357767401
B 0.108588710 0.6558868895 0.046793352
C 0.002195883 0.1322351576 0.843708621
E 0.664189164 0.1558640064 0.001618923
```

Tableau ACP6

```
F 0.254000137 0.6394948446 0.052540485
H 0.237613553 0.0021369413 0.532918555
I 0.977335454 0.0018030494 0.019515246
G 0.225255255 0.7100326262 0.048453453
S 0.367468603 0.5695871402 0.054877172
L 0.449428640 0.0009796007 0.030572609
M 0.009334337 0.0541881209 0.930918885
P 0.030111196 0.5399125630 0.396290884
Y 0.013424420 0.1352982669 0.563255361
D 0.649121573 0.0361900411 0.232927000
T 0.815014466 0.0234956884 0.006544682
Z 0.160435396 0.7694118785 0.002566138
R 0.542620951 0.0069775053 0.428312211

$contrib
        Dim.1        Dim.2       Dim.3
A  0.086809057 24.918240556  2.77697466
Q  0.013354231  3.320148006  6.72792386
U  0.959263358  0.405348182  0.03814559
N  0.698403995  2.668985940  1.28614398
O  0.382913962  2.051885229  2.64152920
B  0.017424029  0.318786476  0.03210965
C  0.005890698  1.074512863  9.67915870   Tableau ACP7
E  1.160771928  0.825102637  0.01209954
F  1.062778731  8.104999281  0.94013550
H  0.816244227  0.022235584  7.82883659
I 79.909714674  0.446550507  6.82366070
G  2.096178253 20.014217475  1.92825951
S  1.446776818  6.792794994  0.92397560
L  0.747130490  0.004932781  0.21734802
M  0.033581859  0.590516971 14.32255754
P  0.294728061 16.007515980 16.58803857
Y  0.007533380  0.229981867  1.35172124
D  1.616360083  0.272965763  2.48038661
T  0.940375100  0.082116609  0.03229326
Z  0.797096286 11.579145499  0.05452279
R  6.906670782  0.269016800 23.31417890

$dist
        A         Q         U         N         O         B         C         E
2.9573426 1.5742264 1.0668530 1.4099552 1.2529130 0.3819612 1.5617667 1.2605640
        F         H         I         G         S         L         M         P    Tableau ACP8
1.9504813 1.7673062 8.6221446 2.9088008 1.8920299 1.2294339 1.8086227 2.9832118
