**Université de Thessalie**
Département d'Aménagement,
D'Urbanisme et Développement Régional

Enseignant : As. Pr. Marie-Noelle Duquenne

# I. Les Méthodes factorielles

La question traitée dans ce document porte sur les techniques de traitement de données multidimensionnelles. Ce sont les analyses factorielles qui permettent le plus efficacement de traiter un grand nombre de données, tant en termes d'individus considérés (n = nombre de lignes) que de variables caractérisant ces individus (p= nombre de colonnes).

L'analyse factorielle a un double objectif :
1. **résumer** le plus possible l'information contenue dans la matrice (n,p)
2. **hiérarchiser** l'information de sorte à perdre le moins possible de l'inertie initiale.

***Résumer*** signifie produire de nouvelles variables synthétiques (axes factoriels) qui sont fonction linéaire de plusieurs variables originales. Il s'agit donc d'un processus de réduction des dimensions. Si les variables initiales n'étaient pas du tout corrélées entre elles (variables totalement indépendantes) alors il ne serait possible de créer les axes factoriels. Ainsi, plus les variables initiales sont étroitement corrélées et plus il sera possible de procéder à une bonne condensation des données de base.

***Hiérarchiser*** signifie ordonner les axes factoriels selon leur plus ou moins grand degré de contribution è l'information, à l'inertie totale du nuage de points. Le premier axe est celui qui intègre le maximum d'information et qui présente la plus grande dispersion du nuage de points. Il est le meilleur résumé dans un espace à une dimension. Le deuxième axe est celui qui correspond au maximum d'information résiduelle, non contenue dans le 1$^{er}$ axe etc… Les deux premiers axes offrent donc la meilleure information possible dans le plan à deux dimensions.

Comme le souligne clairement Tabachnik et Fidell (1989), l'analyse factorielle répond à un certain nombre de questions :

- *Combien de facteurs (metadata) sont nécessaire pour donner une juste représentation des données, tout en réduisant le nombre de dimensions?*
- *Quelle proportion de la variance des données peut être expliquée par un certain nombre de dimensions (facteurs) majeures?*
- *Quelle est la signification de ces facteurs, comment peut-on les interpréter?*
- *Jusqu'à quel point la solution factorielle est conforme à la théorie que nous souhaitons vérifier?*
- *La structure factorielle reste-t-elle la même pour divers groupes?*

## 1. Analyse en Composantes Principales: ACP

L'analyse en Composantes Principales est une méthode d'analyse des données multi variées. Elle permet de décrire et d'explorer les relations qui existent entre plusieurs variables simultanément à la différence des méthodes bi variées qui étudient les relations potentielles entre deux variables. L'ACP cherche une solution pour refléter au mieux l'ensemble de la variance des variables mesurées, de sorte que les composantes soient orthogonales, c'est-à-dire indépendantes entre elles.

La procédure s'appuie sur un tableau R comprenant n individus et p variables ($R_{i,j}$). Le but est de trouver un nombre plus réduit de variables pour décrire efficacement les phénomènes structurant d'un groupe de données.

En d'autres termes, nous cherchons les corrélations existant entre les p variables initiales pour rapprocher au sein de "composantes" (facteurs), les variables les plus proches entre elles. On regroupe donc les variables pour qu'elles composent des dimensions dans le but de réduire le nombre de caractéristiques décrivant les individus afin de mieux interpréter les données. Les composantes (facteurs) sont en quelque sorte des méta-variables.

En pratique, chaque facteur qui correspond à une dimension du phénomène est défini par la meilleure combinaison linéaire de variables initiales expliquant la variance non expliquée par la dimension précédente. La première dimension (1er facteur) est celle qui correspond à la plus forte combinaison linéaire de certaines des variables initiales. On entend par plus forte combinaison, celle qui contribue le plus (en pourcentage) à l'explication de la variance totale, c'est-à-dire de l'inertie totale (information totale).
La deuxième dimension (2eme facteur) est celle qui correspond à 2ème combinaison linéaire de variables qui explique cette fois le meilleur pourcentage de l'information non expliquée par la 1ère dimension (variance résiduelle), etc.

### 1.1. La méthode ACP en quelques mots

1. Construction de la matrice initiale : R = $[r_{ij}]$, i = 1…n (individus) j=1…p (variables)

2. Calcul de la matrice : X = $[x_{ij}]$, matrice des données transformées par centrage et réduction et multiplication par la constante $1/\sqrt{n}$ .

$$x_{i,j} = \frac{r_{i,j} - \bar{r}_j}{\sigma_j} . \frac{1}{\sqrt{n}}$$

Cette transformation permet d'éviter des distorsions dans les représentations, ce qui arrive lorsque les écarts-types entre les variables sont très différents, notamment lorsque les variables sont exprimées dans des unités très différentes (âge, revenu, etc). Cette transformation réduit donc l'effet des variables très dispersées sur la distance entre individus. Nous procédons donc à une **ACP normée**.

3. Calcul de ma matrice C =$X^t$.X = matrice des corrélations linéaires entre les variables deux à deux. Cette matrice est symétrique de dimension (p,p) et sa diagonale est formée de 1. La somme des éléments de la diagonale = trace de C = p (nombre de variables initiales).

4. Détermination des axes factoriels : pour cela, nous sommes amenés à calculer les valeurs propres (eigenvalues) et les vecteurs propres associés à la matrice de corrélation C.
   Les valeurs propres λ sont obtenues en **diagonalisant la matrice des corrélations**. Diagonaliser la matrice C signifie calculer le vecteur λ tel que:
   | C- λ.I | = 0
   avec I = matrice unitaire prenant les valeurs 1 sur la diagonale et 0 ailleurs.
   Si nous avons p variables initiales, la matrice | C- λ.I | est de dimension (p,p) et **nous obtenons alors p valeurs propres** : $\lambda_i$ avec i=1,…p
   A la plus grande des valeurs propres est associé le 1er axe, à la seconde valeur propre est associé le 2ème axe etc.
   Les vecteurs propres sont obtenus grâce au calcul du système d'équations :
   (C- λ.I) . U = 0
   A chaque valeur propre est donc associé un vecteur propre.

5. **Pourcentage d'inertie** : l'inertie totale qui mesure la dispersion du nuage est égale à la trace de la matrice de corrélation. Comme tous les termes de la diagonale de cette matrice sont nécessairement égaux à 1, il en résulte que la trace de la matrice C correspond au nombre de variables initiales.

   **➔ Inertie totale = tr C = p** **[1]**

   Si nous voulons effectuer une analyse factorielle sur la base de 5 variables initiales, nous avons donc une inertie totale = 5.
   Par ailleurs, on montre également que :

   **➔ tr C =** $\sum_j \lambda_j$ **[2]**

   A partir des équations [1] et [2], il en résulte que :

   **Inertie totale =** $\sum_j \lambda_j$ **= p**

   De ce fait, nous pouvons calculer les pourcentages d'inertie associés à chaque axe. Ces pourcentages indiquent la part de l'inertie totale du nuage que restitue chaque axe factoriel, c'est-à-dire la contribution de chaque axe factoriel dans l'inertie totale.. Ils sont définis par :

   Axe j avec j = 1…p, $p_j = \frac{\lambda_j}{tr\ C} = \frac{\lambda_j}{p}$

   Par définition le taux d'inertie le plus élevé est attaché au premier axe factoriel et ainsi de suite jusqu'au dernier axe.

6. **Sélection des axes** : Plusieurs critères peuvent être utilisés.
   - Critère de Kaiser : on admet que seules les valeurs propres > 1 doivent être retenues. Cela correspond au coude que l'on observe sur la carte factorielle (changement de pente).
   - On se fixe pour chaque axe, une proportion minimale d'explication de la variance totale (inertie) : $\lambda_i$ / p > X%. Mais quel pourcentage minimum retenir ??
   - On se fixe un pourcentage cumulé d'inertie totale à atteindre afin de ne pas perdre trop d'information. Quel seuil retenir ?? On admet souvent qu'il est nécessaire de retenir au moins 75% de l'inertie totale (environ). Il ne s'agit que d'un ordre de grandeur et non pas d'un critère strict et cela dépend également de la complexité des phénomènes étudiés.

7. De là, il est possible de calculer les **coordonnées factorielles des variables sur les axes** afin de pouvoir déterminer la contribution relative de chacune des variables initiales à la formation de chacun des axes.
   La coordonnée de la variable initiale j sur l'axe i est donnée par la formule suivante : coord(j,i) = $(\lambda_i)^{1/2}$ . $U_{j,i}$

   Comme nous le verrons par la suite, ces coordonnées peuvent être directement lues dans les résultats fournis par SPSS dans le tableau intitulé ***Component Matrix***.

   Par ailleurs, on peut observer que pour chaque axe factoriel, la somme des carrés des coordonnées des variables initiales est égale à la valeur propre associée à l'axe.
   Si donc, nous avons p variables initiales, nous aurons au niveau de l'axe 1, la relation suivante:

   $\sum_{j=1}^{p} coord^2(j,1) = \lambda_1$ avec $\lambda_1$ = valeur propre associée au 1er axe

   Par conséquent, la **contribution absolue** de la variable j à la formation du premier axe n'est autre que :

   $$CTA(j,1) = \frac{coord^2(j,1)}{\lambda_1}$$

   **Remarque :**
   Les contributions absolues des variables d'une ACP normée ne sont pas souvent fournies par les logiciels (c'est d'ailleurs le cas du logiciel SPSS) car leur calcul n'est pas indispensable pour interpréter les axes. En effet, les coordonnées des variables initiales sur les axes – coordonnées fournies par le tableau Component Matrix – correspondent aux coefficients de corrélation entre chaque variable et chaque axe.

8. Selon un procédé identique, il est également possible de calculer les **coordonnées factorielles des individus sur les axes**. La projection de ces coordonnées dans les plans à deux dimensions permet de mettre en évidence les groupes d'individus qui présentent des comportements de même type, donc des distances peu éloignées.

