# UN GROUPE D'INDIVIDUS DÉCRITS PAR PLUSIEURS (>2) VARIABLES QUANTITATIVES

# CONSTRUIRE UN RÉSUMÉ FACTORIEL DES DONNÉES
# CLASSER LES INDIVIDUS

**Mots-clés** : Coefficient de corrélation de Bravais-Pearson ; Analyse factorielle ; Valeur propre ; Analyse en Composantes Principales (ACP) ; Classification automatique ; Classification Ascendante Hiérarchique (CAH) ; Cluster Analysis.

*Ce document a été établi en indiquant comment obtenir les différents résultats avec le logiciel SES-Pegase (version 6.1). Cependant, il peut être utilisé comme guide méthodologique et d'interprétation, quel que soit le logiciel utilisé.*

## TYPE DES DONNÉES ANALYSÉES

Nous présenterons l'analyse d'un dossier particulier, le dossier ONU. Mais cette analyse s'applique à toutes données de la forme suivante (Tableau 1).

**Tableau 1: Structure du tableau de données individuelles**

| INDIVIDU | X1 | X2 | X3 | X4 | (…) |
|---|---|---|---|---|---|
| i01 | | | | | |
| i02 | | | | | |
| i03 | | | | | |
| i04 | | | | | |
| i05 | | | | | |
| i06 | | | | | |
| i07 | | | | | |
| i08 | | | | | |
| i09 | | | | | |
| i10 | | | | | |
| i11 | | | | | |
| (…) | | | | | |

On a recueilli des données sur des individus (i01, i02…). Ces « individus » peuvent être, des personnes, des pays, des animaux, des voitures… Parmi les données recueillies, on a plusieurs variables quantitatives (temps, effectifs, salaires, températures, notes… ).
Ces variables peuvent être de natures différentes : l'une des variables peut être un temps, l'autre une note, une autre un salaire, etc.

## QUESTIONS

1. Peut-on résumer les données par un nombre plus réduit de variables ?
2. Peut-on constituer des classes d'individus se ressemblant du point de vue de leurs profils sur l'ensemble des variables ?

# UN EXEMPLE : LE DOSSIER ONU

On dispose des résultats d'une enquête menée par l'ONU[1] sur les Budgets-temps (temps passé dans différentes activités au cours de la journée, en moyenne) de 14 personnes.

La base de données est constituée des temps observés dans 10 activités : Profession, Transport, Ménage, Enfants, Courses, Toilette, Repas, Sommeil, Télé, Loisirs.

Les temps sont notés en heures et centièmes d'heures. Ainsi la première valeur en haut à gauche (6,10) indique que cette personnes (i01) passe en moyenne 6 heures et 6 minutes (6 heures + 10% de 60mn) par jour en activité professionnelle (PROF).

On dispose également de deux autres informations sur la personne interrogée : son sexe, son pays ou bloc de pays. Toutefois, on s'intéressera ici uniquement aux 10 variables quantitatives.

## Les données

**Tableau 2: Données ONU**

| INDIVIDU | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS | SEX | PAY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| i01 | 6.10 | 1.40 | 0.60 | 0.10 | 1.20 | 0.95 | 1.15 | 7.60 | 1.75 | 3.15 | H | USA |
| i02 | 4.75 | 0.90 | 2.50 | 0.30 | 1.40 | 1.20 | 1.00 | 7.75 | 1.15 | 3.05 | F | USA |
| i03 | 0.10 | 0.00 | 4.95 | 1.10 | 1.70 | 1.10 | 1.30 | 7.85 | 1.60 | 4.30 | F | USA |
| i04 | 6.15 | 1.40 | 0.65 | 0.10 | 1.15 | 0.90 | 1.15 | 7.65 | 1.80 | 3.05 | H | USA |
| i05 | 1.79 | 0.29 | 4.21 | 0.87 | 1.61 | 1.12 | 1.19 | 7.76 | 1.43 | 3.73 | F | USA |
| i06 | 5.85 | 1.15 | 0.50 | 0.00 | 1.50 | 1.05 | 1.00 | 7.60 | 1.50 | 3.85 | H | USA |
| i07 | 4.82 | 0.94 | 1.96 | 0.18 | 1.41 | 1.30 | 0.96 | 7.75 | 1.32 | 3.36 | F | USA |
| i08 | 6.53 | 1.00 | 0.95 | 0.07 | 0.57 | 0.85 | 1.50 | 8.08 | 1.15 | 3.30 | H | Ouest |
| i09 | 5.11 | 0.70 | 3.07 | 0.30 | 0.80 | 0.95 | 1.42 | 8.16 | 0.87 | 2.62 | F | Ouest |
| i10 | 0.20 | 0.07 | 5.68 | 0.87 | 1.12 | 0.90 | 1.80 | 8.43 | 1.25 | 3.68 | F | Ouest |
| i11 | 6.56 | 0.97 | 0.97 | 0.10 | 0.52 | 0.85 | 1.52 | 8.08 | 1.22 | 3.21 | H | Ouest |
| i12 | 1.68 | 0.22 | 5.28 | 0.69 | 1.02 | 0.83 | 1.74 | 8.24 | 1.19 | 3.11 | F | Ouest |
| i13 | 6.43 | 1.05 | 0.72 | 0.00 | 0.62 | 0.77 | 1.40 | 8.13 | 1.00 | 3.88 | H | Ouest |
| i14 | 4.29 | 0.34 | 2.62 | 0.14 | 0.92 | 0.97 | 1.47 | 8.49 | 0.84 | 3.92 | F | Ouest |

Source : Adapté d'une enquête de l'ONU (1967)

## Questions

1. Peut-on résumer les données par un nombre réduit de nouvelles variables ?
2. Peut-on constituer des classes d'individus se ressemblant du point de vue du temps qu'ils passent dans les différentes activités ?

## Type et statut des variables

Les 10 variables sont de même type : toutes quantitatives. Plus précisément il s'agit de variables de rapport.
Elles ont toutes le même statut de variable à analyser ou variable dépendante (VD).

## Ouverture du fichier

```
SES-Pegase
Menu Fichier – Ouvrir un dossier SES (*.SES)
Sélectionner le dossier ONU.SES
Menu Données à analyser
Sélectionner les 10 variables numériques comme variables dépendantes (VD).
```

[1] *Source : ONU (1967)*

# ANALYSER LES VARIABLES UNE À UNE

On commence par analyser les variables une à une, avant de chercher à mettre en relation les variables.
On se reportera à l'analyse du dossier NOTEBAC pour un exemple d'analyse détaillée d'une variable quatitative.

Pour cette partie, on se reportera à l'analyse du dossier NOTEBAC où l'on procède à l'analyse détaillée d'une variable numérique.

## Analyser la forme des distributions

On prend l'exemple de Profession.

```
SES-Pegase :
- Menu Données à analyser : sélectionner la variable PROF comme VD
- Menu Analyse – Voir la distribution des valeurs de PROF - Distribution en effectifs et
%

- Bouton Graphique
```

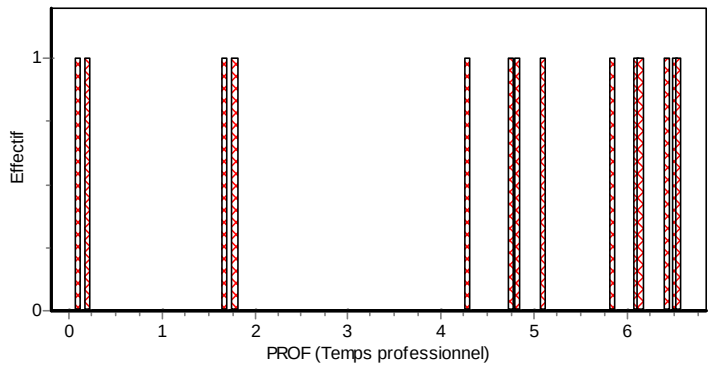


***Figure 1 : Distribution des valeurs de PROF***

Cette figure montre que la distribution de PROF présente deux, voire trois, groupes de valeurs, des valeurs faibles (dont certaines proches de 0) et des valeurs fortes (supérieures à 4 heures).