        Y         D         T         Z         R
0.7143063 1.5046773 1.0242483 2.1254096 3.4019149
```

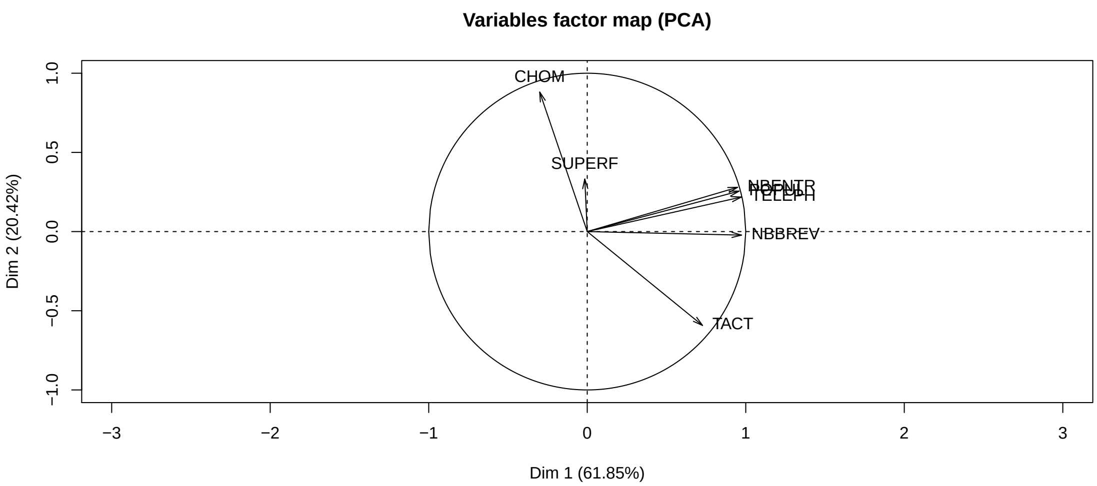



Figure 1 – Graphique des variables sur le premier plan principal

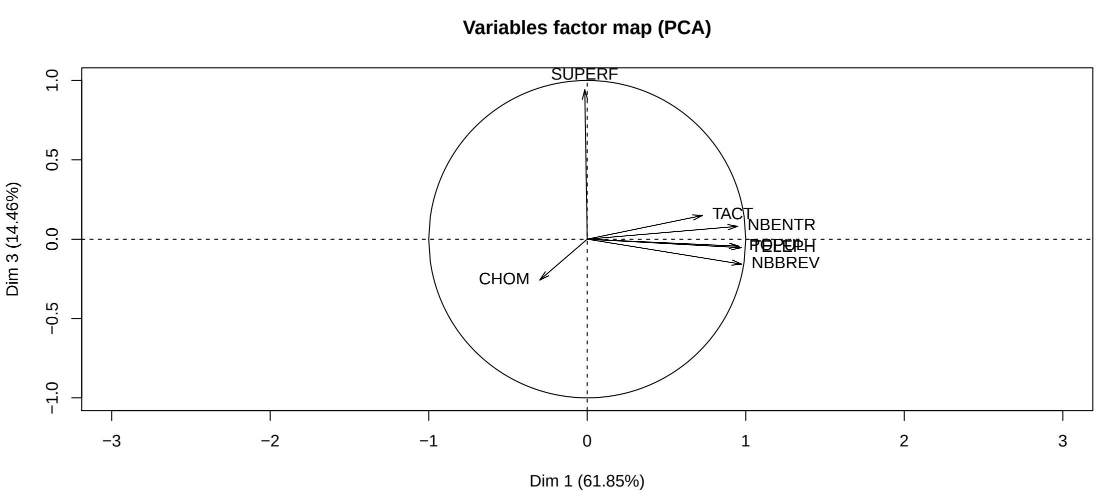



Figure 2 – Graphique des variables sur le plan (C1,C3)

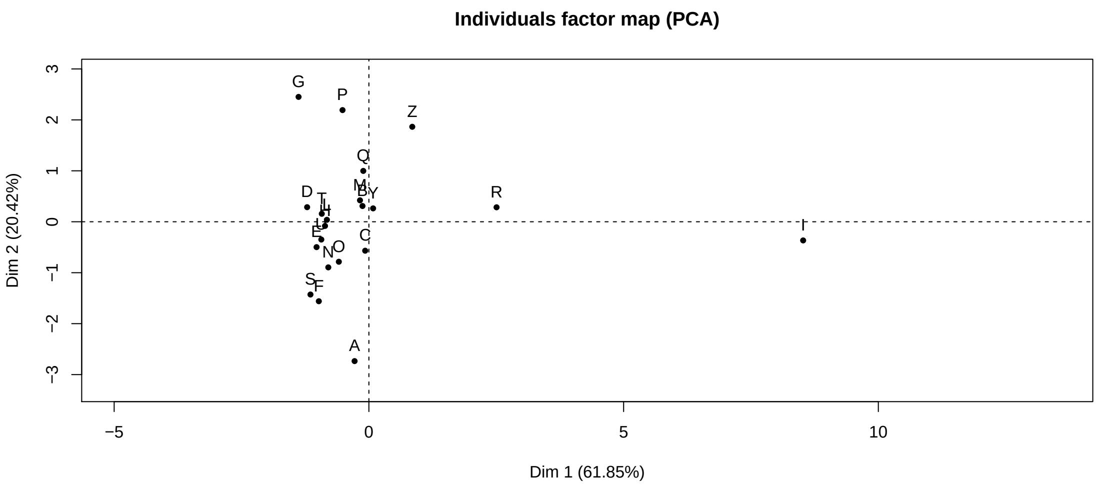



FIGURE 3 – Graphique des individus sur le premier plan principal

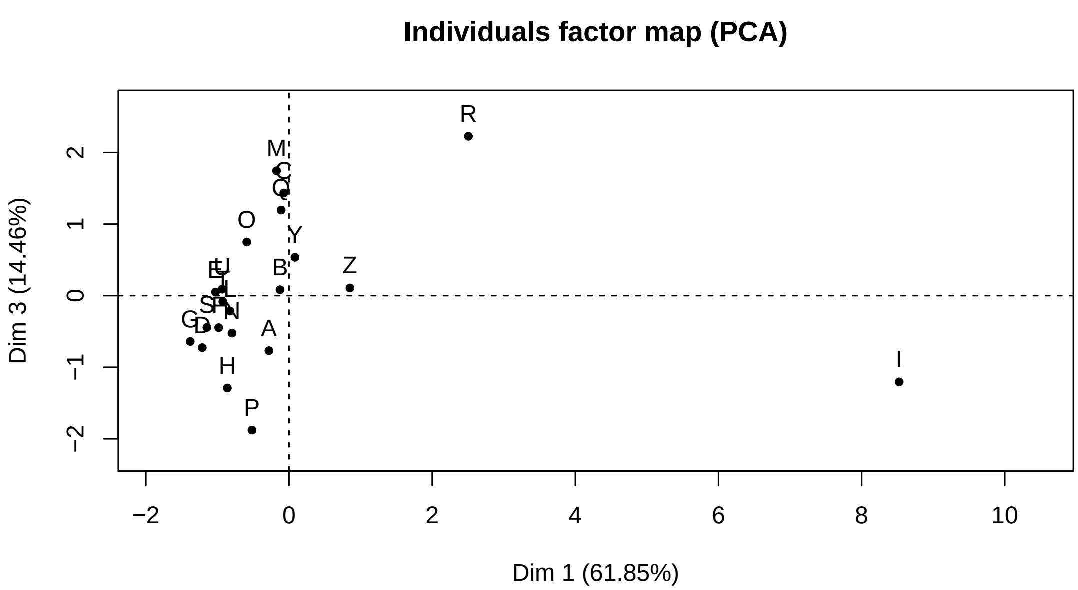



FIGURE 4 – Graphique des individus sur le plan (C1,C3)

## SORTIES SAS : ACP sur le fichier des régions

### Les données

| Obs. | 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 |
| **7** | C | Centre | 2370 | 38.78 | 39151 | 56753 | 229 | 7.9 | 1100 |
| **8** | E | Champ-Ar | 1340 | 37.85 | 25606 | 24060 | 155 | 9.3 | 550 |
| **9** | F | Fr-Comte | 1090 | 37.27 | 16202 | 27481 | 159 | 7.1 | 450 |
| **10** | H | Hte-Norm | 1730 | 37.80 | 12317 | 37461 | 181 | 10.8 | 750 |
| **11** | I | Ile-de-F | 10660 | 46.04 | 12012 | 273604 | 6722 | 7.3 | 5800 |
| **12** | G | Lang-Rou | 2110 | 32.12 | 27376 | 62202 | 179 | 13.2 | 1000 |
| **13** | S | Limousin | 720 | 38.06 | 16942 | 21721 | 73 | 7.9 | 350 |
| **14** | L | Lorraine | 2300 | 34.34 | 23547 | 48353 | 185 | 8.6 | 950 |
| **15** | M | Midi-Pyr | 2430 | 37.14 | 45348 | 78771 | 237 | 9.0 | 1100 |
| **16** | P | Nord-PdC | 3960 | 32.05 | 12414 | 78504 | 278 | 12.6 | 1600 |
| **17** | Y | Pays-Loi | 3060 | 37.93 | 32082 | 72027 | 339 | 9.6 | 1300 |
| **18** | D | Picardie | 1810 | 34.39 | 19399 | 36285 | 139 | 9.8 | 750 |
| **19** | T | Poit-Cha | 1590 | 36.82 | 25809 | 44598 | 133 | 10.1 | 750 |
| **20** | Z | Pr-Cte-A | 4260 | 34.96 | 31400 | 132552 | 610 | 11.0 | 2300 |
| **21** | R | Rh-Alpes | 5350 | 39.44 | 48698 | 159634 | 1474 | 7.4 | 2500 |

Procédure PRINCOMP

| | |
|---|---|
| **Observations** | 21 |
| **Variables** | 7 |