## 1.2. Procédure ACP à l'aide du SPSS

### Premier Exemple : Les 22 régions de France

Considérons l'exemple suivant tiré de Giannelloni et Vernette (1994), Etudes de marché, Eds Vuibert. Cet exemple traite de certains indicateurs caractérisant les 22 régions de France. Il s'agit d'un exemple bien connu, présenté ici pour raisons pédagogiques mais en réalité, le nombre d'individus (régions) est sujet à critique (trop faible par rapport au nombre de variables).

***Données :***
population : Population en milliers
pop_active : Population active en pourcentage de la population totale
superficie : Superficie en km2
nbre_entreprises : Nombre d'entreprises = nombre d'unités
nbre_brevets : Nombre de brevets = nombre déposé au cours de l'année
Taux de chômage est en pourcentage
Nombre de lignes téléphoniques en milliers
Nos données sont quantitatives et pour cela, l'analyse en composantes principales est une méthode appropriée de réduction des dimensions afin de déterminer une structure sous-jacente caractérisant les diverses régions de France.

Données introduites dans SPSS
(fichier: **DB1_Regions_France.sav**)

Indicateurs economiques Regions de France - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

23 : Code

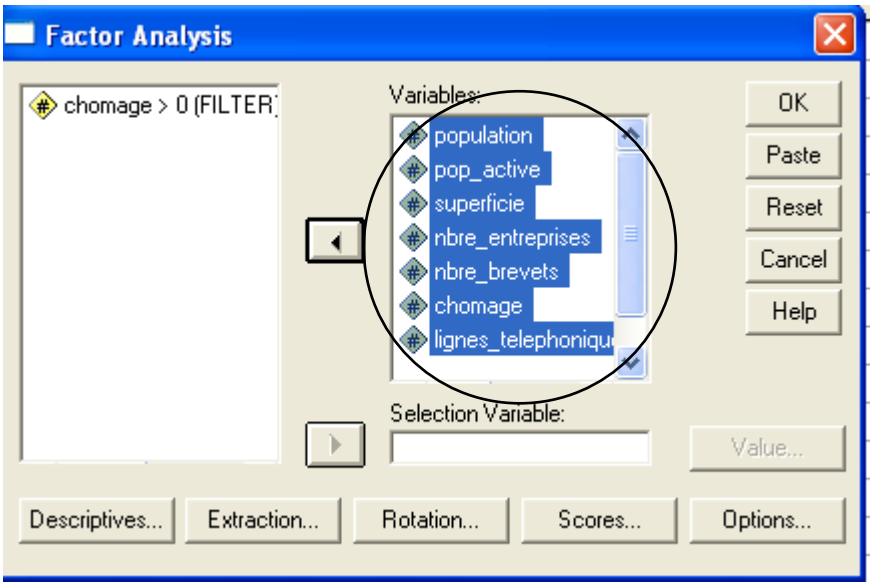
| | Code | region | population | pop_active | superficie | nbre_entreprises | nbre_brevets | chomage | lignes_telephoniques | var |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Alsace | 1624 | 39,14 | 8280 | 35976 | 241 | 5,20 | 700 | |
| 2 | 2 | Aquitaine | 2795 | 36,62 | 41308 | 85351 | 256 | 10,20 | 1300 | |
| 3 | 3 | Auvergne | 1320 | 37,48 | 26013 | 40494 | 129 | 9,30 | 600 | |
| 4 | 4 | Basse Normandie | 1390 | 38,63 | 17589 | 35888 | 91 | 9,00 | 600 | |
| 5 | 5 | Bourgogne | 1600 | 38,26 | 31582 | 40714 | 223 | 8,10 | 750 | |
| 6 | 6 | Bretagne | 2795 | 36,62 | 27208 | 73763 | 296 | 9,50 | 1300 | |
| 7 | 7 | Centre | 2370 | 38,78 | 39151 | 56753 | 229 | 7,90 | 110 | |
| 8 | 8 | Champagne | 1340 | 37,85 | 25606 | 24060 | 155 | 9,30 | 550 | |
| 9 | 9 | Corse | 240 | . | 8680 | 8273 | . | . | . | |
| 10 | 10 | Franche Comte | 1090 | 37,27 | 16202 | 27481 | 159 | 7,10 | 450 | |
| 11 | 11 | Haute Normandie | 1730 | 37,80 | 12317 | 37461 | 181 | 10,80 | 750 | |
| 12 | 12 | Ile de France | 10660 | 46,04 | 12012 | 273604 | 6722 | 7,30 | 5800 | |
| 13 | 13 | Languedoc Roussillon | 2110 | 32,12 | 27376 | 62202 | 179 | 13,20 | 1000 | |
| 14 | 14 | Limousin | 720 | 38,06 | 16942 | 21721 | 73 | 7,90 | 350 | |
| 15 | 15 | Lorraine | 2300 | 34,34 | 23547 | 48353 | 185 | 8,60 | 950 | |
| 16 | 16 | Midi-Pyrennees | 2430 | 37,14 | 45348 | 78771 | 237 | 9,00 | 1100 | |
| 17 | 17 | Nord Pas de Calais | 3960 | 32,05 | 12414 | 78504 | 278 | 12,60 | 1600 | |
| 18 | 18 | Pays de la Loire | 3060 | 37,93 | 32082 | 72027 | 339 | 9,60 | 1300 | |
| 19 | 19 | Picardie | 1810 | 34,39 | 19399 | 36285 | 139 | 9,80 | 750 | |
| 20 | 20 | Poitou Charentes | 1590 | 36,82 | 25809 | 44592 | 133 | 10,10 | 750 | |
| 21 | 21 | Provence Cote d Azur | 4260 | 34,96 | 31400 | 132552 | 610 | 11,00 | 2300 | |
| 22 | 22 | Rhine Alpes | 5350 | 39,44 | 48698 | 159634 | 1474 | 7,40 | 2500 | |
| 23 | | | | | | | | | | |

1ère étape : Examen des données
On observe qu'une région, la Corse, ne dispose pas de toutes les données par conséquent, il est nécessaire d'éliminer cette région de l'analyse car en cas contraire, cela nous amènerait à procéder à une analyse avec seulement 3 variables, donc nous perdrions une grande part de l'information et pas des moindres.

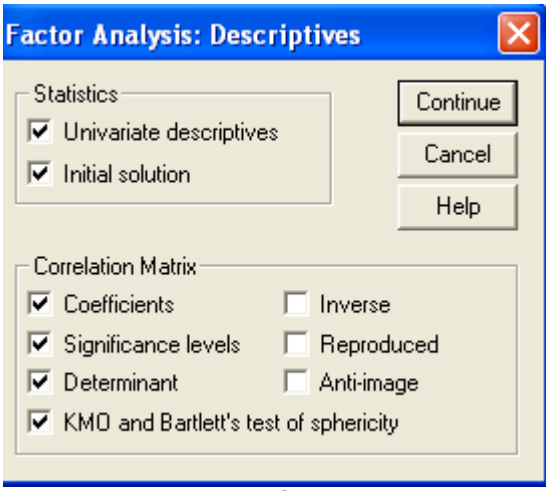
2ème étape : Réalisation de l'analyse

Sélectionner Analyze > Data Reduction > Factor

**1er** : Sélectionner toutes les variables sauf la dernière variable qui correspond au filtre de sélection utilisé pour ôter de l'analyse la région de Corse.




**2ème : Descriptives** : cette fenêtre nous donne la possibilité de calculer la moyenne et l'écart-type associés à chaque variable initiale, afin d'observer leur degré de dispersion (Univariate descriptives), la matrice des coefficients de **corrélation linéaire** entre les variables initiales , de même que certains tests.




Nous allons ainsi vérifier que les conditions nécessaires à la réalisation d'une ACP sont réunies :

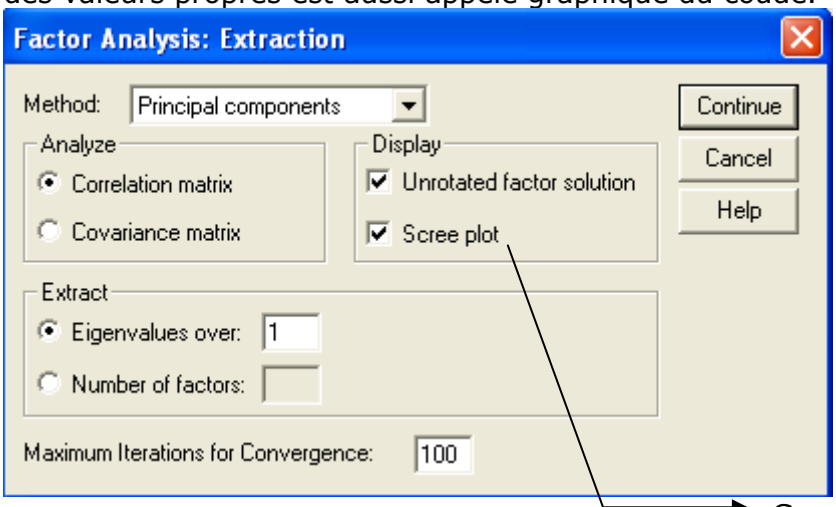
1. Les variables initiales doivent présenter un degré de variabilité suffisant. Un fort degré d'homogénéité (CV très faible) ne permettra pas à la variable initiale de contribuer efficacement à la formation des axes principaux.
2. Les variables initiales doivent être corrélées entre elles (en prenant comme risque $\alpha \leq 5\%$). On choisit également de calculer le déterminant de la matrice C pour vérifier que celui-ci est très petit : un déterminant très faible est un bon indice de l'existence de bonnes corrélations entre les variables. Il vient donc confirmer la lecture de la matrice des corrélations.