## Analyser les tendances centrales

```
SES-Pegase
Menu Analyse – Analyser les tendances centrales - Moyennes
```

***Tableau 3 : Moyennes des temps passés dans les différentes activités (en heures et centièmes d'heure)***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| Moy | 4.31 | 0.74 | 2.48 | 0.34 | 1.11 | 0.98 | 1.33 | 7.97 | 1.29 | 3.44 |

## Analyser les dispersions

```
SES-Pegase
Menu Analyse – Analyser les dispersions – Minimum et maximum
```

***Tableau 4 : Minimum et maximum***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| Min | 0.10 | 0.00 | 0.50 | 0.00 | 0.52 | 0.77 | 0.96 | 7.60 | 0.84 | 2.62 |
| Max | 6.56 | 1.40 | 5.68 | 1.10 | 1.70 | 1.30 | 1.80 | 8.49 | 1.80 | 4.30 |

SES-Pegase
Menu **Analyse – Analyser les dispersions – Écarts-type**

***Tableau 5 : écarts-type et écarts-type non corrigés***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| Ety | 2.27 | 0.46 | 1.82 | 0.36 | 0.37 | 0.15 | 0.26 | 0.29 | 0.29 | 0.44 |
| EtyC | 2.36 | 0.48 | 1.89 | 0.37 | 0.39 | 0.15 | 0.27 | 0.30 | 0.30 | 0.46 |

# Rédiger le compte rendu de l'analyse

Tous les calculs doivent être faits avec la précision maximale, ce que font très bien les logiciels statistiques. Par contre lorsque l'on communique les résultats, il n'est pas nécessaire d'indiquer un grand nombre de décimales. Trop de précision dans la présentation des résultats peut gêner la lecture, et la mémorisation, sans apporter une information réellement utile. On s'autorisera par exemple, lorsque le temps moyen de Ménage dans l'échantillon est de 2.48, à l'arrondir à 2.5 et à dire qu'il est « presque deux heures et demie » ou « un peu moins de deux heures et 30 mn ».[2]

On s'intéresse aux temps moyens, par jour, passés dans différentes activités.
Sur cet échantillon de 14 individus,
on constate (cf. Tableau 6) que :

Les activités qui, en moyenne, occupent le plus de temps sont, dans l'ordre :
Le Sommeil (m = 8.0h), la Profession (m = 4.3h), les Loisirs (m = 3.4h) et le Ménage (m = 2.5h).
Certaines activités occupent entre 1h et 2h par jour, en moyenne. C'est le cas de Repas, Télé, Courses, Transport, Enfants.
Enfin certaines activités occupent, en moyenne, moins d'une heure par jour : Toilette, Transport, Enfants.

Il existe peu de différences individuelles pour les temps Sommeil (de 7.6h à 8.5h, ety = 0.3h) et de Loisirs (de 2.6h à 4.3h, ety = 0.4h).
Par contre, il existe de plus grandes différences individuelles pour les temps Professionnel (de 0.1h à 6.6h, ety = 2.3h) et de Ménage (de 0.5h à 5.7h, ety= 1.8h).

***Tableau 6 : Minimum, maximum, moyennes et écarts-type***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| Min | 0.10 | 0.00 | 0.50 | 0.00 | 0.52 | 0.77 | 0.96 | 7.60 | 0.84 | 2.62 |
| Max | 6.56 | 1.40 | 5.68 | 1.10 | 1.70 | 1.30 | 1.80 | 8.49 | 1.80 | 4.30 |
| Moy | 4.31 | 0.74 | 2.48 | 0.34 | 1.11 | 0.98 | 1.33 | 7.97 | 1.29 | 3.44 |
| Ety | 2.27 | 0.46 | 1.82 | 0.36 | 0.37 | 0.15 | 0.26 | 0.29 | 0.29 | 0.44 |

Remarque : Dans le cas où, comme ici, les variables sont toutes sur la même échelle (ici des temps en heures) il est possible de présenter en résumé de l'analyse un graphique (cf. Figure 2) qui présente à la fois un indice de tendance centrale (ici les médianes) et des indices de dispersion (ici les minima, maxima et répartition en quartiles).

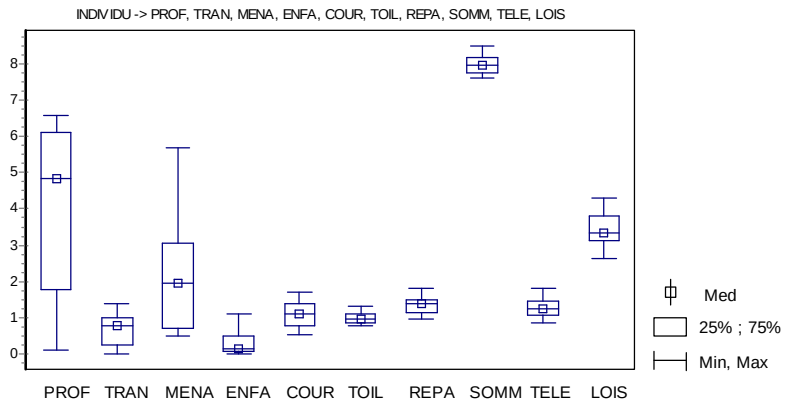


***Figure 2 : Répartition des valeurs aux différentes variables par quartiles***

[2] *Rappelons que les données sont présentées en heures et centièmes d'heures. Ainsi 2.48 se traduit par 2 heures et 29 minutes (0.48 × 60 mn = 28.8 mn).*

# ANALYSER LES LIAISONS ENTRE LES VARIABLES DEUX À DEUX

Pour cette partie, on se reportera à l'analyse du dossier INTELLIGENCE où l'on procède à l'analyse détaillée de la liaison entre deux variables quantitatives.

## Forme des liaisons ?

SES-Pegase
Menu **Analyse – Analyser les liaisons bivariées – Forme des liaisons ?**

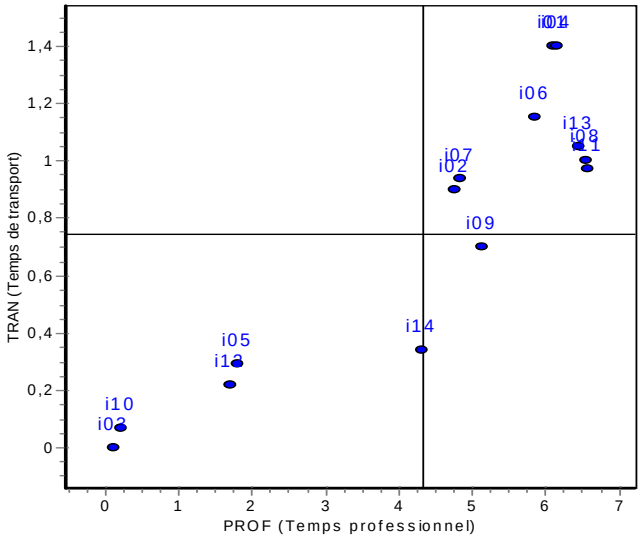


**Figure 3 : Nuage bivarié des individus pour Profession et Transport**

L'examen de l'ensemble des nuages bivariés montre :
- l'existence de liaisons non linéaires. C'est le cas par exemple de Transport et Télé, Enfants et Loisirs.
- l'existence de sous-groupes hétérogènes. C'est le cas de la plupart des couples avec la variable Profession, ou avec Enfants, mais aussi pour Transport et Loisirs, Courses et Sommeil, Toilette et Sommeil, Repas et Télé, Sommeil et Loisirs.