**Simple Statistics**

| | POPUL | TACT | SUPERF | NBENTR | NBBREV | CHOM | TELEPH |
|---|---|---|---|---|---|---|---|
| **Mean** | 2681.142857 | 37.22571429 | 25727.76190 | 69827.23810 | 587.095238 | 9.185714286 | 1261.904762 |
| **StD** | 2151.172454 | 2.90653672 | 11348.95478 | 58161.01681 | 1436.473700 | 1.845070653 | 1178.548340 |

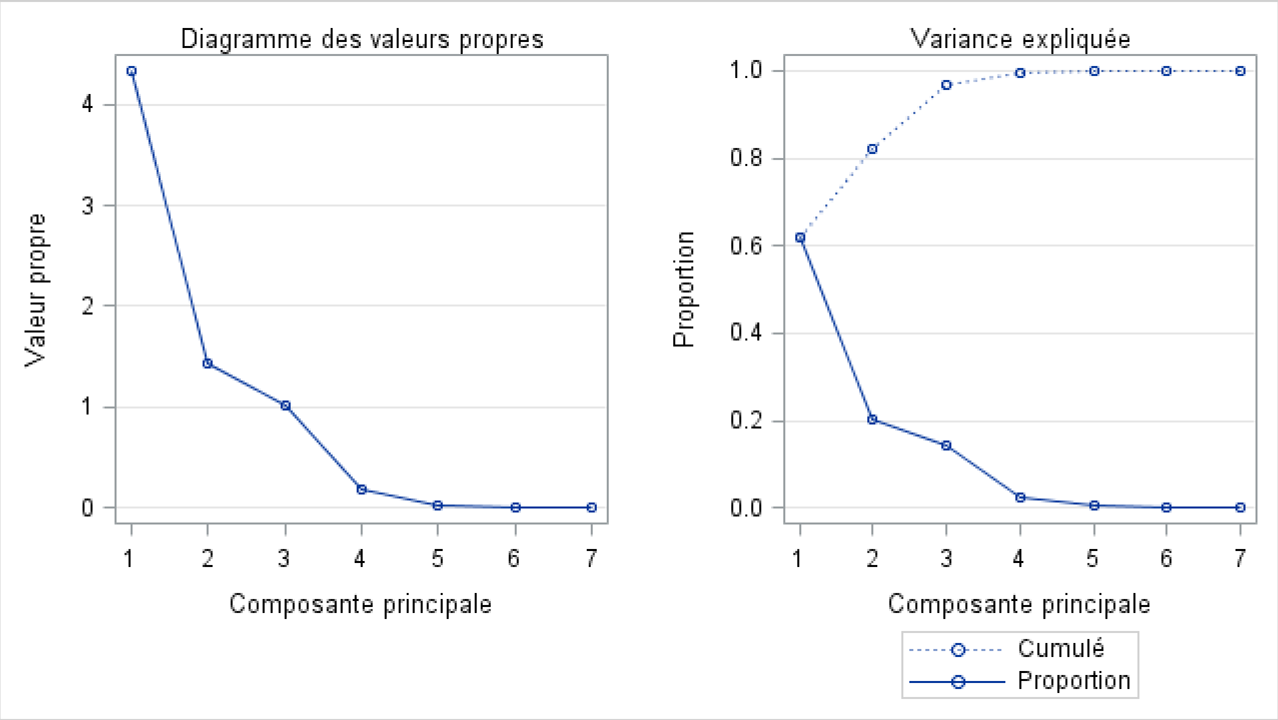
**Correlation Matrix**

| | POPUL | TACT | SUPERF | NBENTR | NBBREV | CHOM | TELEPH |
|---|---|---|---|---|---|---|---|
| **POPUL** | 1.0000 | 0.5138 | 0.0244 | 0.9810 | 0.9214 | -.0731 | 0.9939 |
| **TACT** | 0.5138 | 1.0000 | -.0593 | 0.5157 | 0.7085 | -.6985 | 0.5553 |
| **SUPERF** | 0.0244 | -.0593 | 1.0000 | 0.1493 | -.1640 | 0.0621 | 0.0048 |
| **NBENTR** | 0.9810 | 0.5157 | 0.1493 | 1.0000 | 0.8916 | -.0780 | 0.9829 |
| **NBBREV** | 0.9214 | 0.7085 | -.1640 | 0.8916 | 1.0000 | -.2566 | 0.9444 |
| **CHOM** | -.0731 | -.6985 | 0.0621 | -.0780 | -.2566 | 1.0000 | -.0983 |
| **TELEPH** | 0.9939 | 0.5553 | 0.0048 | 0.9829 | 0.9444 | -.0983 | 1.0000 |

**Eigenvalues of the Correlation Matrix**

| | Valeur propre | Différence | Proportion | Cumulé |
|---|---|---|---|---|
| **1** | 4.32967589 | 2.90029372 | 0.6185 | 0.6185 |
| **2** | 1.42938216 | 0.41694538 | 0.2042 | 0.8227 |
| **3** | 1.01243678 | 0.82967105 | 0.1446 | 0.9674 |
| **4** | 0.18276574 | 0.15000942 | 0.0261 | 0.9935 |
| **5** | 0.03275632 | 0.02203572 | 0.0047 | 0.9981 |
| **6** | 0.01072060 | 0.00845809 | 0.0015 | 0.9997 |
| **7** | 0.00226251 | | 0.0003 | 1.0000 |

| | Eigenvectors | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** |
| **POPUL** | 0.460322 | 0.213687 | -.044160 | -.207781 | -.233027 | 0.703738 | 0.384408 |
| **TACT** | 0.349483 | -.495690 | 0.148384 | 0.719834 | -.300519 | 0.040055 | 0.007678 |
| **SUPERF** | -.007459 | 0.277585 | 0.936223 | 0.074837 | 0.184603 | 0.077282 | -.026946 |