On choisit également de produire l'indice de KMO (Kaiser-Meyer-Olkin) et le test de sphéricité de Bartlett. **L'indice de KMO** est très utile car il permet d'évaluer dans quelle mesure l'ensemble des variables sélectionnées est un ensemble cohérent qui permet de définir une solution pertinente en termes conceptuels. Plus cet indice est élevé et plus la solution factorielle obtenue est satisfaisante. C. Durand (1997) propose la lecture suivante du KMO :

| Valeur du KMO | Solution |
|---|---|
| < 0,5 | Inacceptable |
| < 0,6 | Médiocre |
| < 0,7 | Moyenne |
| < 0,8 | méritoire |
| 0,9 | merveilleuse |

Le test de sphéricité de Bartlett est utilisé pour vérifier si toutes les corrélations sont ou non égales à zéro. On acceptera que toutes les corrélations ne soient pas égales à zéro, si l'indice de significativité (le risque) est inférieur à 5%. Mais ce test est très sensible au nombre de cas et il est presque toujours significatif lorsque l'on a un très grand nombre de cas. En cela, il perd bien souvent de son sens.

**3ème** : **Extraction** = Sélection du type de méthode de factorisation. Conserver la méthode par défaut qui est Principal components
Choisir aussi de produire le graphique des valeurs propres (scree plot) pour voir leur évolution et donner pour maximum d'itérations = 100. Le graphique des valeurs propres est aussi appelé graphique du coude.

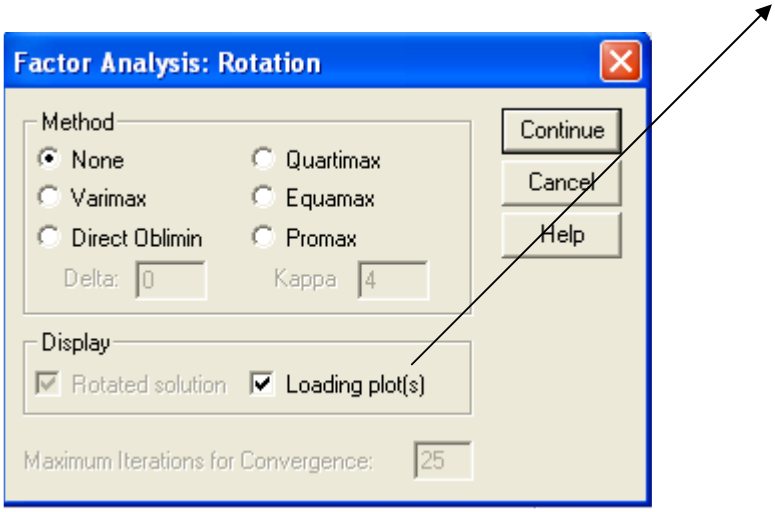



Graphique du coude

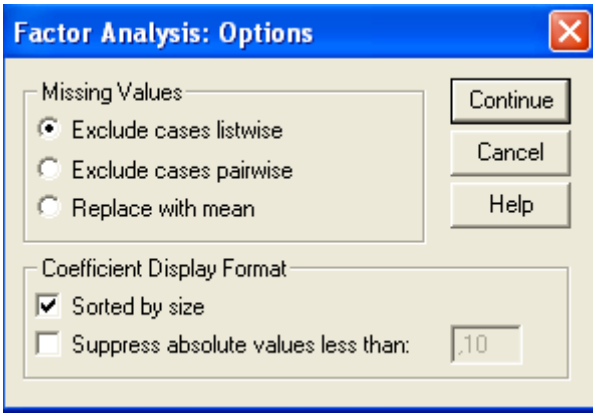
D'autres méthodes alternatives peuvent être employées mais nous verrons que pour de telles données socio-économiques quantitatives, la méthode en ACP est la plus robuste.

**4ème** : **Rotation.** Il est possible de procéder à une rotation des axes, lorsque la matrice des composantes [c'est-à-dire la matrice de définition des nouvelles dimensions (axes) = matrice de coordonnées des variables initiales sur les axes], ne permet pas de discerner clairement quelles variables composent quelles dimensions. On a souvent alors recours à une rotation des axes pour modifier les coordonnées des variables par rapport aux axes.
Dans un premier temps, on ne procède pas à une telle rotation, On commence toujours par l'analyse de base sans rotation.

Par contre, il faut toujours sélectionner la carte factorielle (loading plots),

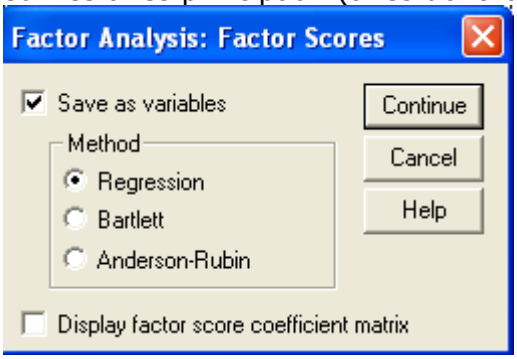



**5ème : Options.** Cette boite permet de sélectionner des options de présentation, pour classer les variables les plus importantes et cacher celles qui n'expliquent pas les dimensions. La lecture des tableaux de résultats est ainsi facilitée.




Présentation des coordonnées des variables initiales selon l'ordre de grandeur des coordonnées

**6ème : Scores** : Cette boîte permet de sauver les coordonnées des individus sur les axes principaux (axes dont la valeur propre est supérieure à 1).




Désormais, nous pouvons étudier les résultats de l'ACP pour voir dans quelle mesure, nous avons pu réduire les dimensions initiales et structurer les données selon certaines thématiques fondamentales.

Compte tenu des sélections précédentes, la procédure avec SPSS propose les tableaux et graphiques suivants:

## 1er tableau : étude de la dispersion des 7 variables initiales

**Descriptive Statistics**

| | Mean | Std. Deviation | Analysis N |
|---|---|---|---|
| Population | 2681,1 | 2151,2 | 21 |
| pop_active | 37,2 | 2,9 | 21 |
| Superficie | 25727,8 | 11349,0 | 21 |
| nbre_entreprises | 69818,4 | 58158,7 | 21 |
| nbre_brevets | 587,1 | 1436,5 | 21 |
| Chomage | 9,2 | 1,8 | 21 |
| lignes_telephoniques | 1214,8 | 1204,9 | 21 |

Toutes les variables initiales à l'exception de la variable population active ont une forte dispersion.

## 2ème Tableau : Coefficients de corrélation linéaire

**Correlation Matrix**

| | | population | pop_active | superficie | nbre_ entreprises | nbre_brevets | chomage | lignes_ telephoni ques |
|---|---|---|---|---|---|---|---|---|
| Correlation | population | 1,000 | ,514 | ,024 | ,981 | ,921 | -,073 | ,978 |
| | pop_active | ,514 | 1,000 | -,059 | ,516 | ,708 | -,699 | ,521 |
| | superficie | ,024 | -,059 | 1,000 | ,149 | -,164 | ,062 | -,044 |
| | nbre_entreprises | ,981 | ,516 | ,149 | 1,000 | ,892 | -,078 | ,971 |
| | nbre_brevets | ,921 | ,708 | -,164 | ,892 | 1,000 | -,257 | ,934 |
| | chomage | -,073 | -,699 | ,062 | -,078 | -,257 | 1,000 | -,068 |
| | lignes_telephoniques | ,978 | ,521 | -,044 | ,971 | ,934 | -,068 | 1,000 |
| Sig. (1-tailed) | population | | ,009 | ,458 | ,000 | ,000 | ,376 | ,000 |
| | pop_active | ,009 | | ,399 | ,008 | ,000 | ,000 | ,008 |
| | superficie | ,458 | ,399 | | ,259 | ,239 | ,395 | ,425 |
| | nbre_entreprises | ,000 | ,008 | ,259 | | ,000 | ,368 | ,000 |
| | nbre_brevets | ,000 | ,000 | ,239 | ,000 | | ,131 | ,000 |
| | chomage | ,376 | ,000 | ,395 | ,368 | ,131 | | ,386 |
| | lignes_telephoniques | ,000 | ,008 | ,425 | ,000 | ,000 | ,386 | |

a. Determinant = 5,206E-0

Il est souhaitable de calculer le déterminant de la matrice C pour vérifier que celui-ci est très petit car un déterminant très petit est un bon indice de l'existence de bonnes corrélations entre les variables. Il vient donc confirmer la lecture de la matrice des corrélations.

La plupart des variables sont significativement corrélées entre elles. Seule la variable superficie présente des coefficients non significatifs. Le déterminant de la matrice est égale à 0,000005206, soit une valeur excessivement petite, ce qui confirme bien l'existence de réelles corrélations entre les variables.

On choisit également de produire l'indice de KMO (Kaiser-Meyer-Olkin) et le test de sphéricité de Bartlett. L'indice de KMO est très utile car il permet d'évaluer dans quelle mesure l'ensemble des variables sélectionnées est un ensemble cohérent qui permet de définir une solution pertinente en termes conceptuels. Plus cet indice est élevé et plus la solution factorielle obtenue est satisfaisante.

C. Durand (1997) propose la lecture suivante du KMO :

| Valeur du KMO | Solution |
|---|---|
| < 0,5 | Inacceptable |
| < 0,6 | Médiocre |
| < 0,7 | Moyenne |
| < 0,8 | méritoire |
| 0,9 | merveilleuse |

Le test de sphéricité de Bartlett est utilisé pour vérifier si toutes les corrélations sont ou non égales à zéro. On acceptera que toutes les corrélations ne soient pas égales à zéro, si l'indice de significativité (le risque) est inférieur à 5%. Mais ce test est très sensible au nombre de cas et il est presque toujours significatif lorsque l'on a un très grand nombre de cas. En cela, il perd bien souvent de son sens.

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,70046 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 204,789 |
| | df | 21 |
| | Sig. | ,000 |

L'indice de KMO est de l'ordre de 0,7, la solution n'est ni mauvaise ni excellente.

### 3ème Tableau : La qualité de la représentation

Question : dans quelles mesures les variables initiales sont-elles prises en compte par les variables extraites?