Il semble ne pas exister de valeurs atypiques.

## Sens des liaisons linéaires dans l'échantillon ?

Si la liaison est de type linéaire, cela signifie que l'on peut ajuster le nuage de points par une droite. On s'intéresse tout d'abord à la pente (positive ou négative) de cette droite.

SES-Pegase
Menu Analyse – Analyser les liaisons bivariées – Sens des liaisons linéaires ?
Cliquer sur le bouton Inférer

***Tableau 7: Signes des liaisons linéaires entre variables***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| PROF | + | + | - | - | - | - | - | - | - | - |
| TRAN | + | + | - | - | - | - | - | - | + | - |
| MENA | - | - | + | + | + | + | + | + | - | + |
| ENFA | - | - | + | + | + | + | + | + | + | + |
| COUR | - | - | + | + | + | + | - | - | + | + |
| TOIL | - | - | + | + | + | + | - | - | + | + |
| REPA | - | - | + | + | - | - | + | + | - | + |
| SOMM | - | - | + | + | - | - | + | + | - | + |
| TELE | - | + | - | + | + | + | - | - | + | + |
| LOIS | - | - | + | + | + | + | + | + | + | + |

Ce tableau affiche :
- un signe positif (+) si le coefficient de corrélation linéaire de Bravais-Pearson est positif,
- un signe négatif (-) si le coefficient de corrélation linéaire de Bravais-Pearson est négatif.

## Sens des liaisons linéaires dans la population ?

***Tableau 8: Résultats du test de Student sur les corrélations entre variables***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| PROF | + | + | - | - | ? | ? | ? | ? | ? | ? |
| TRAN | + | + | - | - | ? | ? | - | - | ? | ? |
| MENA | - | - | + | + | ? | ? | ? | ? | ? | ? |
| ENFA | - | - | + | + | ? | ? | ? | ? | ? | ? |
| COUR | ? | ? | ? | ? | + | + | - | - | + | ? |
| TOIL | ? | ? | ? | ? | + | + | - | ? | ? | ? |
| REPA | ? | - | ? | ? | - | - | + | + | ? | ? |
| SOMM | ? | - | ? | ? | - | ? | + | + | - | ? |
| TELE | ? | ? | ? | ? | + | ? | ? | - | + | ? |

Ce tableau affiche :
- un signe positif (+) si la liaison est positive dans la population (corrélation observée positive et test significatif) ;
- un signe négatif (-) si la liaison est négative dans la population (corrélation observée négative et test significatif) ;
- un point d'interrogation (?) si on ne peut pas conclure sur le sens de la liaison dans la population (test non significatif).

Le test utilisé est le test *t* de Student pour inférence sur le coefficient de corrélation de Bravais-Pearson, avec le seuil repère $p = .05$.

A titre d'exemple, il semble que, dans la population d'où est extrait l'échantillon :
- il existe une liaison linéaire positive entre PROF et TRAN (t[12]=7.58, $p < .0001$))
- il existe une liaison linéaire négative entre PROF et MENA (t[12]=-11.26, $p < .0001$))
- on ne peut pas conclure sur l'existence d'une liaison linéaire entre PROF et COUR (t[12]=-2.13, $p = .0541$).

## Force des liaisons linéaires dans l'échantillon ?

Lorsque le groupe observé est un échantillon d'un ensemble plus vaste (la population) on cherchera à connaître le sens de la liaison dans la population. Pour cela on analyse, pour chaque couple de variable, les résultats des tests « *t* de Student pour une corrélation ».

SES-Pegase
Menu **Analyse – Analyser les liaisons bivariées – Force des liaisons linéaires ?**
Cliquer sur le bouton Inférer

***Tableau 9: Coefficients de corrélation linéaire de Bravais-Pearson observés sur l'échantillon***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| **PROF** | +1.00 | +.91 | -.96 | -.96 | -.52 | -.22 | -.37 | -.28 | -.07 | -.42 |
| **TRAN** | +.91 | +1.00 | -.94 | -.86 | -.25 | -.09 | -.54 | -.57 | +.30 | -.43 |
| **MENA** | -.96 | -.94 | +1.00 | +.92 | +.35 | +.14 | +.50 | +.44 | -.15 | +.21 |
| **ENFA** | -.96 | -.86 | +.92 | +1.00 | +.52 | +.21 | +.32 | +.17 | +.15 | +.34 |
| **COUR** | -.52 | -.25 | +.35 | +.52 | +1.00 | +.77 | -.57 | -.58 | +.57 | +.36 |
| **TOIL** | -.22 | -.09 | +.14 | +.21 | +.77 | +1.00 | -.72 | -.51 | +.20 | +.12 |
| **REPA** | -.37 | -.54 | +.50 | +.32 | -.57 | -.72 | +1.00 | +.86 | -.44 | +.01 |
| **SOMM** | -.28 | -.57 | +.44 | +.17 | -.58 | -.51 | +.86 | +1.00 | -.76 | +.09 |
| **TELE** | -.07 | +.30 | -.15 | +.15 | +.57 | +.20 | -.44 | -.76 | +1.00 | +.10 |
| **LOIS** | -.42 | -.43 | +.21 | +.34 | +.36 | +.12 | +.01 | +.09 | +.10 | +1.00 |

Ce tableau montre les valeurs des coefficients de corrélation linéaire de Bravais-Pearson (*r*) dans l'échantillon des 14 individus. Pour chacune d'elle on peut ainsi conclure sur la force de la liaison linéaire.

On considère qu'une liaison linéaire est faible si la valeur absolue du coefficient de corrélation est inférieure à .20. C'est le cas, par exemple, des liaisons entre Profession et Télé ($r$ = -.07) et entre Sommeil et Enfants ($r$ = +.17).
On considère qu'une liaison linéaire est modérée si la valeur absolue du coefficient de corrélation est comprise entre .20 et .40. C'est le cas, par exemple, des liaisons entre Profession et Toilette ($r$ = .22) et entre Transport et Télé ($r$ = +.30).
On considère enfin qu'une liaison linéaire est forte si la valeur absolue du coefficient de corrélation est supérieure à .40. C'est le cas, par exemple, des liaisons entre Profession et Transport ($r$ = +.91) et entre Profession et Courses ($r$ = -.52).

## Force des liaisons linéaires dans la population ?

Lorsque le groupe observé est un échantillon d'un ensemble plus vaste (la population) on cherchera à connaître la force de la corrélation dans la population. Pour cela on analyse, pour chaque couple de variable, l'intervalle de confiance sur le coefficient de corrélation.