| **NBENTR** | 0.456005 | 0.233990 | 0.080407 | -.178842 | -.329468 | -.699630 | 0.317532 |
| **NBBREV** | 0.467849 | -.018723 | -.156560 | 0.144916 | 0.838511 | -.077938 | 0.161495 |
| **CHOM** | -.144107 | 0.737032 | -.256323 | 0.605428 | -.058670 | 0.004026 | 0.017725 |
| **TELEPH** | 0.467247 | 0.182368 | -.053299 | -.116350 | -.078848 | 0.040043 | -.851014 |



| Obs. | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 | Prin7 |
|---|---|---|---|---|---|---|---|
| **1** | -0.27417 | -2.66900 | -0.74987 | -0.72205 | -0.21315 | -0.04008 | 0.00131 |
| **2** | -0.10754 | 0.97424 | 1.16718 | 0.18910 | -0.01331 | -0.03244 | 0.01250 |
| **3** | -0.91141 | -0.34041 | 0.08789 | 0.34316 | 0.06521 | -0.08437 | 0.02417 |
| **4** | -0.77767 | -0.87350 | -0.51032 | 0.47755 | -0.18486 | -0.04617 | 0.02741 |
| **5** | -0.57583 | -0.76589 | 0.73135 | 0.14625 | 0.12656 | 0.05063 | -0.04503 |
| **6** | -0.12283 | 0.30188 | 0.08063 | -0.09335 | -0.13039 | 0.00941 | -0.02050 |
| **7** | -0.07142 | -0.55424 | 1.39997 | 0.10168 | 0.00809 | 0.17943 | -0.09045 |
| **8** | -1.00257 | -0.48567 | 0.04950 | 0.48827 | 0.12978 | 0.11908 | -0.02101 |
| **9** | -0.95932 | -1.52218 | -0.43631 | -0.41537 | 0.12347 | -0.08430 | 0.02531 |
| **10** | -0.84072 | -0.07973 | -1.25906 | 0.78446 | -0.24527 | 0.00294 | 0.02618 |
| **11** | 8.31845 | -0.35729 | -1.17546 | 0.24735 | 0.18437 | 0.00423 | -0.01112 |
| **12** | -1.34728 | 2.39199 | -0.62486 | 0.12691 | 0.31143 | -0.13225 | 0.02071 |
| **13** | -1.11929 | -1.39352 | -0.43254 | 0.10231 | 0.05758 | -0.11712 | -0.00170 |
| **14** | -0.80434 | 0.03755 | -0.20979 | -0.82817 | 0.23060 | 0.08895 | -0.01340 |
| **15** | -0.17053 | 0.41087 | 1.70298 | 0.02463 | 0.11692 | -0.04423 | 0.03291 |
| **16** | -0.50519 | 2.13920 | -1.83272 | -0.46404 | -0.18072 | 0.18771 | 0.04775 |
| **17** | 0.08077 | 0.25641 | 0.52317 | 0.28012 | -0.18350 | 0.16611 | 0.01506 |