**Communalities**

| | Initial | Extraction |
|---|---|---|
| population | 1,000 | ,981 |
| pop_active | 1,000 | ,903 |
| superficie | 1,000 | ,998 |
| nbre_entreprises | 1,000 | ,987 |
| nbre_brevets | 1,000 | ,972 |
| chomage | 1,000 | ,931 |
| lignes_telephoniques | 1,000 | ,985 |

Extraction Method: Principal Component Analysis.

Toutes les variables sont bien prises en compte, la qualité de leur représentation est de plus de 90%.
Cela veut dire que plus de 90% de la variance de chaque variable initiale est prise en compte par l'une des variables (facteurs) extraites

Communalities = % de variance expliquée dans les dimensions extraites.

## 4ème Tableau : La variance expliquée totale

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 4,299 | 61,409 | 61,409 | 4,299 | 61,409 | 61,409 |
| 2 | 1,437 | 20,532 | 81,941 | 1,437 | 20,532 | 81,941 |
| 3 | 1,020 | 14,578 | **96,519** | 1,020 | 14,578 | 96,519 |
| 4 | ,184 | 2,631 | 99,150 | | | |
| 5 | ,033 | ,475 | 99,625 | | | |
| 6 | ,019 | ,275 | 99,900 | | | |
| 7 | ,007 | ,100 | 100,000 | | | |

Extraction Method: Principal Component Analysis.

1ère dimension: $\lambda_1 = 4,299$, ainsi $\frac{\lambda_1}{p} = \frac{4,299}{7} = 0,641$. Le 1er axe explique 61,4% de l'inertie totale du phénomène.

2ème dimension explique 20,5%

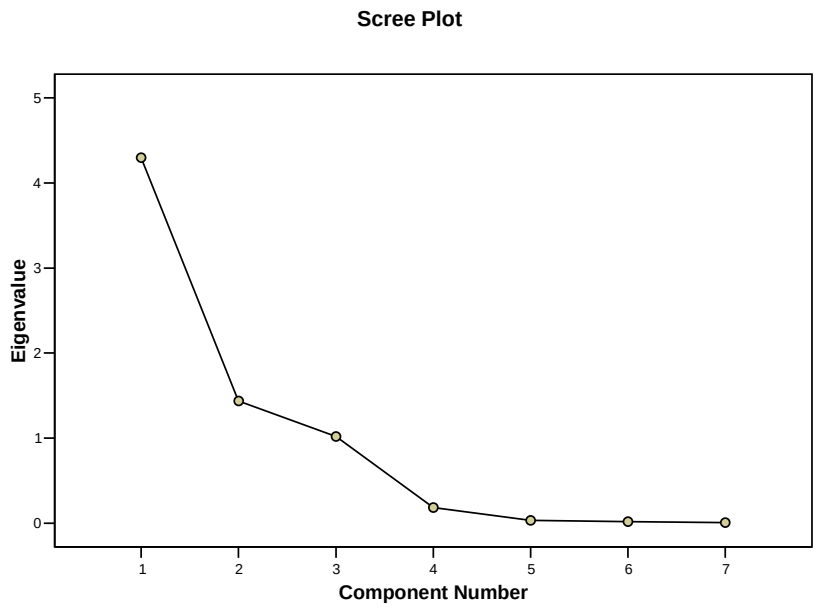
Le plan (axe 1 et axe 2) explique à lui seul 82% environ de l'inertie totale (variance totale).

Le 3ème axe est lui aussi significatif (sa valeur propre est supérieure à 1).

Vous pouvez également vérifier que la somme des valeurs propres est bien égale à 7 (nombre de variables).

Ainsi, à partir de 7 dimensions, nous avons réduit à 3 dimensions le phénomène.

### Graphique du coude :



Entre la 2ème et 3ème valeur propre, nous avons une faible pente, alors qu'entre la 3ème et 4ème, la pente est très prononcée. L'apport du 4ème axe à l'inertie totale est relativement peu importante et en cela, il peut être négligé.

### 5ème Tableau : La matrice des coordonnées

Elle nous donne les coefficients de corrélation de chaque variable initiale aux variables extraites, aux axes factoriels.

**Component Matrix[a]**

| | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| nbre_brevets | ,975 | -,022 | -,142 |
| lignes_telephoniques | ,960 | ,235 | -,088 |
| population | ,958 | ,251 | -,020 |
| nbre_entreprises | ,949 | ,273 | ,105 |
| pop_active | ,721 | -,602 | ,142 |
| chomage | -,293 | ,890 | -,231 |
| superficie | -,030 | ,300 | ,953 |

Extraction Method: Principal Component Analysis.

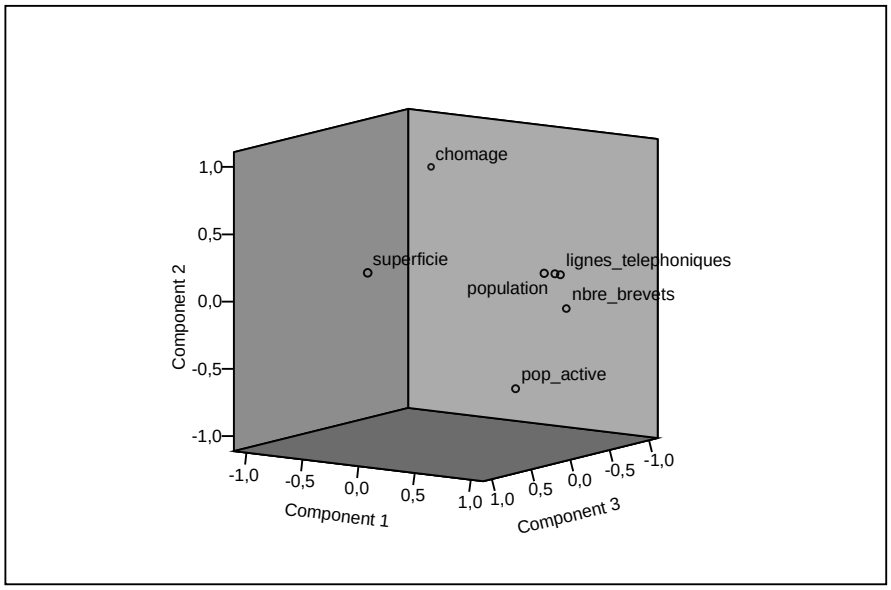
a. 3 components extracted.

Les 4 premières variables sont fortement corrélées avec l'axe 1 alors qu'elles ne le sont pas avec les deux autres axes.
Les deux variables suivantes sont fortement corrélées avec l'axe 2
La dernière variable n'est quant à elle fortement corrélée qu'avec l'axe 3.

**Toutes les variables initiales entrent dans le modèle. Cela est normal vu les résultats du premier tableau.**

L'axe 1 peut être nommé «axe de potentiel de développement» car ce sont bien des indicateurs de développement et d'industrialisation qui entrent dans la formation de cet axe.
L'axe 2 peut être nommé «axe de niveau d'activité» car tant la population active que le niveau de chômage expriment dans quelle mesure l'activité économique est en expansion ou non et il est parfaitement normal que les coefficients de ces 2 variables soient inversés.
L'axe 3 peut être nommé «axe de taille» des régions.

## Projection des variables sur les 3 axes factoriels




Nous pouvons vérifier notre interprétation des axes, en calculant les Contributions Absolues des 7 variables sur les 3 axes.

Pour cela, reprenons notre matrice des coordonnées sur les 3 axes principaux :

| | Axes | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| nbre_brevets | 0,975 | -0,022 | -0,142 |
| lignes_telephoniques | 0,960 | 0,235 | -0,088 |
| population | 0,958 | 0,251 | -0,020 |
| nbre_entreprises | 0,949 | 0,273 | 0,105 |
| pop_active | 0,721 | -0,602 | 0,142 |
| chomage | -0,293 | 0,890 | -0,231 |
| superficie | -0,030 | 0,300 | 0,953 |

Puis calculons le carré **de chacune de ces coordonnées : coord$^2$(j,i)**

| | Axes | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| nbre_brevets | 0,951 | 0,000 | 0,020 |
| lignes_telephoniques | 0,922 | 0,055 | 0,008 |
| population | 0,918 | 0,063 | 0,000 |
| nbre_entreprises | 0,901 | 0,074 | 0,011 |
| pop_active | 0,520 | 0,362 | 0,020 |
| chomage | 0,086 | 0,792 | 0,054 |
| superficie | 0,001 | 0,090 | 0,907 |
| **Somme** | **4,299** | **1,437** | **1,020** |

$0,951 = (0,975)^2$ de même $0,922 = (0,960)^2$

On vérifie bien également que la somme des carrés des coordonnées pour l'axe 1 est égale à $\lambda_1$ = 4,299. Il en va de même pour les deux autres axes.