***Tableau 10: Intervalles de confiance sur les corrélations***

| | PROF | TRAN | MENA | ENFA | COUR | TOIL | REPA | SOMM | TELE | LOIS |
|---|---|---|---|---|---|---|---|---|---|---|
| **PROF** | [+1.00] | | | | | | | | | |
| **TRAN** | [+.72;+.97] | [+1.00] | | | | | | | | |
| **MENA** | [-.99;-.85] | [-.98;-.79] | [+1.00] | | | | | | | |
| **ENFA** | [-.99;-.88] | [-.95;-.58] | [+.75;+.97] | [+1.00] | | | | | | |
| **COUR** | [-.82;+.03] | [-.69;+.33] | [-.24;+.74] | [-.04;+.81] | [+1.00] | | | | | |
| **TOIL** | [-.67;+.36] | [-.59;+.46] | [-.43;+.62] | [-.37;+.66] | [+.38;+.92] | [+1.00] | | | | |
| **REPA** | [-.74;+.22] | [-.83;+.01] | [-.07;+.81] | [-.26;+.72] | [-.84;-.04] | [-.90;-.28] | [+1.00] | | | |
| **SOMM** | [-.70;+.30] | [-.84;-.04] | [-.13;+.78] | [-.40;+.64] | [-.84;-.05] | [-.81;+.05] | [+.60;+.95] | [+1.00] | | |
| **TELE** | [-.57;+.48] | [-.29;+.71] | [-.63;+.42] | [-.42;+.62] | [+.03;+.84] | [-.38;+.66] | [-.78;+.13] | [-.91;-.36] | [+1.00] | |
| **LOIS** | [-.77;+.16] | [-.78;+.14] | [-.37;+.66] | [-.25;+.73] | [-.23;+.74] | [-.44;+.61] | [-.52;+.54] | [-.46;+.59] | [-.46;+.59] | [+1.00] |

Pour chaque couple de variables, on a les deux limites de l'intervalle de confiance (au seuil .05) soit au niveau de confiance 95%). Cet intervalle indique l'incertitude sur la vraie valeur de la corrélation dans la population (corrélation parente).
Remarque : ce tableau est symétrique (l'IC pour PROF et TRAN est égal à l'IC pour TRAN et PROF).

On utilise les mêmes valeurs repères (.20 et .40) que celles utilisées pour qualifier la force des corrélations observées. Il est donc possible de conclure :
- à une liaison linéaire faible lorsque l'intervalle de confiance ne comprend que des valeurs faibles (< .20) que celles-ci soient toutes positives (par exemple [+.03 ; + .19], ou toutes négatives (par exemple [-.16 ; -.06], ou à la fois positives et négatives (par exemple [-.17 ; +.14]).
- à une liaison linéaire modérée lorsque l'intervalle de confiance ne comprend que des valeurs comprises, en valeur absolue, entre .20 et .40 (par exemple [+.23 ; + .32], ou toutes négatives (par exemple [-.36 ; -.24].
- à une liaison linéaire forte lorsque l'intervalle de confiance ne comprend que des valeurs supérieures, en valeur absolue, à .40. C'est le cas, par exemple, des liaisons entre Profession et Transport (IC95% [+.72;+.97]) et entre Profession et Ménage (IC95% [-.99;-.85]).

## Rédiger le compte rendu de l'analyse

La présentation du tableau des intervalles de confiance suffit pour la présentation des résultats.[3] En effet :
1. L'intervalle de confiance indique, comme le test $t$ de Student, si on peut conclure à une liaison parente linéaire et positive (lors que l'intervalle ne comprend que des valeurs positives) ou à une liaison négative (si l'intervalle ne comprend que des valeurs négatives) ou si on ne peut pas conclure sur le sens de la liaison (lorsque l'intervalle comprend des valeurs négatives et positives). Si l'on prend le seuil repère .05 pour le test et le niveau de confiance 95% pour l'intervalle de confiance, l'intervalle de confiance est nécessairement cohérent avec le résultat du test $t$ de Student qui indique si l'on peut conclure à une liaison parente positive ou négative (cas d'un test significatif) ou si on ne peut pas conclure (cas d'un test non significatif).
2. L'intervalle de confiance présente l'avantage supplémentaire, par rapport au test $t$, de fournir une information sur la force de la liaison dans population.

[3] *Il peut remplacer la traditionnelle matrice de corrélations où l'on associe une, deux ou trois étoiles à chaque corrélation selon que le test* t *de Student est significatif avec des valeurs p inférieures à .05, .01, ou .001, et aucune étoile lorsque le test n'est pas significatif.*

On s'intéresse aux liaisons entre les temps passés dans un certain nombre d'activités quotidiennes,
dans la population d'où est extrait cet échantillon de 14 personnes,

1. Concernant le signe des liaisons linéaires,
Il semble qu'il existe (au seuil .05)
- des liaisons linéaires positives entre :
(PROF et TRAN), (MENA et ENFA), (COUR et TOIL), (COUR et TELE), (REPA et SOMM).
- des liaisons linéaires négatives entre :
(PROF et MENA), (PROF et ENFA), (TRAN et MENA), (TRAN et ENFA), (TRAN et REPA), (TRAN et SOMM), (COUR et REPA), (COUR et SOMM), (TOIL et REPA), (SOMM et TELE).
- il n'est pas possible de se prononcer sur le sens des liaisons linéaires pour les autres couples d'activités.

**< insérer** Tableau 10**>**

2. Concernant la force des liaisons linéaires,
il semble que (au niveau de confiance 95%),

Parmi les liaisons linéaires positives :
- on peut conclure à des liaisons fortes entre PROF et TRAN (IC95% [.72; +.97]), entre MENA et ENFA (IC95% [+.75; +.97]), entre REPA et SOMM (IC95% [+.60 ; +.95]),
- on peut conclure à une liaison modérée à forte entre COUR et TOIL (IC95% [+.38; +.92]),
- on ne peut pas conclure sur la force de liaison entre COUR et TELE (IC95% [+.03; +.84]).

Parmi les liaisons linéaires négatives :
- on peut conclure à des liaisons fortes entre PROF et MENA (IC95% [-.99;-.85]), entre PROF et ENFA (IC95% [-.99;-.88]), entre TRAN et MENA (IC95% [-.98;-.79]), entre TRAN et ENFA (IC95% [-.95;-.58]),
- on peut conclure à une liaison modérée à forte entre TOIL et REPA (IC95% [-.90;-.28]), entre SOMM et TELE (IC95% [-.91;-.36])
- on ne peut pas conclure sur la force de liaison entre TRAN et REPA (IC95% [-.83;+.01]), entre TRAN et SOMM (IC95% [-.84;-.04]), entre COUR et REPA (IC95% [-.84;-.04]), entre COUR et SOMM (IC95% [-.84;-.05]).

# CONSTRUIRE UN RÉSUMÉ FACTORIEL DES DONNÉES

L'objectif est de construire un nombre restreint de nouvelles variables (2, 3, 4… ?) à partir d'un grand nombre de variables initiales. Pour ce type de données (plusieurs variables quantitatives recueillies sur des individus) la méthode la plus classique est une méthode d'analyse factorielle : l'analyse en composantes principales (ACP) normée. Les nouvelles variables sont appelées variables factorielles (VF).

## Combien de variables factorielles retenir ?

L'ACP construit autant de variables factorielles que de variables initiales, soit ici 10. La première étape va consister à se demander quel est le nombre minimal de variables factorielles à retenir pour résumer les données contenues dans les 10 variables initiales sans trop perdre d'information.

Pour cela, on utilise deux types d'information :
- les « valeurs propres » (les variances des variables factorielles),
- les « qualités de représentation » (qualité de représentation des individus et des variables initiales par les variables factorielles).

### Les valeurs propres

SES-Pegase
Menu Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) – Valeurs propres

Cliquer sur l'icône 

***Tableau 11: Valeurs propres de l'ACP***

| VF | VP | Pct | %Cum |
|---|---|---|---|
| VF1 | 4,53276 | 45 % | 45 % |
| VF2 | 3,46904 | 35 % | 80 % |
| VF3 | 0,93425 | 9 % | 89 % |
| VF4 | 0,86605 | 9 % | 98 % |
| VF5 | 0,09326 | 1 % | 99 % |
| VF6 | 0,07155 | 1 % | 100 % |
| VF7 | 0,01672 | 0 % | 100 % |
| VF8 | 0,01365 | 0 % | 100 % |
| VF9 | 0,00271 | 0 % | 100 % |
| VF10 | 0,00000 | 0 % | 100 % |

VP est l'acronyme de Valeur Propre. Une valeur propre est la variance de la variable factorielle. Le total des VP est égal au nombre de variables initiales, soit ici 10.