| **18** | -1.18307 | 0.27935 | -0.70869 | -0.34984 | 0.22778 | 0.04458 | -0.00610 |
| **19** | -0.90239 | 0.15322 | -0.08086 | 0.38776 | 0.04449 | -0.04927 | -0.00662 |
| **20** | 0.83080 | 1.81940 | 0.10507 | -0.37395 | -0.31360 | -0.19263 | -0.12446 |
| **21** | 2.44555 | 0.27732 | 2.17275 | -0.45278 | -0.16149 | -0.03023 | 0.10707 |

Procédure CORR

| | |
|---|---|
| **7 Avec les variables :** | POPUL TACT SUPERF NBENTR NBBREV CHOM TELEPH |
| **3 Variables :** | Prin1 Prin2 Prin3 |

| Statistiques simples | | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | **N** | **Moyenne** | **Ecart-type** | **Somme** | **Minimum** | **Maximum** |
| **POPUL** | 21 | 2681 | 2151 | 56304 | 720.00000 | 10660 |
| **TACT** | 21 | 37.22571 | 2.90654 | 781.74000 | 32.05000 | 46.04000 |
| **SUPERF** | 21 | 25728 | 11349 | 540283 | 8280 | 48698 |
| **NBENTR** | 21 | 69827 | 58161 | 1466372 | 21721 | 273604 |
| **NBBREV** | 21 | 587.09524 | 1436 | 12329 | 73.00000 | 6722 |
| **CHOM** | 21 | 9.18571 | 1.84507 | 192.90000 | 5.20000 | 13.20000 |
| **TELEPH** | 21 | 1262 | 1179 | 26500 | 350.00000 | 5800 |
| **Prin1** | 21 | 0 | 2.08079 | 0 | -1.34728 | 8.31845 |
| **Prin2** | 21 | 0 | 1.19557 | 0 | -2.66900 | 2.39199 |
| **Prin3** | 21 | 0 | 1.00620 | 0 | -1.83272 | 2.17275 |

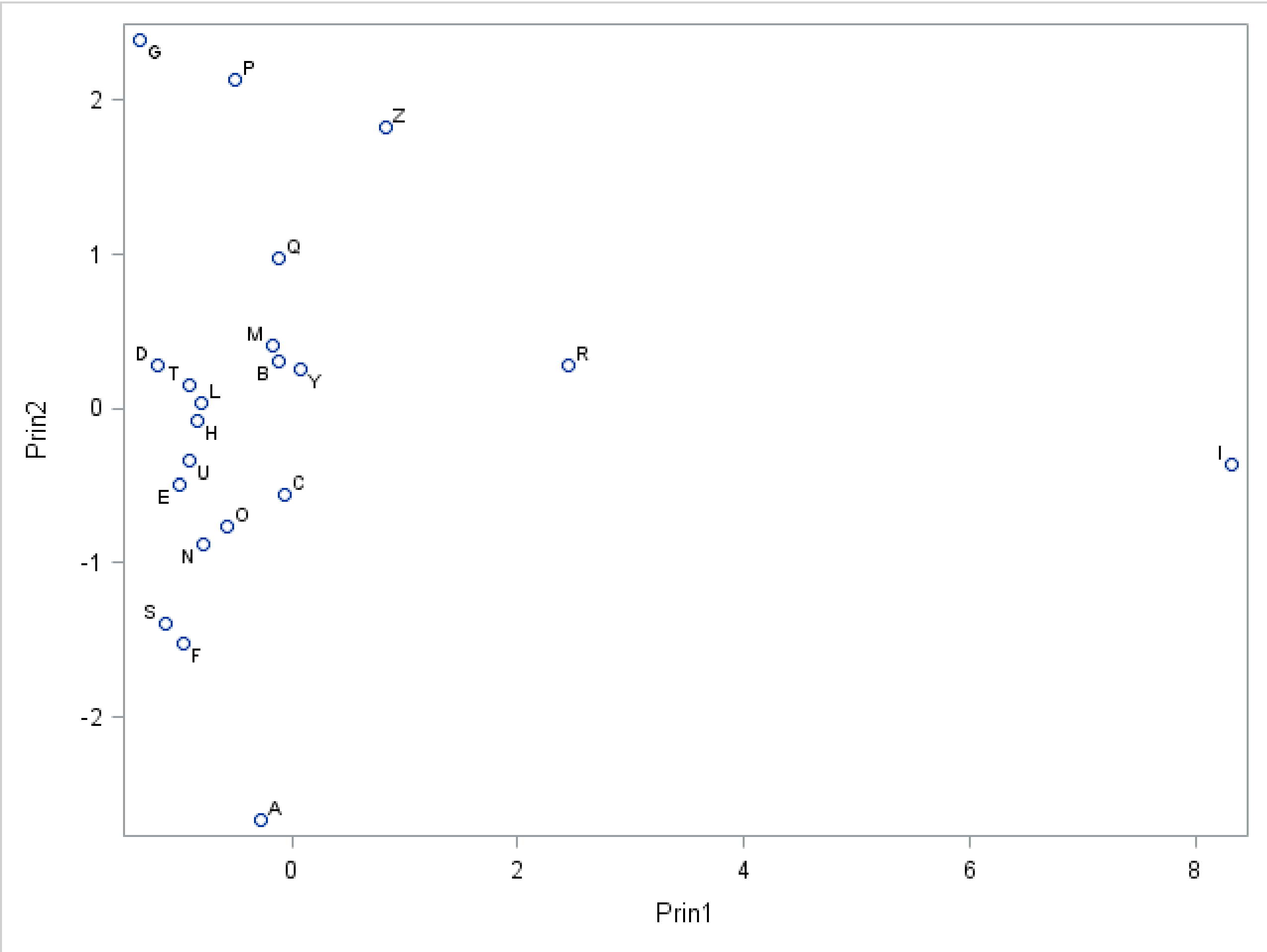
| Coefficients de corrélation de Pearson, N = 21<br>Proba > \|r\| sous H0: Rho=0 | | | |
|---|---|---|---|
| | **Prin1** | **Prin2** | **Prin3** |
| **POPUL** | 0.95783 | 0.25548 | -0.04443 |
| | <.0001 | 0.2637 | 0.8483 |
| **TACT** | 0.72720 | -0.59263 | 0.14930 |
| | 0.0002 | 0.0046 | 0.5183 |
| **SUPERF** | -0.01552 | 0.33187 | 0.94203 |
| | 0.9468 | 0.1416 | <.0001 |
| **NBENTR** | 0.94885 | 0.27975 | 0.08091 |
| | <.0001 | 0.2194 | 0.7274 |
| **NBBREV** | 0.97349 | -0.02238 | -0.15753 |
| | <.0001 | 0.9233 | 0.4953 |
| **CHOM** | -0.29986 | 0.88117 | -0.25791 |