Dans ces conditions, le tableau des Contributions Absolues (CTA) est donné par :

$$\frac{coord^2(j,i)}{\lambda_i}, \; j = 1...7 \; et \; i = 1,..3$$

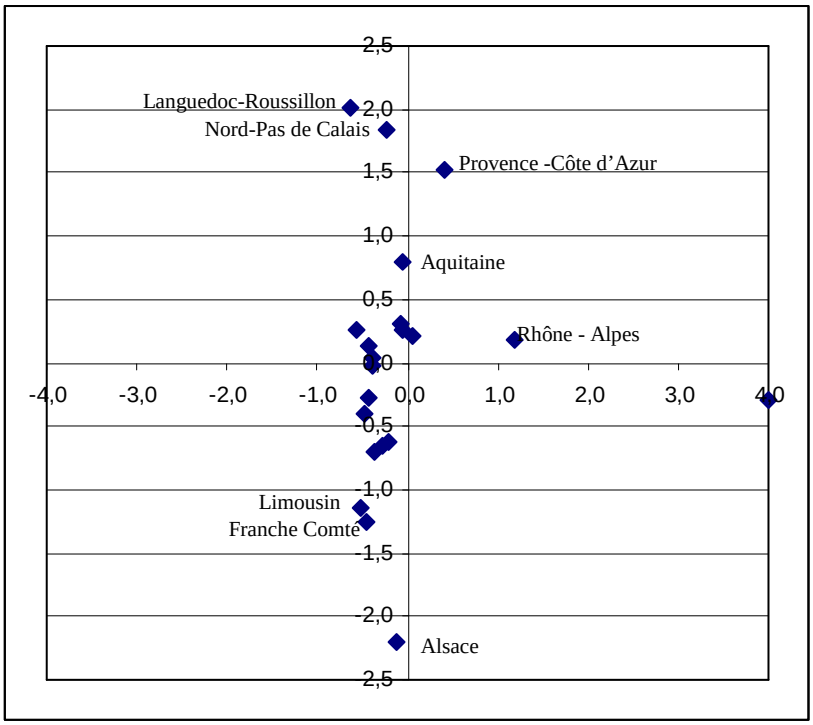
| | Axes | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| nbre_brevets | 0,221 | 0,000 | 0,020 |
| lignes_telephoniques | 0,214 | 0,038 | 0,008 |
| population | 0,214 | 0,044 | 0,000 |
| nbre_entreprises | 0,210 | 0,052 | 0,011 |
| pop_active | 0,121 | 0,252 | 0,020 |
| chomage | 0,020 | 0,551 | 0,052 |
| superficie | 0,000 | 0,062 | 0,889 |
| **Somme** | 1,000 | 1,000 | 1,000 |

0,221 = 0,951 / 4,299 : la variable nombre de brevets contribue à hauteur de 22% à la formation de l'axe 1. Les 4 premières variables expliquent près de 86% du premier axe factoriel.

Les deux variables population active et chômage expliquent à elles seules 80% du 2ème axe et la superficie 89% du 3ème axe.

Nos résultats précédents n'en sont que confortés.

**Projection des 21 régions sur le plan principal (facteur 1, facteur 2)**



On vérifie bien ici qu'au début des années 90, l'Ile de France (Région de Paris) et le Rhône-Alpes (Région de Lyon) se distinguent par leur niveau de développement. Trois autres régions sont marquées par un problème de niveau d'activité (chômage et faible pourcentage de population active). Il s'agit du Languedoc-Roussillon, Nord et Pas de Calais ainsi que de la Provence – Côte d'Azur) alors qu'au contraire l'Alsace est caractérisée par le plus faible taux de chômage.

## 2ème Exemple d'ACP : les entreprises de Thessalie

**Fichier : DB3_Thessalie_Entreprises**

L'étude porte sur 390 entreprises de Thessalie, réparties dans 5 communes différentes : Larissa, Volos, Farsala, Skopelos et Livadi. Cette enquête s'est tenue en 2003 et avait pour objectif de mettre en évidence le degré de dynamisme des entreprises étudiées ainsi que leur capacité à créer des postes de travail. Nous nous intéressions aux caractères internes de l'entreprise et non pas à sa position sur le marché.

Après un premier traitement des données provenant des questionnaires, 11 variables ont pu être retenues. Il s'agit brièvement de :

**Type d'entreprise** :

m1 : Degré de spécialisation du chef d'entreprise par rapport à l'activité dominante de celle-ci.
m3 : Emploi ou non de personnel (entreprise individuelle ou non)

**Fonctionnement de l'entreprise** :

m2 : Degré de valorisation de la force de travail, en pourcentage. Le chef d'entreprise a évalué lui- même sur une échelle de 0 à 100% dans quelle mesure, le personnel de l'entreprise – y compris lui- même – est employé à pleine capacité ou bien est sous-employé.
m4 : Nombre d'employés permanents
m5 : Degré de couverture des besoins de l'entreprise en personnel qualifie, nécessaire au bon fonctionnement de l'entreprise (sur une échelle de 0 à 10)
m6 : Recours ou non au travail informel
m10 : Pourcentage d'employés jeunes (18 à 40 ans)
m11 : Pourcentage de techniciens et ouvriers

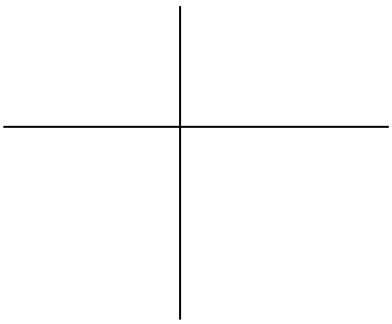
**Perspectives perçues par le chef d'entreprise (ou son comptable)**

m7 : Rentabilité économique sur une échelle de 0 à 10, telle que perçue par le propriétaire lui même
m8 : Perspectives pour le secteur, sur une échelle de –1 à 1
m9 : Perspectives pour l'entreprise, sur une échelle de –1 à 1

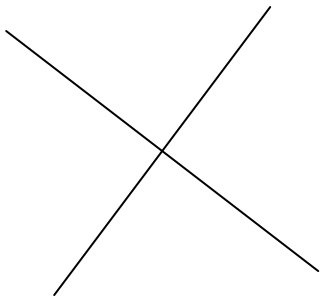
L'objectif du travail est de produire une ACP normée de façon à résumer efficacement les 11 variables ci-dessus présentées en un nombre limité de variables composites (axes principaux). Ces axes principaux nous permettront dans un second temps de procéder à une typologie des entreprises.

L'analyse que nous avons effectuée à l'aide de SPSS est la même que celle suivie lors de l'exemple précédent (**même procédure de choix des commandes**). Simplement, nous allons ajouter une commande de plus, concernant la rotation des axes. Nous demandons effectivement une rotation Varimax des axes pour mieux lire les coordonnées des variables initiales sur les axes. Cette procédure ne modifie pas le nuage, il change la projection des variables sur les axes principaux.

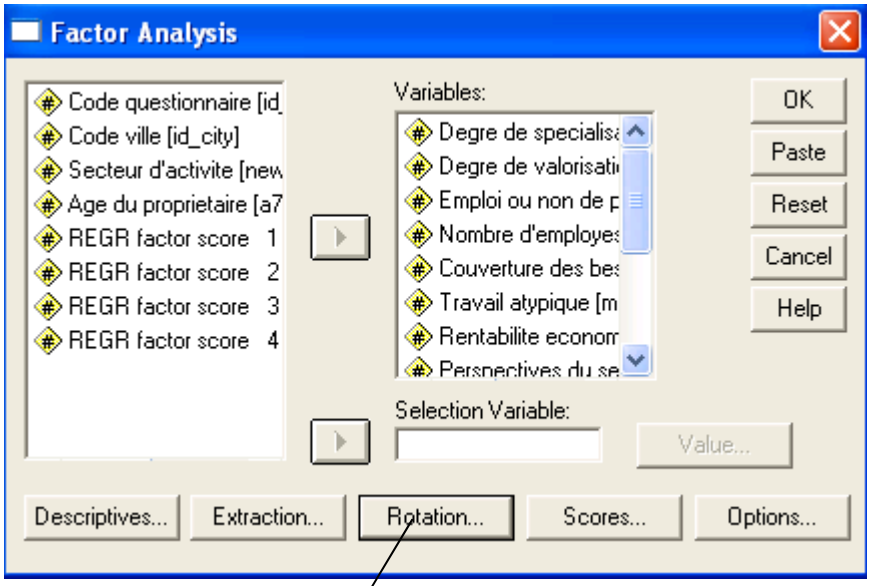
Présentation sans rotation

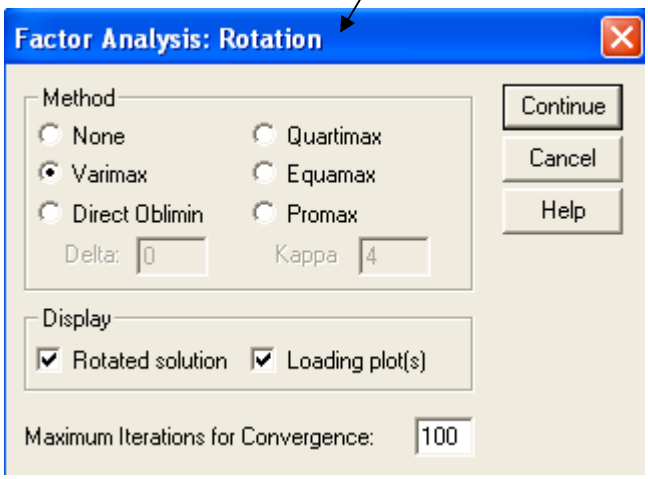


Présentation avec rotation



Pour cela, nous sélectionnons, une fois la commande Analyze > Data reduction > factor choisie, la sous-commande Rotation qui permet de choisir le type voulu de rotation. Généralement, la procédure la plus efficace est celle du Varimax.




Hormis cette commande supplémentaire, toutes les autres commandes sont les mêmes que celles présentées dans le précédent exercice.

## RESULTAT DE L' ANALYSE AVEC 11 VARIABLES

**Descriptive Statistics**

| | Mean | Std. Deviation | Analysis N |
|---|---|---|---|
| Degre de specialisation du chef d'entreprise | ,882 | 1,676 | 390 |
| Degre de valorisation du personnel | 93,397 | 15,522 | 390 |
| Emploi ou non de personnel | ,462 | ,499 | 390 |
| Nombre d'employes permanents | 2,397 | 6,738 | 390 |
| Couverture des besoins en qualification | 8,236 | 1,459 | 390 |
| Travail atypique | ,644 | ,480 | 390 |
| Rentabilite economique | 6,662 | 1,729 | 390 |
| Perspectives du secteur | ,036 | ,826 | 390 |
| Perspectives de l'entreprise | ,341 | ,716 | 390 |
| Pourcentage de jeunes employes | 35,902 | 45,984 | 390 |
| Techniciens-ouvriers | 29,305 | 43,866 | 390 |

Toutes les variables initiales ont une relativement forte variabilité, ce qui est assez satisfaisant. On peut également vérifier que le déterminant de la matrice de corrélation est égale à 0,0569, ce qui est un niveau assez petit, vu le nombre de variables et d'individus.