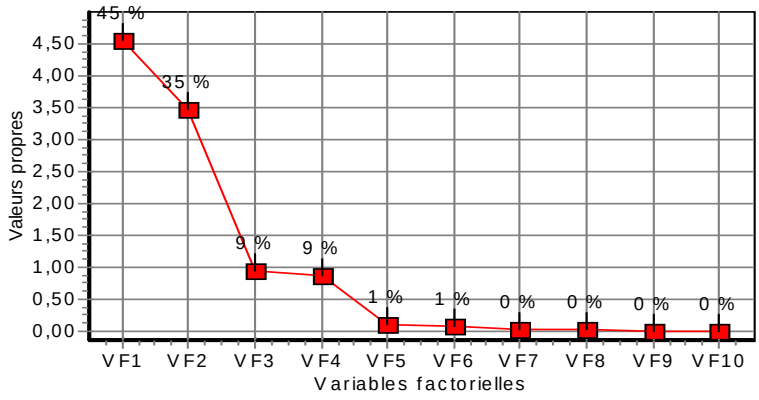


***Figure 4 : Décroissance des valeurs propres de l'ACP***

Ce graphique montre que les deux premières variables factorielles (VF1 et VF2) représentent deux parts importantes de la variance totale des données (respectivement 45% et 35%). Toutefois, les deux variables suivantes (VF3 et VF4) représentent une part non négligeable de cette variance (9% et 9%). Enfin, on peut

considérer que les variables suivantes (VF5 à VF10) ne présentent pas d'intérêt pour l'analyse car elles représentent de trop faibles parts de variances (1% au maximum).
Le Tableau 11 indique que les deux premières variables factorielles (VF1 et VF2) cumulent 80% de la variance totale. En y ajoutant les deux variables factorielles suivantes (VF3 et VF4), on cumulerait 98% de la variance.

Combien de variables factorielles retenir ? Il existe un critère simple, celui du pourcentage moyen. Soit ici 10%. En effet, sachant qu'il y a 10 valeurs propres, si les valeurs propres étaient identiques pour les 10 variables factorielles, on aurait 10% (100% / 10) pour chaque variable factorielle. On retiendra donc les variables factorielles pour lesquelles le pourcentage de valeurs propres est supérieur à 10%. Soit ici deux variables factorielles (VF1 et VF2).

## Qualité de représentation des variables par les variables factorielles (ACP)

Menu **Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) – Aides à l'interprétation des facteurs sur les variables – Qualité de représentation (cumulée) des variables par les variables factorielles.**

***Tableau 12 : Qualité de représentation (cumulée) des variables par les variables factorielles (ACP)***

| Variables | VF1 | VF2 | VF3 | VF4 | VF5 | VF6 | VF7 | VF8 | VF9 | VF10 |
|---|---|---|---|---|---|---|---|---|---|---|
| PROF | 91% | 98% | 99% | 99% | 100% | 100% | 100% | 100% | 100% | 100% |
| TRAN | 97% | 97% | 98% | 99% | 99% | 99% | 100% | 100% | 100% | 100% |
| MENA | 93% | 94% | 94% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| ENFA | 82% | 92% | 94% | 97% | 99% | 100% | 100% | 100% | 100% | 100% |
| COUR | 8% | 97% | 97% | 97% | 97% | 100% | 100% | 100% | 100% | 100% |
| TOIL | 0% | 64% | 97% | 97% | 98% | 100% | 100% | 100% | 100% | 100% |
| REPA | 33% | 91% | 98% | 99% | 99% | 99% | 99% | 100% | 100% | 100% |
| SOMM | 29% | 92% | 95% | 97% | 99% | 100% | 100% | 100% | 100% | 100% |
| TELE | 4% | 52% | 96% | 98% | 100% | 100% | 100% | 100% | 100% | 100% |
| LOIS | 17% | 24% | 27% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Ce tableau nous indique qu'en retenant deux variables factorielles (VF1 et VF2) on cumule au moins, pour chaque variable, 90 % de sa variance, à l'exception de Toilette (64%), Télé (52%) et Loisirs (24%). Il nous indique qu'il faudrait considérer quatre variables factorielles pour cumuler, pour toutes les variables, plus de 90% de la variance des 10 variables.
cf. Annexe, Tableau 22 pour la construction de ce tableau.

## Qualité de représentation des individus par les variables factorielles (ACP)

SES-Pegase
Menu **Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) – Aides à l'interprétation des facteurs sur les unités – Qualité de représentation (cumulée) des individus par les variables factorielles.**

| INDIVIDU | VF1 | VF2 | VF3 | VF4 | VF5 | VF6 | VF7 | VF8 | VF9 | VF10 |
|---|---|---|---|---|---|---|---|---|---|---|
| i01 | 68 % | 81 % | 96 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i02 | 11 % | 48 % | 90 % | 99 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i03 | 58 % | 93 % | 96 % | 99 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i04 | 65 % | 72 % | 93 % | 99 % | 99 % | 99 % | 99 % | 100 % | 100 % | 100 % |
| i05 | 40 % | 98 % | 98 % | 98 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i06 | 44 % | 83 % | 83 % | 97 % | 97 % | 99 % | 99 % | 100 % | 100 % | 100 % |
| i07 | 14 % | 68 % | 96 % | 96 % | 99 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i08 | 28 % | 96 % | 97 % | 98 % | 98 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i09 | 0 % | 46 % | 67 % | 98 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i10 | 89 % | 96 % | 98 % | 98 % | 100 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i11 | 27 % | 93 % | 95 % | 95 % | 95 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i12 | 63 % | 82 % | 85 % | 99 % | 99 % | 100 % | 100 % | 100 % | 100 % | 100 % |
| i13 | 17 % | 69 % | 70 % | 96 % | 99 % | 99 % | 100 % | 100 % | 100 % | 100 % |
| i14 | 14 % | 53 % | 69 % | 97 % | 99 % | 99 % | 100 % | 100 % | 100 % | 100 % |

Comme le tableau équivalent pour les variables (cf. Tableau 12) ce tableau indique que tous les individus seraient représentés à plus 95%, seulement si l'on considérait quatre variables factorielles (cf. colonne VF4).

En conclusion, on retiendra deux variables factorielles pour résumer les données, comme le suggère l'examen des valeurs propres, mais on n'oubliera pas que ce résumé factoriel, comme tout résumé, laisse de côté une partie des informations.

## Visualiser les liaisons entre les variables

```
SES-Pegase
Menu Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) –
Variables factorielles des variables

Cliquer sur l'icône
Si nécessaire, modifier le numéro des variables factorielles (VF) à représenter
Si nécessaire, redimensionner le graphique de manière à obtenir un cercle.
```

Les graphiques suivants sont construits à partir des valeurs du Tableau 23 (cf. Annexe)

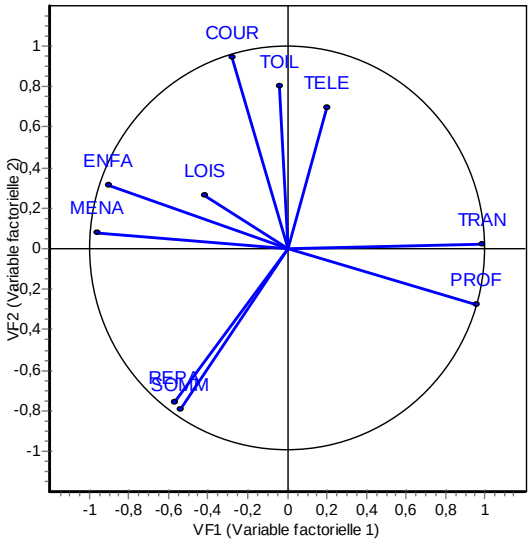


**Figure 5 : Représentation des variables dans l'espace des deux premières variables factorielles (ACP)**

La Figure 5 peut être vue comme une représentation graphique approchée de la matrice des corrélations (cf. Tableau 9). Chaque variable initiale est représentée par un vecteur (ici un segment). L'angle formé par les vecteurs donne une indication de la corrélation (linéaire) entre les variables.[4] La corrélation (linéaire) est nulle si l'angle est de 90°. Un angle inférieur à 90° indique une corrélation positive entre ces variables. Elle est d'autant plus forte que l'angle est proche de 0°. Un angle supérieur à 90° indique une corrélation négative. Elle est d'autant plus forte que l'angle est proche de 180°.
On ne considère que les variables pour lesquelles les extrémités du vecteur sont proches du cercle de corrélation car, sauf cas particulier, l'angle donne une mauvaise indication de la corrélation. On exclut donc ici les variables Loisir, Toilette et Télé.