**Coefficients de corrélation de Pearson, N = 21**
**Proba > |r| sous H0: Rho=0**

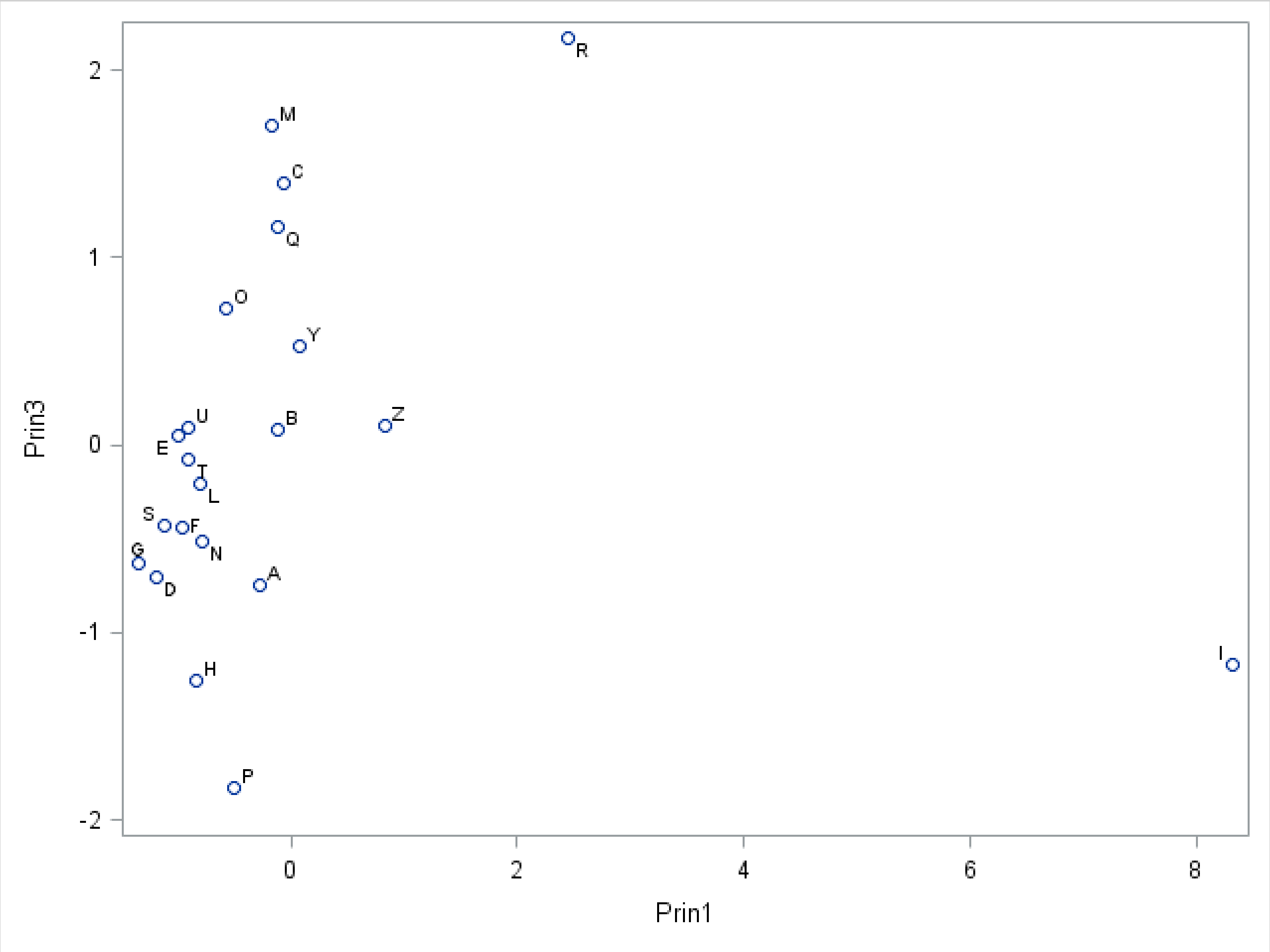
| | Prin1 | Prin2 | Prin3 |
|---|---|---|---|
| | 0.1866 | <.0001 | 0.2590 |
| **TELEPH** | 0.97224 | 0.21803 | -0.05363 |
| | <.0001 | 0.3424 | 0.8174 |

---

| Obs. | REGION | rap1carre | rap2carre | rap3carre |
|---|---|---|---|---|
| **1** | Alsace | 0.00902 | 0.85523 | 0.06751 |
| **2** | Aquitain | 0.00490 | 0.40215 | 0.57721 |
| **3** | Auvergne | 0.76631 | 0.10690 | 0.00713 |
| **4** | Bas-Norm | 0.31943 | 0.40300 | 0.13755 |
| **5** | Bourgogn | 0.22179 | 0.39235 | 0.35777 |
| **6** | Bretagne | 0.10859 | 0.65589 | 0.04679 |
| **7** | Centre | 0.00220 | 0.13224 | 0.84371 |
| **8** | Champ-Ar | 0.66419 | 0.15586 | 0.00162 |
| **9** | Fr-Comte | 0.25400 | 0.63949 | 0.05254 |
| **10** | Hte-Norm | 0.23761 | 0.00214 | 0.53292 |
| **11** | Ile-de-F | 0.97734 | 0.00180 | 0.01952 |
| **12** | Lang-Rou | 0.22526 | 0.71003 | 0.04845 |
| **13** | Limousin | 0.36747 | 0.56959 | 0.05488 |
| **14** | Lorraine | 0.44943 | 0.00098 | 0.03057 |
| **15** | Midi-Pyr | 0.00933 | 0.05419 | 0.93092 |
| **16** | Nord-PdC | 0.03011 | 0.53991 | 0.39629 |
| **17** | Pays-Loi | 0.01342 | 0.13530 | 0.56326 |
| **18** | Picardie | 0.64912 | 0.03619 | 0.23293 |
| **19** | Poit-Cha | 0.81501 | 0.02350 | 0.00654 |
| **20** | Pr-Cte-A | 0.16044 | 0.76941 | 0.00257 |
| **21** | Rh-Alpes | 0.54262 | 0.00698 | 0.42831 |



Prin2
2
1
0
-1
-2
0
2
4
6
8
Prin1
G
P
Z
Q
M
D
T
B
Y
R
L
H
I
U
C
E
O
N
S
F
A




**CODE SAS et COMMENTAIRES**

```
/*ACP sur le fichier des régions*/
proc print data=anadon.regions;
run;
/* Vérifier que l'ACP est justifiée vérifiant que certaines variables sont
bien corrélées entre elles*/
proc princomp data=anadon.regions out=anadon.comp;
var popul tact superf nbentr nbbrev chom teleph;
run;
/*L'ACP crée autant de composantes principales que de variables initiales