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,727 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 1102,068 |
| | df | 55 |
| | Sig. | ,000 |

L'indice de KMO est égal à 0,727 ce qui est assez satisfaisant et signifie bien que le choix des variables initiales est pertinent.

**Communalities**

| | Initial | Extraction |
|---|---|---|
| Degre de specialisation du chef d'entreprise | 1,000 | ,640 |
| Degre de valorisation du personnel | 1,000 | ,374 |
| Emploi ou non de personnel | 1,000 | ,848 |
| Nombre d'employes permanents | 1,000 | ,300 |
| Couverture des besoins en qualification | 1,000 | ,636 |
| Travail atypique | 1,000 | ,639 |
| Rentabilite economique | 1,000 | ,546 |
| Perspectives du secteur | 1,000 | ,616 |
| Perspectives de l'entreprise | 1,000 | ,690 |
| Pourcentage de jeunes employes | 1,000 | ,838 |
| Techniciens-ouvriers | 1,000 | ,700 |

Extraction Method: Principal Component Analysis.

Sur la base des axes principaux sélectionnés par le programme, il apparaît que deux variables présentent une relativement mauvaise qualité en termes

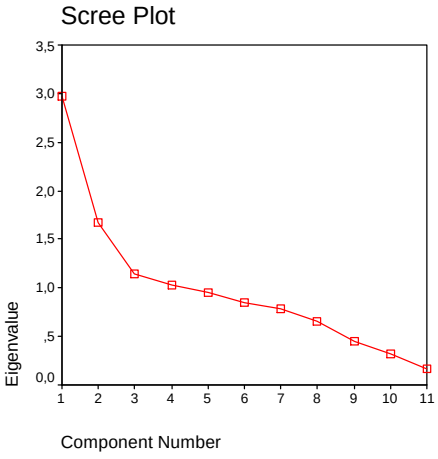
de pourcentage de variance prise en compte : degré de valorisation du personnel (37,4%) et nombre d'employés permanents (30%).

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2,975 | 27,043 | 27,043 | 2,975 | 27,043 | 27,043 | 2,696 | 24,505 | 24,505 |
| 2 | 1,675 | 15,231 | 42,274 | 1,675 | 15,231 | 42,274 | 1,794 | 16,307 | 40,812 |
| 3 | 1,143 | 10,392 | 52,666 | 1,143 | 10,392 | 52,666 | 1,193 | 10,843 | 51,655 |
| 4 | 1,032 | 9,386 | 62,052 | 1,032 | 9,386 | 62,052 | 1,144 | 10,397 | 62,052 |
| 5 | ,952 | 8,659 | 70,711 | | | | | | |
| 6 | ,853 | 7,753 | 78,464 | | | | | | |
| 7 | ,787 | 7,157 | 85,620 | | | | | | |
| 8 | ,654 | 5,943 | 91,564 | | | | | | |
| 9 | ,445 | 4,041 | 95,605 | | | | | | |
| 10 | ,322 | 2,923 | 98,528 | | | | | | |
| 11 | ,162 | 1,472 | 100,00 | | | | | | |

Extraction Method: Principal Component Analysis.

Quatre axes principaux sont extraits et couvrent 62% de la variance totale. Ce pourcentage est relativement faible mais il est assez satisfaisant si l'on tient compte de la nature des données employées. Il serait en fait possible d'envisager l'introduction du 5ème axe qui a une valeur propre inférieure à 1 mais assez proche de 1 et sa contribution est de l'ordre de 8,7%. Ainsi, nous aurions près de 71% de l'inertie totale.




La pente entre la 4ème et 5ème valeur propre ne change pas vraiment, voilà pourquoi, le 5ème axe pourrait être introduit.

**Component Matrix[a]**

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Degre de specialisation du chef d'entreprise | ,200 | ,229 | **-,582** | -,457 |
| Degre de valorisation du personnel | ,111 | ,428 | -,211 | ,365 |
| Emploi ou non de personnel | **,876** | -,283 | ,033 | -,003 |
| Nombre d'employes permanents | ,499 | -,165 | ,139 | ,058 |
| Couverture des besoins en qualification | ,157 | ,129 | -,294 | **,713** |
| Travail atypique | -,246 | ,091 | **,754** | ,029 |
| Rentabilite economique | ,403 | **,535** | ,160 | ,268 |
| Perspectives du secteur | ,340 | **,645** | ,095 | -,276 |
| Perspectives de l'entreprise | ,473 | **,636** | ,191 | -,161 |
| Pourcentage de jeunes employes | **,878** | -,253 | ,002 | -,054 |
| Techniciens-ouvriers | **,741** | -,370 | ,113 | ,038 |

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

**Rotated Component Matrix[a]**

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Degre de specialisation du chef d'entreprise | -,023 | ,241 | **,755** | -,108 |
| Degre de valorisation du personnel | -,095 | ,260 | ,092 | **,537** |
| Emploi ou non de personnel | **,906** | ,108 | ,119 | ,028 |
| Nombre d'employes permanents | ,538 | ,073 | -,064 | ,026 |
| Couverture des besoins en qualification | ,082 | -,087 | ,018 | **,788** |
| Travail atypique | -,155 | ,166 | **-,728** | -,238 |
| Rentabilite economique | ,165 | **,582** | -,140 | ,401 |
| Perspectives du secteur | ,023 | **,773** | ,129 | -,032 |
| Perspectives de l'entreprise | ,165 | **,812** | ,022 | ,053 |
| Pourcentage de jeunes employes | **,889** | ,139 | ,169 | ,003 |
| Techniciens-ouvriers | **,837** | -,008 | ,000 | -,008 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

A la lecture du dernier tableau, il est possible de voir la signification des 4 axes principaux.

Interprétation des axes factoriels

Le 1er axe est relatif à la structure du personnel employé par les entreprises lorsqu'elles en emploient : **dimension structurelle** de l'entreprise

Le 2ème axe est l'axe relatif au futur de l'entreprise, à la perception de son avenir par le chef d'entreprise, perception liée à la bonne santé de l'entreprise : **dimension économique** de l'entreprise.

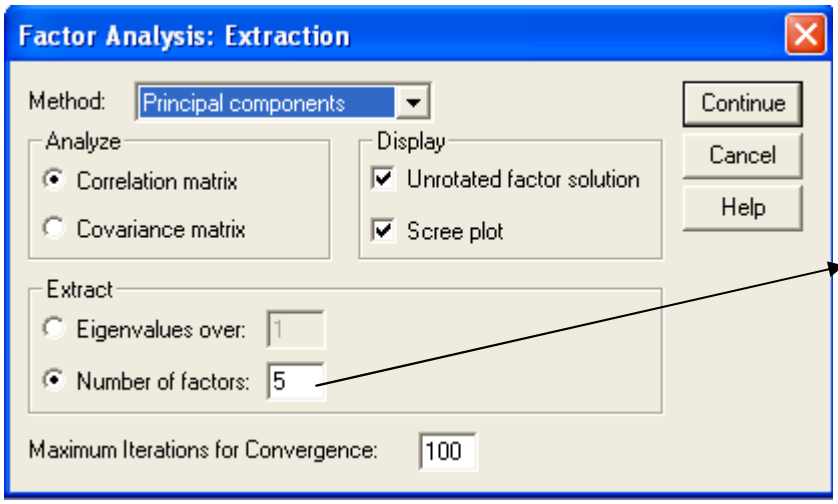
Le 3ème axe est l'axe de recours au travail familial qui est également lié à la qualification ou non du chef d'entreprise. **Dimension familiale** de l'entreprise.

Le 4ème axe est un axe qui renvoie à une «bonne» qualification du personnel employé : **dimension organisationnelle de l'entreprise** puisqu'il renvoie

au degré plus ou moins fort de couverture des besoins de l'entreprise en qualifications nécessaire à son bon fonctionnement.

Il résulte de l'analyse précédente que les 4 axes ont une signification assez précise et renvoient à des dimensions spécifiques. Une variable n'intervient pas. Il s'agit du nombre d'employés permanents qui reflète la taille de l'entreprise. Cela vient conforter le premier résultat que nous avions émis, concernant cette variable. Ce résultat n'est pas vraiment surprenant dans la mesure ou les entreprises analysées sont de petites entreprises. Rares sont celles qui ont plus de 10 employés (environ 5% de l'échantillon).

Par ailleurs, si l'on extrait 5 axes et non pas 4, en utilisant la commande suivante dans extraction:




Dans la commande Extraction, nous avons choisi d'extraire 5 axes et non plus d'extraire les seuls axes avant

il apparaît finalement que la variable «nombre d'employés permanents» **n'entre toujours pas dans la structure factorielle**. Donc, la solution de 5 axes n'est pas efficiente.

Il suffit pour cela de regarder la qualité des variables avec une représentation sur 5 axes puis directement la matrice des coordonnées des variables initiales sur les 5 axes sélectionnées (matrice après rotation des axes) pour vérifier ce résultat. De plus, nous avons choisi de n'imprimer que les valeurs des coordonnées supérieures à 0,5 (dans Options, on choisi de supprimer les valeurs inférieures à 0,5).
Qualité des variables quand on retient 5 axes au lieu de 4.

**Communalities**

| | Initial | Extraction |
|---|---|---|
| Degre de specialisation du chef d'entreprise | 1,000 | ,645 |
| Degre de valorisation du personnel | 1,000 | ,889 |
| Emploi ou non de personnel | 1,000 | ,849 |
| Nombre d'employes permanents | 1,000 | ,360 |
| Couverture des besoins en qualification | 1,000 | ,889 |
| Travail atypique | 1,000 | ,639 |
| Rentabilite economique | 1,000 | ,550 |
| Perspectives du secteur | 1,000 | ,621 |
| Perspectives de l'entreprise | 1,000 | ,743 |
| Pourcentage de jeunes employes | 1,000 | ,842 |
| Techniciens-ouvriers | 1,000 | ,751 |

→ Qualité continue d'être faible

Extraction Method: Principal Component Analysis.