La Figure 5 montre, par exemple, que :
- les variables Repas et Sommeil sont fortement corrélées positivement. De même les variables Enfants et Ménage, ainsi que Profession et Transport.
- les variables Transport et Ménage sont fortement corrélées négativement. De même les variables Profession et Enfants.
- les variables Transport et Courses sont corrélées négativement, mais faiblement.

La représentation factorielle fournit une représentation synthétique des corrélations entre les variables. Toutefois, lorsqu'il s'agit de connaître précisément la corrélation entre deux variables, on se référera toujours *in fine* à la matrice de corrélation.

[4] *Précisément, la corrélation est égale au cosinus de l'angle.*

SES-Pegase
Modifier les numéros des variables factorielles (VF) à représenter (VF3 et VF4)

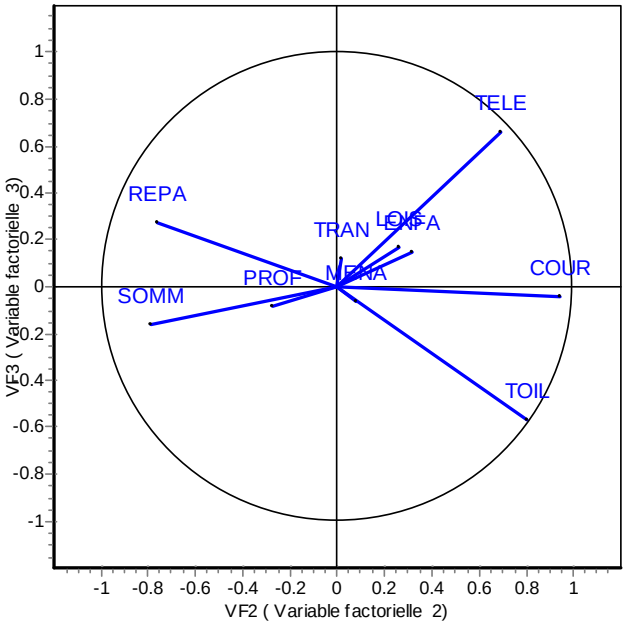


**Figure 6 : Représentation des variables dans l'espace des variables factorielles 2 et 3 (ACP)**

Même si on a décidé de ne considérer que les deux premières variables factorielles pour l'analyse, on présente la Figure 6 qui représente un autre plan factoriel constitué des variables factorielles 2 et 3 (VF2 et VF3). Elle permet de représenter la corrélation entre les variables Télé et Toilette, toutes deux mal représentées sur la figure précédente. On constate que l'angle est légèrement inférieur à 90°, ce qui suggère une corrélation positive mais faible entre ces deux variables. Cela suggère également une corrélation également positive, mais plus forte que la précédente, entre Courses et Toilette.

# Interpréter les variables factorielles

Les 10 variables initiales ont été remplacées par deux variables factorielles (VF). Il s'agit maintenant de voir en quoi chacune de ces deux VF résument les 10 variables initiales.

## Contributions des variables initiales aux variables factorielles (ACP)

SES-Pegase
Menu Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) –Aides à l'interprétation des facteurs sur les variables – Contributions des variables initiales aux variables factorielles.

***Tableau 13 : Contributions relatives des variables initiales aux premières variables factorielles***

| | VF1 | VF2 | VF3 | VF4 |
|---|---|---|---|---|
| PROF | **20%** | 2% | 1% | 1% |
| TRAN | **21%** | 0% | 2% | 0% |
| MENA | **20%** | 0% | 0% | 7% |
| ENFA | **18%** | 3% | 2% | 3% |
| COUR | 2% | **26%** | 0% | 0% |
| TOIL | 0% | **19%** | **34%** | 0% |
| REPA | 7% | **17%** | 8% | 1% |
| SOMM | 7% | **18%** | 3% | 2% |
| TELE | 1% | **14%** | **47%** | 2% |
| LOIS | 4% | 2% | 3% | **84%** |
| Total | **100%** | **100%** | **100%** | **100%** |

Ce tableau affiche en gras et rouge les contributions supérieures à la contribution moyenne. Pour cet exemple comprenant dix variables, cette contribution moyenne est de 100% / 10 = 10%.

Pour l'interprétation de chaque variable factorielle, on construira à la main un tableau qui regroupe les principales contributions des variables initiales à cette variable factorielle.

***Tableau 14 : Principales contributions des variables initiales à la première variable factorielle***

| Coordonnées négatives (-) | Coordonnées positives (+) |
|---|---|
| MENA (20%) | TRAN (21%) |
| ENFA (18%) | PROF (20%) |

Ce tableau montre les caractéristiques principales (puisque nous considérons la première variable factorielle) de ces données :
1. MENA et ENFA sont corrélées positivement. TRAN et PROF sont corrélés positivement
2. une opposition entre, d'une part, le temps passé à faire le ménage et à s'occuper des enfants et, d'autre part, le temps passé dans les transports et au travail. Dis autrement, cela signifie que les individus qui passent, relativement aux autres, beaucoup de temps à faire le ménage et à s'occuper des enfants passent, toujours relativement aux autres, peu de temps dans les transports et au travail. A l'inverse, les individus qui passent beaucoup de temps dans les transports et au travail, passent peu de temps à faire le ménage et à s'occuper des enfants.

On peut résumer la nature de cette première variable factorielle : elle traduit une opposition entre activités au domicile et activités à l'extérieur. On peut la désigner, par exemple, par la variable « Intérieur/Extérieur ».

***Tableau 15 : Principales contributions des variables initiales à la deuxième variable factorielle***

| Coordonnées négatives (-) | Coordonnées positives (+) |
|---|---|
| SOMM (18%) | COUR (26%) |
| REPA (17%) | TOIL (19%) |
| | TELE (14%) |

Ce tableau montre d'autres caractéristiques (moins importantes puisque nous considérons la deuxième variable factorielle) de ces données :
1. une corrélation positive entre SOMM et REPA et une corrélation positive entre COUR, TOIL et TELE,
2. une opposition entre, d'une part, le temps passé à dormir (SOMM) et dans les repas (REPA) et, d'autre part, le temps passé à faire des courses (COUR), faire sa toilette (TOIL) et regarder la télévision (TELE). Dis autrement, cela signifie que les individus qui passent, relativement aux autres, beaucoup de temps à dormir et dans les repas passent, relativement aux autres, peu de temps à faire des courses, faire sa toilette et regarder la télévision. A l'inverse, les individus qui passent beaucoup de temps à faire des courses, faire sa toilette et regarder la télévision, passent peu de temps à dormir et dans les repas.
On peut résumer la nature de cette deuxième variable factorielle : elle traduit une opposition entre deux manières d'occuper son temps au domicile, des occupations tranquilles (Sommeil et Repas) et des occupations plus actives (Courses, Toilette et, dans une moindre mesure, Télé). On peut la désigner comme étant la variable « Niveau d'activité à domicile ».