donc ici p=7 mais on veut éliminer les dernières c.p.
(celles qui apportent très peu d'information)
 2 critères pour déterminer le nombre de composantes principales à retenir
:
- Critère de la part d'inertie expliquée : si on retient C1, C2 et C3 on
explique 94,76% de l'inertie (on perd très peu d'information)
En général , on cherche à conserver au moins 80% de l'inertie. Donc dans
cet exemple, on retiendrait C1 et C2. Mais
dans cet exemple retenir 3 axes permet de bien réduire la dimension du
tableau de données tout en perdant très peu d'information).
-Critère de Kaiser : on rtielt les cc.p. de variance >1 donc celles de
valeur propre >1. On retient donc encore C1, C2 et C3.
Remarque: les 2 critères ne fournissent pas tout le temps le même nombre de
c.p. à retenir. Il faut alors arbitrer entre les 2 critères.*/
/*Tableau des composantes principales*/
```

```
proc print data=anadon.comp (keep=prin1 prin2 prin3 prin4 prin5 prin6
prin7);
run;
/*Corrélations entre C1, C2, C3 et les variables de départ pour interpréter
les composantes principales*/
proc corr data=anadon.comp;
var prin1 prin2 prin3;
with popul tact superf nbentr nbbrev chom teleph;
run;

/*C1 est fortement corrélée positivememnt avec POPUL, NBENTR, NBBREV et
TELEPH et (TACT)
C1>>0 équivaut à POPUL, NBENTR, NBBREV et TELEPH élevés
C1>>0 équivaut à POPUL, NBENTR, NBBREV et TELEPH faibles
C1 autour de 0 équivaut à POPUL, NBENTR, NBBREV et TELEPH moyens
C1 mesure le potentiel de développement économique de la région
C2 est fortement corrélé positivement avec CHOM et négativement avec TACT
C2>>0 équivaut à CHOM élévé et TACT faible
C2<<0 équivaut à CHOM faible et TACT élevé
C2 autour de 0 équivaut à CHOM moyen et TACT moyen
C2 mesure le dynamisme économique de la region (attention, un région
économiquement dynamique a un C2<<0)
C3 est très corrélé positivement à SUPERF.