Matrice des coordonnées des variables initiales sur les 5 axes.

**Rotated Component Matrix[a]**

| | Component | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Degre de specialisation du chef d'entreprise | | | ,755 | | |
| Degre de valorisation du personnel | | | | ,934 | |
| Emploi ou non de personnel | ,907 | | | | |
| Nombre d'employes permanents | ,512 | | | | |
| Couverture des besoins en qualification | | | | | ,940 |
| Travail atypique | | | -,732 | | |
| Rentabilite economique | | ,570 | | | |
| Perspectives du secteur | | ,770 | | | |
| Perspectives de l'entreprise | | ,846 | | | |
| Pourcentage de jeunes employes | ,893 | | | | |
| Techniciens-ouvriers | ,853 | | | | |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Il reste donc deux solutions possibles :
Solution 1: ôter la variable la moins pertinente (nombre d'employés permanents) et travailler avec 10 variables initiales (10 dimensions) au lieu de 11.
Solution 2: ôter les deux variables mal représentées (en termes de pourcentage de leur variance), à savoir le degré de valorisation du personnel et le nombre d'employés permanents et travailler avec 9 variables.

**Solution 1 :**
**Analyse avec 10 variables (suppression du nombre d'employés permanents)**

On constate alors, sur la base des deux tableaux qui suivent, que :

(a) l'indice de KMO ne se modifie pas substantiellement

(b) la proportion d'inertie totale expliquée passe de 62% à 66% toujours pour 4 axes principaux
(c) la signification des axes ne s'est pas modifiée par rapport à l'analyse précédente. Cela confirme bien que nos 4 dimensions composites sont explicatives du degré de dynamisme des entreprises étudiées

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2,795 | 27,954 | 27,954 | 2,795 | 27,954 | 27,954 | 2,497 | 24,974 | 24,974 |
| 2 | 1,659 | 16,592 | 44,546 | 1,659 | 16,592 | 44,546 | 1,796 | 17,958 | 42,932 |
| 3 | 1,137 | 11,369 | 55,915 | 1,137 | 11,369 | 55,915 | 1,182 | 11,817 | 54,749 |
| 4 | 1,032 | 10,317 | 66,232 | 1,032 | 10,317 | 66,232 | 1,148 | 11,483 | 66,232 |
| 5 | ,942 | 9,421 | 75,652 | | | | | | |
| 6 | ,821 | 8,205 | 83,858 | | | | | | |
| 7 | ,681 | 6,808 | 90,666 | | | | | | |
| 8 | ,445 | 4,448 | 95,114 | | | | | | |
| 9 | ,326 | 3,260 | 98,374 | | | | | | |
| 10 | ,163 | 1,626 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.

**Rotated Component Matrix[a]**

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Degre de specialisation du chef d'entreprise | -,005 | ,242 | **,747** | -,111 |
| Degre de valorisation du personnel | -,083 | ,251 | ,067 | **,548** |
| Emploi ou non de personnel | **,915** | ,121 | ,094 | ,035 |
| Couverture des besoins en qualification | ,077 | -,091 | ,033 | **,782** |
| Travail atypique | -,157 | ,164 | **-,743** | -,231 |
| Rentabilite economique | ,145 | **,582** | -,137 | ,408 |
| Perspectives du secteur | ,014 | **,774** | ,131 | -,028 |
| Perspectives de l'entreprise | ,152 | **,815** | ,024 | ,058 |
| Pourcentage de jeunes employes | **,907** | ,152 | ,137 | ,010 |
| Techniciens-ouvriers | **,869** | ,003 | -,045 | ,001 |

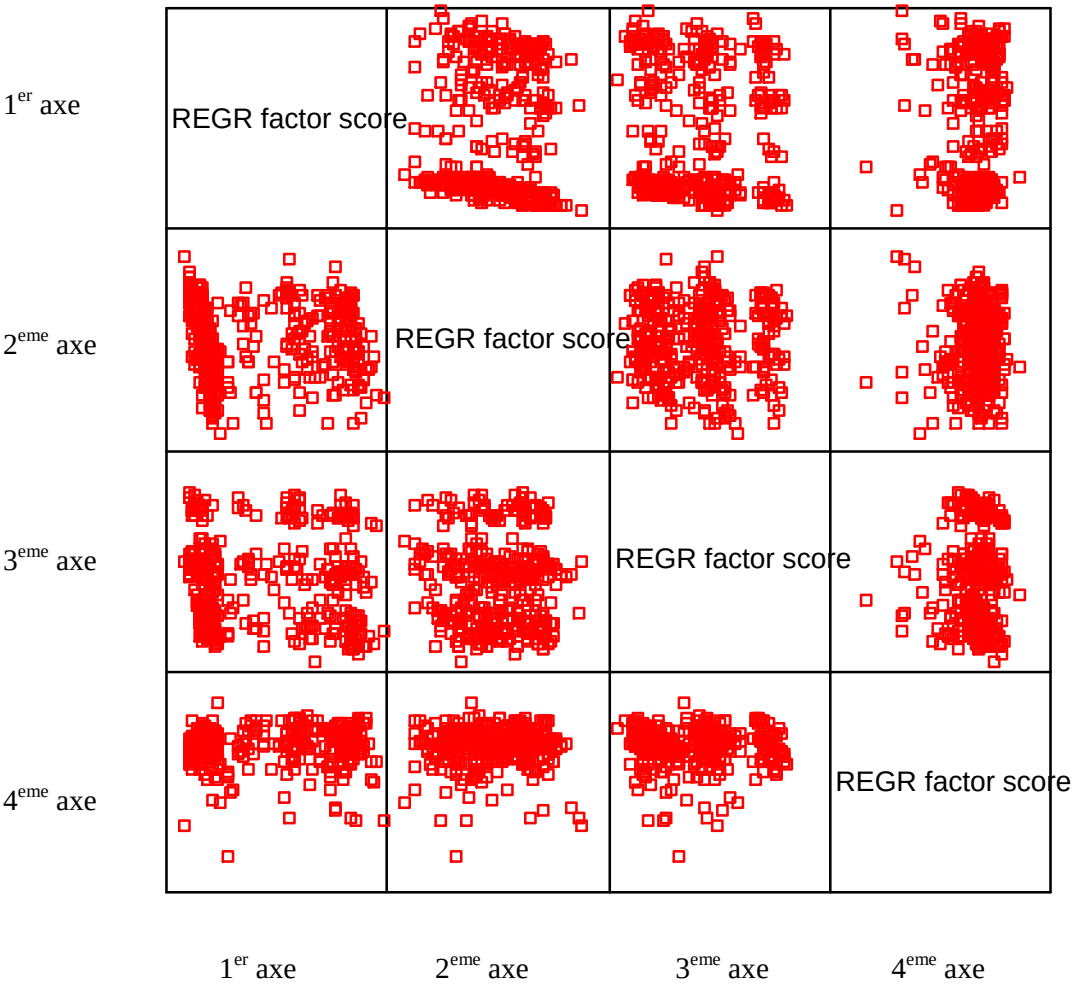
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Nous pouvons à l'aide de la commande Graphs Scatter, produire un graphique qui reproduit la projection des 390 entreprises sur les 4 axes principaux pris deux à deux. Nous avons en effet, sauvé les coordonnées des 390 individus sur chacun des axes. Ce sont les 4 nouvelles variables apparaissant dans la feuille de données.

Ainsi, nous obtenons la projection des individus sur les axes 1 et 2, 1 et 3, 1 et 4, puis 2 et 3 etc…

Représentation des individus dans les plans (1, 2), (1,3) etc.




On vérifie ainsi que les axes ne sont pas linéairement corrélés deux à deux entre eux, ce qui est le but recherché.

Sur la base de ces résultats, il est alors possible de procéder à une analyse de classification (clusters) pour faire émerger des groupes homogènes d'entreprises, en fonction des 4 variables composites que nous avons pu produire, 4 variables qui correspondent à 4 dimensions spécifiques.

## Solution 2 :
## Analyse avec 9 variables (suppression du nombre d'employés permanents et du degré de valorisation du personnel)

**Communalities**

| | Initial | Extraction |
|---|---|---|
| Degre de specialisation du chef d'entreprise | 1,000 | ,618 |
| Emploi ou non de personnel | 1,000 | ,859 |
| Couverture des besoins en qualification | 1,000 | ,903 |
| Travail atypique | 1,000 | ,644 |
| Rentabilite economique | 1,000 | ,510 |
| Perspectives du secteur | 1,000 | ,618 |
| Perspectives de l'entreprise | 1,000 | ,718 |
| Pourcentage de jeunes employes | 1,000 | ,864 |
| Techniciens-ouvriers | 1,000 | ,777 |

Extraction Method: Principal Component Analysis.

**Total Variance Explained**

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2,783 | 30,922 | 30,922 | 2,783 | 30,922 | 30,922 | 2,475 | 27,502 | 27,502 |
| 2 | 1,584 | 17,602 | 48,524 | 1,584 | 17,602 | 48,524 | 1,801 | 20,015 | 47,518 |
| 3 | 1,129 | 12,548 | 61,072 | 1,129 | 12,548 | 61,072 | 1,188 | 13,202 | 60,720 |
| 4 | 1,014 | 11,263 | 72,335 | 1,014 | 11,263 | 72,335 | 1,045 | 11,615 | 72,335 |
| 5 | ,823 | 9,139 | 81,474 | | | | | | |
| 6 | ,718 | 7,973 | 89,447 | | | | | | |
| 7 | ,455 | 5,057 | 94,504 | | | | | | |
| 8 | ,332 | 3,687 | 98,191 | | | | | | |
| 9 | ,163 | 1,809 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.