REMARQUE : Les noms proposés pour les variables factorielles peuvent être discutés. Nous sommes au-delà du strict résultat statistique, dans le domaine de l'interprétation. La statistique passe la main au chercheur (sociologue, psychologue…).

## Représenter les individus dans l'espace des variables factorielles

Après avoir interprété les variables factorielles, il va être possible de décrire les individus en fonction de ces variables factorielles.

Menu **Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) – Variables factorielles (ACP) des individus**

Cliquer sur l'icône 
Si nécessaire, modifier le numéro des deux variables factorielles (VF) à représenter.
Si nécessaire, redimensionner le graphique de manière à ce que l'échelle soit équivalente sur les deux axes.

***Tableau 16: Coordonnées factorielles (ACP) des individus sur les deux premières variables factorielles***

| INDIVIDU | VF1 | VF2 |
|---|---|---|
| i01 | 2.531 | 1.102 |
| i02 | 0.822 | 1.520 |
| i03 | -3.450 | 2.709 |
| i04 | 2.570 | 0.846 |
| i05 | -1.951 | 2.321 |
| i06 | 2.021 | 1.897 |
| i07 | 1.096 | 2.137 |
| i08 | 1.343 | -2.080 |
| i09 | 0.091 | -1.807 |
| i10 | -3.915 | -1.056 |
| i11 | 1.337 | -2.105 |
| i12 | -2.654 | -1.467 |
| i13 | 1.262 | -2.201 |
| i14 | -1.104 | -1.816 |

Ce tableau indique, pour chaque individu, ses coordonnées sur les 2 variables factorielles qui peuvent remplacer les 10 variables initiales.

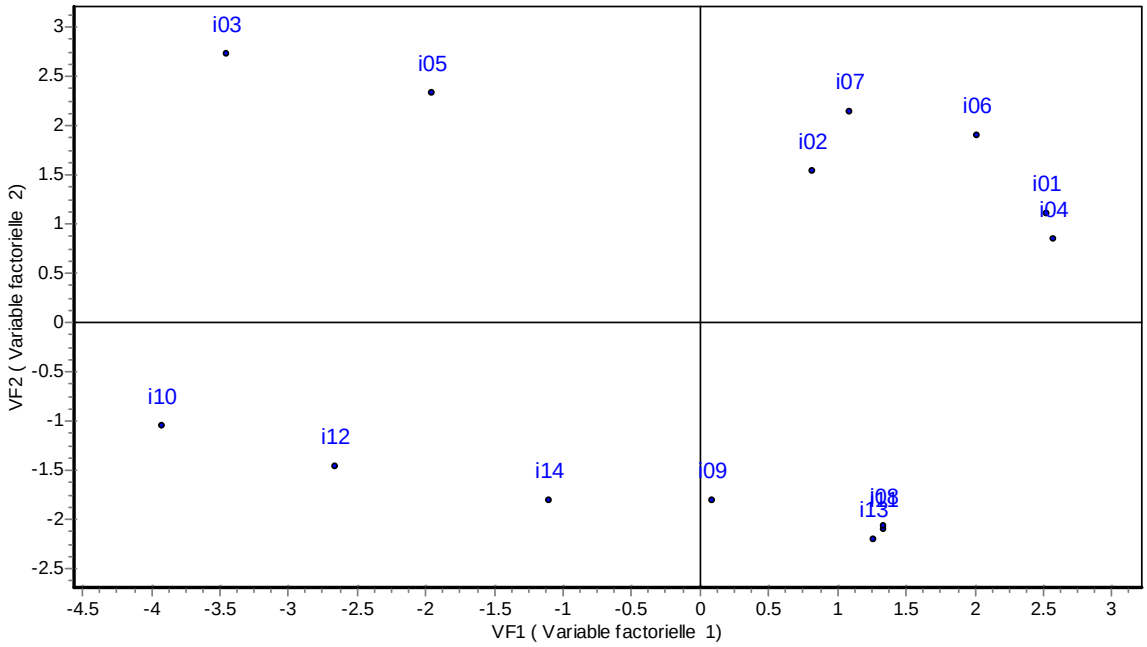


**Figure 7 : Représentation des individus dans l'espace des deux premières variables factorielles (ACP)**

SES-Pegase

Pour enregistrer ces VF dans la base de données, cliquer sur 


## Analyser les individus selon les variables factorielles

SES-Pegase
Menu **Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) – Aides à l'interprétation des facteurs sur les variables – Contributions des individus aux variables factorielles.**

***Tableau 17 : Contributions relatives des individus aux premières variables factorielles***

| | VF1 | VF2 | VF3 | VF4 | VF5 | VF6 | VF7 | VF8 | VF9 | VF10 |
|---|---|---|---|---|---|---|---|---|---|---|
| i01 | **10%** | 3% | **11%** | 3% | 3% | 0% | 0% | 0% | 3% | 4% |
| i02 | 1% | 5% | **20%** | 5% | 3% | 1% | 5% | 0% | 4% | 1% |
| i03 | **19%** | **15%** | 4% | 5% | 4% | 8% | 1% | 3% | 1% | 6% |
| i04 | **10%** | 1% | **16%** | 5% | 4% | 2% | 0% | 28% | 3% | 1% |
| i05 | 6% | **11%** | 0% | 0% | **14%** | 1% | 3% | 1% | 1% | 3% |
| i06 | 6% | **7%** | 0% | **11%** | 0% | 18% | 5% | 25% | 5% | 0% |
| i07 | 2% | **9%** | **18%** | 0% | **15%** | 9% | 7% | 2% | 20% | 18% |
| i08 | 3% | **9%** | 0% | 0% | 0% | 13% | 0% | 7% | 4% | 3% |
| i09 | 0% | 7% | **12%** | **18%** | 7% | 1% | 2% | 5% | 4% | 5% |
| i10 | **24%** | 2% | 3% | 0% | **17%** | 0% | 23% | 0% | 17% | 7% |
| i11 | 3% | **9%** | 1% | 0% | 0% | 29% | 6% | 2% | 0% | 8% |
| i12 | **11%** | 4% | 2% | **13%** | 0% | 11% | 3% | 15% | 30% | 26% |
| i13 | 3% | **10%** | 1% | **20%** | **18%** | 2% | 28% | 2% | 7% | 4% |
| i14 | 2% | 7% | **11%** | **19%** | **15%** | 4% | 17% | 11% | 0% | 13% |
| **Total** | **100%** | **100%** | **100%** | **100%** | **100%** | **100%** | **100%** | **100%** | **100%** | **100%** |

Ce tableau affiche en gras et rouge les contributions supérieures à la contribution moyenne. Pour cet exemple comprenant dix variables, cette contribution moyenne est de 100% / 14 = 7%.

***Tableau 18 : Principales contributions des individus à la première variable factorielle***

| Coordonnées négatives (-) | Coordonnées positives (+) |
|---|---|
| i03 (19%) | i01 (10%) |
| i05 (24%) | i04 (10%) |
| i12 (11%) | |

Ce tableau indique que la première variable factorielle oppose, du point de vue de leurs profils, les individus i03, i10 et i12 d'une part, aux individus i01 et i04 d'autre part.

La mise en correspondance du Tableau 14 et du Tableau 18 (première variable factorielle) montre que i03, i10 et i12 passent plus de temps que les autres au domicile (Ménage et Enfants) et moins de temps à l'extérieur (Profession et Transport). A l'inverse, i01 et i04 passent plus de temps que les autres à l'extérieur (Profession et Transport) et moins de temps au domicile (Ménage et Enfants).