C3>>0 équivaut à SUPERF élevé etc...
C3 est une mesure de la superficie.
/* Avant de commenter le graphique des régions, il faut étudier la qualité
de représentation des régions, car pour un axe donné, on ne peut commenter
que les régions bien représentées sur cet axe (celles pour lesquelles on
peut faire confiance à la c.p. comme résumé de
leurs caractéristiques initiales*/


/*Qualité de représentation des régions*/
data anadon.qualite (keep = region rap1carre rap2carre rap3carre);
set anadon.comp;
norminitcarre=prin1**2+prin2**2+prin3**2+prin4**2+prin5**2+prin6**2+prin7**
2;
rap1carre=prin1**2/norminitcarre;
rap2carre=prin2**2/norminitcarre;
rap3carre=prin3**2/norminitcarre;
run;
proc print;
run;
data anadon.qualiteaxe1;
set anadon.qualite;
if rap1carre>0.5;
run;
/*Règle : une région est bien représentée sur C1 si rap1>0.5
Donc les régions bien représentées sur C1 sont: Auvergne, Champagne-
Ardennes,
Ile de France, Picardie, Poitou-Charentes et Rhône-Alpes.*/
data anadon.qualiteaxe2;
set anadon.qualite;
if rap2carre>0.25;
run;
/*Règle : une région est bien représentée sur C2 si rap2>0.25
Donc les régions bien représentées sur C2 sont: Alsace, Aquitaine, Basse-
Normandie, Bourgogne,
Bretagne, Franche-Comté, Languedoc-Roussillon, Limousin, Nord-Pas-de-Calais
et Provence-Côte d'Azur.*/
data anadon.qualiteaxe3;
```

```
set anadon.qualite;
if rap3carre>0.15;
run;

/*Règle : une région est bien représentée sur C3 si rap3>0.15
Donc les régions bien représentées sur C3 sont: Aquitaine,
Bourgogne, Centre, Haute-Normandie, Midi-Pyrénées, Nord-pas de Calais,
Pays de Loire, Picardie et Rhône-Alpes.

/*Graphiques des régions*/
proc plot data=anadon.comp;
plot prin2*prin1=nom /hpos=40 vpos=20;
run;
quit;


%plotit(data=anadon.comp,labelvar=region,
           plotvars=prin2 prin1, color=black);
   run;
   quit;
%plotit(data=anadon.comp,labelvar=region,
           plotvars=prin3 prin1, color=black);
   run;
   quit;
   proc sgplot data=anadon.comp;
scatter x= prin1  y=prin2 /datalabel = nom;
run;
quit;
 proc sgplot data=anadon.comp;
scatter x= prin1  y=prin3 /datalabel = nom;
run;
quit;

/*Commentaires des graphiques
Pour un axe donné, on commente uniquement les régions bien représentées sur
cet axe.
- Axe 1 (C1): U, E, I, D, T, et R  peuvent être interprétés.
   On repère sur le graphique que l'Ile de France (I) est seule à droite
sur l'axe 1.
   Ce qui signifie qu'elle a C1 élevée et donc (voir question 3)
   que Idf se caractérise par : popul, nbbrev, nbentr et teleph : élevées
(puisque toutes les corr. Sont positives).
   La région Rhône-Alpes (R) est aussi relativement à droite sur le
graphique par rapport aux autres régions.
Donc l'idf  et Rhône-Alpes s'opposent à l'Auverge, la Bourgogne, la
Picardie et Poitou-Charentes
   grâce à  un développement démographique et économique important.
   De plus cette opposition est sur l'axe 1 c'est donc l'information
   la plus importante de ce fichier de données (inertie de l'axe 1 = 61,85%
de
   l'information contenue dans les données) ; l'avantage de l'ACP est non
seulement de résumer mais aussi de classer les résumés.
IDF est appelée individu «  atypique ». elle « se détache » des autres
régions
   (l'ACP permet aussi de repérer des individus atypiques ; on l'a déjà
étudié dans les boîtes à moustaches de la partie anadon. Des.)
- Axe 2 (C2): A, Q, N, O, B, F, G, S,  P et Z
 A, S, F sont opposées à G, P, Z et Q.
Retour question précédente pour savoir pourquoi elles sont opposées :
   A, S et F se caractérisent par un tact élevé et un chômage faible
```

alors que G, P, Z et Q se caractérisent par un chômage élevé et un taux d’activité faible.

Enfin, N, O et P sont des régions moyennement économiquement dynamiques. Cette information est la deuxième la plus importante, inertie =20.4%

- Axe 3 (C3) : P, H et D  sont opposées aux régions Q, M, R et C

P, H et D sont de  superficie faible alors que C, Q, M et R ont une superficie relativement élevée.

O et Y ont une superficie moyenne.