**Component Matrix[a]**

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Degre de specialisation du chef d'entreprise | | | -,632 | |
| Emploi ou non de personnel | ,877 | | | |
| Couverture des besoins en qualification | | | | ,885 |
| Travail atypique | | | ,751 | |
| Rentabilite economique | | ,514 | | |
| Perspectives du secteur | | ,667 | | |
| Perspectives de l'entreprise | | ,682 | | |
| Pourcentage de jeunes employes | ,887 | | | |
| Techniciens-ouvriers | ,752 | | | |

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

**Rotated Component Matrix[a]**

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Degre de specialisation du chef d'entreprise | | | ,747 | |
| Emploi ou non de personnel | ,909 | | | |
| Couverture des besoins en qualification | | | | ,946 |
| Travail atypique | | | -,748 | |
| Rentabilite economique | | ,641 | | |
| Perspectives du secteur | | ,761 | | |
| Perspectives de l'entreprise | | ,836 | | |
| Pourcentage de jeunes employes | ,904 | | | |
| Techniciens-ouvriers | ,880 | | | |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Globalement, en considérant cette dernière alternative, nous constatons les points suivants :

- d'une part, toutes les variables sont bien représentées sur les 4 axes principaux, leur qualité est d'au moins 50%.
- D'autre part, les 4 axes principaux rendent 72% de l'information initiale, ce qui est un niveau tout à fait satisfaisant pour ce type d'analyse et les deux premiers axes à eux seuls fournissent près de 50% de l'information.
- L'essence de l'analyse ne s'est pas modifiée puisque les quatre axes principaux conservent une signification précise.

De là, il est possible de réaliser une classification (cluster) pour dégager des groupes d'entreprises à comportement homogène.

**Classification des 390 entreprises de Thessalie sur la base des 4 axes factoriels obtenus à partir des variables initiales (2ème solution adoptée)**

Il suffit de procéder à une classification (k-means) en prenant comme variables, les coordonnées des 390 entreprises sur les 4 axes, variables qui ont été sauvées lors de la procédure d'Analyse en Composantes Principales (variables Fac1, fac2, Fac3 et Fac4).

Il est souhaitable dans un premier temps de choisir un nombre relativement élevé de classes et de comparer les résultats obtenus, au fur et à mesure que l'on diminue le nombre de classes.

La commande à utiliser est la suivante : Analyze > Classify > K-means cluster.
Il faut ensuite spécifier le nombre de classes. Il est suggérer de commencer avec au moins 10 classes.
Il ne faut pas oublier de préciser que le nombre d'itérations peut aller jusqu'à 100 et enfin sauver la variable qui indique à quel groupe (classe) appartient chaque entreprise.

Faire tourner le programme successivement pour 10, 9, 8, 7 et 6 classes. Ci-dessous, nous fournissons les résultats pour 6 classes.

Nous remarquons que le nombre d'itérations est de 17, ce qui est un résultat assez satisfaisant compte tenu du nombre d'entreprises que nous avons. La répartition des 390 entreprises est assez homogènes, seul un groupe contient un très petit nombre d'entreprises mais à la lecture des résultats de l'analyse avec un nombre de groupes plus élevé, il semble bien que ces 13 entreprises aient un comportement bien spécifique qui les distingue des autres entreprises.

Rappelons que l'axe 1 renvoie à des entreprises employant du personnel dont un pourcentage significatif de jeunes, l'axe 2 renvoie aux entreprises présentant des perspectives de développement, l'axe 3 renvoie aux entreprises familiales avec une forte présence de travail atypique tandis que l'axe 4 se réfère à la capacité de couverture des besoins en qualification.

## Classification des entreprises sur la base de 4 axes - choix de 6 groupes

**Initial Cluster Centers**

| | Cluster | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| REGR factor score 1 for analysis 1 | 1,69361 | ,54773 | -,73208 | -,84310 | ,68422 | -1,09179 |
| REGR factor score 2 for analysis 1 | -1,42133 | 1,68517 | -1,96969 | 1,35495 | -1,44173 | 1,21551 |
| REGR factor score 3 for analysis 1 | -,81039 | -1,28924 | ,65098 | ,42485 | ,64941 | 1,68414 |
| REGR factor score 4 for analysis 1 | -,43151 | 1,38493 | 1,49989 | -5,59997 | -3,41466 | -,11815 |

**Iteration History[a]**

| | Change in Cluster Centers | | | | | |
|---|---|---|---|---|---|---|
| Iteration | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1,351 | 1,563 | 1,462 | 1,517 | ,989 | 1,203 |
| 2 | ,542 | ,250 | ,258 | ,408 | ,326 | ,160 |
| 3 | ,225 | ,300 | ,066 | ,309 | ,117 | ,105 |
| 4 | ,197 | ,278 | ,067 | ,541 | ,228 | ,072 |
| 5 | ,156 | ,205 | ,048 | ,000 | ,147 | ,040 |
| 6 | ,119 | ,062 | ,087 | ,317 | ,439 | ,022 |
| 7 | ,044 | ,046 | ,030 | ,321 | ,342 | ,024 |
| 8 | ,088 | ,000 | ,024 | ,000 | ,274 | ,000 |
| 9 | ,043 | ,058 | ,069 | ,241 | ,358 | ,022 |
| 10 | ,089 | ,026 | ,069 | ,000 | ,440 | ,032 |
| 11 | ,081 | ,000 | ,025 | ,407 | ,394 | ,077 |
| 12 | ,027 | ,012 | ,055 | ,181 | ,180 | ,038 |
| 13 | ,000 | ,021 | ,033 | ,187 | ,059 | ,018 |
| 14 | ,019 | ,000 | ,016 | ,000 | ,071 | ,018 |
| 15 | ,000 | ,000 | ,000 | ,000 | ,066 | ,041 |
| 16 | ,000 | ,000 | ,000 | ,000 | ,033 | ,021 |
| 17 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 17. The minimum distance between initial centers is 3,471.

**Final Cluster Centers**

| | Cluster | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| REGR factor score 1 for analysis 1 | **1,20464** | -,77541 | -,65657 | ,91517 | **1,17235** | -,60162 |
| REGR factor score 2 for analysis 1 | ,25819 | ,21848 | **-1,16621** | -,07305 | -,31658 | **,88111** |
| REGR factor score 3 for analysis 1 | -,65717 | **-,92913** | ,23322 | -,26434 | **1,22547** | ,85902 |
| REGR factor score 4 for analysis 1 | ,46265 | -,02343 | ,06683 | **-3,75126** | -,05512 | ,12979 |

**ANOVA**

| | Cluster | | Error | | | |
|---|---|---|---|---|---|---|
| | Mean Square | df | Mean Square | df | F | Sig. |
| REGR factor score 1 for analysis 1 | 62,717 | 5 | ,196 | 384 | 319,340 | ,000 |
| REGR factor score 2 for analysis 1 | 36,241 | 5 | ,541 | 384 | 66,971 | ,000 |
| REGR factor score 3 for analysis 1 | 49,555 | 5 | ,368 | 384 | 134,745 | ,000 |
| REGR factor score 4 for analysis 1 | 40,506 | 5 | ,486 | 384 | 83,415 | ,000 |

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

**Number of Cases in each Cluster**

| | | |
|---|---|---|
| Cluster | 1 | 83,000 |
| | 2 | 91,000 |
| | 3 | 79,000 |
| | 4 | 13,000 |
| | 5 | 48,000 |
| | 6 | 76,000 |
| Valid | | 390,000 |
| Missing | | ,000 |

### 1.3. Conclusion sur l'Analyse en Composantes Principales

L'Analyse en Composantes Principales fournit nécessairement une solution unique qui maximise la variance expliquée par les facteurs. Les étapes de l'analyse consistent tout d'abord dans la détermination des variables à introduire pour explorer le phénomène que l'on cherche à étudier grâce à une analyse conjointe des variables qui représentent les diverses dimensions de ce phénomène.

Les variables retenues doivent montrer une certaine variation du positionnement des individus quant à ce qui est mesuré, il est donc souhaitable que les indices de dispersion $\sigma/\bar{x}$ soient suffisamment élevés.

Si le nombre de facteurs retenus est fixé par le critère de la valeur propre supérieure à 1, la solution n'a de sens réel que si les axes retenus reflètent un pourcentage suffisant de l'inertie totale. On admet souvent qu'un pourcentage satisfaisant est de l'ordre des 70 à 75%. Mais en la matière, il n'y a pas de critère scientifique.

Le maintien des variables initiales dépend des résultats fournis par le tableau des communautés (Communalities). La communauté représente l'appartenance de chaque variable à la covariance de l'ensemble des variables. C'est un indicateur du niveau de représentation de la variable initiale dans la solution obtenue. En général, on considère que la communauté minimale doit être au moins de 20% pour justifier le maintien de la variable dans l'analyse.

Le retrait d'une variable peut également se justifier par le fait qu'une variable n'entre pas ou peu dans la formation des axes principaux ou encore entre dans la détermination de tous les axes, traduisant finalement une trop grande complexité de cette variable, ce qui justifie alors son retrait.

Enfin, l'analyse nécessite une bonne décodification de la «signification conceptuelle» des axes. Le regroupement des variables initiales doit avoir un sens facile à comprendre et pertinent par rapport à la théorie et aux hypothèses de base. Comme le souligne à juste titre C. Durand (1997), «Le critère de la justesse de l'analyse est en partie subjectif. Il faut faire particulièrement attention à la tendance qu'ont certains chercheurs à donner aux facteurs des noms qui font du sens et qui impressionnent mais qui ne reflètent pas ce qui a été mesuré».

L'indice de KMO (Kaiser-Meyer-Olkin) est un indice qui nous indique justement à quel point l'ensemble des variables retenues dans l'analyse est un ensemble cohérent et permet de définir de nouvelles variables multidimensionnelles adéquates face aux concepts étudiés. Plus le KMO est élevé plus la solution trouvée est satisfaisante. Au dessous de la valeur 0,5, la solution doit être considérée comme inacceptable.