***Tableau 19 : Principales contributions des individus à la deuxième variable factorielle***

| Coordonnées négatives (-) | Coordonnées positives (+) |
|---|---|
| i03 (15%) | i08 (9%) |
| i05 (11%) | i11 (9%) |
| i06 (7%) | i13 (10%) |
| i07 (9%) | |

La première variable factorielle oppose les individus i03, i05, i06, i07 d'une part, aux individus i08, i11 et i13 d'autre part, du point de vue de leurs profils.

La mise en correspondance du Tableau 15 et du Tableau 19 (deuxième variable factorielle) montre que i03, i05, i06 et i07 passent, relativement aux autres, beaucoup de temps en activités tranquilles (Repas et Sommeil) et peu de temps en activités moins tranquilles (Courses, Toilette et Télés). A l'inverse, i08, i11 et i13 passent, relativement aux autres, beaucoup de temps en Courses, Toilette et Télé et peu de temps en activités plus tranquilles (Repas et Sommeil).

## Rédiger le compte-rendu de l'analyse

Pour ce groupe de 14 individus,
l'analyse en composantes principales (ACP) fait apparaître les points suivants :

1. Du point de vue des temps passés dans les différentes activités (cf. Figure 5)):
- une corrélation entre le temps passé en ménage et avec les enfants,
- une corrélation entre le temps professionnel et le temps de transport,
- une corrélation positive entre le temps passé en sommeil et dans les repas,

- une corrélation entre le temps passés dans les courses, la toilette et devant la télévision.

Cette analyse montre également :
- une opposition principale (45% de la variance totale) entre deux manières d'occuper son temps, au domicile (Ménage et Enfants) ou à l'extérieur (Profession et Transport),
- une opposition secondaire (35% de la variance totale) entre deux manières d'occuper son temps au domicile, avec des occupations tranquilles (Sommeil et Repas) ou des occupations moins tranquilles (Courses, Toilette et Télé).

**< Insérer** Figure 5 **>**

2. Du point de vue des individus (cf. Figure 7),
on observe une opposition principale entre deux groupes :
- i03, i10 et i12 qui passent, relativement aux autres, beaucoup de temps à l'extérieur (Profession et Transport) et peu de temps au domicile (Ménage et Enfants),
- i01 et i04 qui passent, relativement aux autres, peu de temps à l'extérieur (Profession et Transport) et beaucoup de temps au domicile (Ménage et Enfants).

On observe également une opposition secondaire entre deux groupes :
- i03, i05, i06 et i07 qui passent, relativement aux autres, beaucoup de temps en occupations tranquilles (Repas et Sommeil) et peu de temps en occupations moins tranquilles (Courses, Toilette et Télé).
- i08, i11 et i13 qui passent, relativement aux autres, beaucoup de temps en Courses, Toilette et Télé et peu de temps en Repas et Sommeil.

**< Insérer** Figure 7 **>**

# CLASSER LES INDIVIDUS

Référence : Lebart, Piron, Morineau (2006), p.247-328.

Lorsque le tableau comprend un grand nombre d'individus, une classification automatique vise à constituer des classes d'individus en fonction de leurs profils, c'est-dire ici selon leurs « scores » sur les différentes variables. On regroupera dans une même classe les individus ayant des profils proches et dans des classes différentes les individus ayant des profils différents. La méthode utilisée ici est la Classification Ascendante Hiérarchique (CAH).

Il s'agit ici de classer les 14 individus en fonction de leurs profils (leurs temps passés dans les différentes activités) c'est-à-dire de regrouper dans une même classe les individus qui ont des profils voisins et regrouper dans des classes différentes les individus qui ont des profils différents.
Cette méthode particulière (CAH) consiste à comparer, non pas les profils de scores bruts, mais les scores Z (variables centrées-réduites). Ceci permet en particulier de classer les individus en fonction de variables qui seraient sur des échelles différentes.

**Synonymes :** Classification automatique / Cluster Analysis.

## Combien de classes retenir ?

```
SES-Pegase
Menu Analyse – Classer les INDIVIDU - Classification Ascendante Hiérarchique (Addad) –
Combien de classes retenir ?
```

```
SOMME DES INDICES DE NIVEAU   0.10000E+02
 ------------------------------------------------------------------------------
 !  J  ! I(J) ! A(J)! B(J)!T(J)!T(Q)! HISTOGRAMME DES INDICES DE NIVEAU
 ------------------------------------------------------------------------------
 !   27!  3841!   25!   26! 384! 384!****************************************
 !   26!  2533!   23!   24! 253! 637!*************************
 !   25!  1167!   18!   19! 117! 754!************
 !   24!   761!   17!   21!  76! 830!********
 !   23!   539!   20!   22!  54! 884!******
 !   22!   399!   14!    9!  40! 924!****
 !   21!   287!    6!   16!  29! 953!***
 !   20!   146!   13!   15!  15! 967!**
 !   19!   139!    3!    5!  14! 981!*
 !   18!   119!   10!   12!  12! 993!*
 !   17!    55!    2!    7!   5! 999!*
 !   16!     9!    4!    1!   1!1000!*
 !   15!     5!   11!    8!   0!1000!*
```

**Figure 8 : CAH – Histogramme des indices de niveau**

L'histogramme indique qu'il est possible de retenir une première partition en deux classes, puis une partition plus fine en trois classes.[5]
Avec une partition en deux classes (cf. ligne J=27) on « récupère » 38.4% de la variance totale (cf. colonne T(Q)). Avec une partition en trois classes (cf. ligne J=26) on récupère 63.7% (38.4% + 25.3%) de la variance totale.

## Composition des classes ?

```
SES-Pegase
Menu Analyse – Classer les INDIVIDU - Classification Ascendante Hiérarchique (Addad) –
Description des classes
```

[5] *Compte tenu des faibles indices de niveau qui suivent (cf. colonne I(J)), il serait illusoire de chercher à analyser une partition plus fine.*

| J | I(J) | A(J) | B(J) | P(J) | DESCRIPTION DES CLASSES DE LA HIERARCHIE |
|---|---|---|---|---|---|
| 27 | 3841 | 25 | 26 | 14 | |
| 26 | 2533 | 23 | 24 | 10 | i13 i11 i08 i14 i09 i02 i07 i06 i04 i01 |
| 25 | 1167 | 18 | 19 | 4 | i10 i12 i03 i05 |
| 24 | 761 | 17 | 21 | 5 | i02 i07 i06 i04 i01 |
| 23 | 539 | 20 | 22 | 5 | i13 i11 i08 i14 i09 |
| 22 | 399 | 14 | 9 | 2 | i14 i09 |
| 21 | 287 | 6 | 16 | 3 | i06 i04 i01 |
| 20 | 146 | 13 | 15 | 3 | i13 i11 i08 |
| 19 | 139 | 3 | 5 | 2 | i03 i05 |
| 18 | 119 | 10 | 12 | 2 | i10 i12 |
| 17 | 55 | 2 | 7 | 2 | i02 i07 |
| 16 | 9 | 4 | 1 | 2 | i04 i01 |
| 15 | 5 | 11 | 8 | 2 | i11 i08 |

**Figure 9 : CAH – Description des classes de la hiérarchie**

Ce tableau se lit de bas en haut. Il décrit comment les classes (numérotées de 15 à 27) se sont progressivement constituées, par emboîtements successifs. Les premiers individus regroupés (les plus proches du point de vue de leurs profils) sont les individus i11 et i18, puis i04 et i01 (…) A la sixième étape, la première classe (J=20) composée des individus i11 et i08, est regroupée avec i13 pour constituer une nouvelle classe comprenant 3 individus, etc.

## Représenter l'arbre des classes

SES-Pegase
Menu Analyse – Classer les INDIVIDU - Classification Ascendante Hiérarchique (Addad) – Représentation de l'arbre

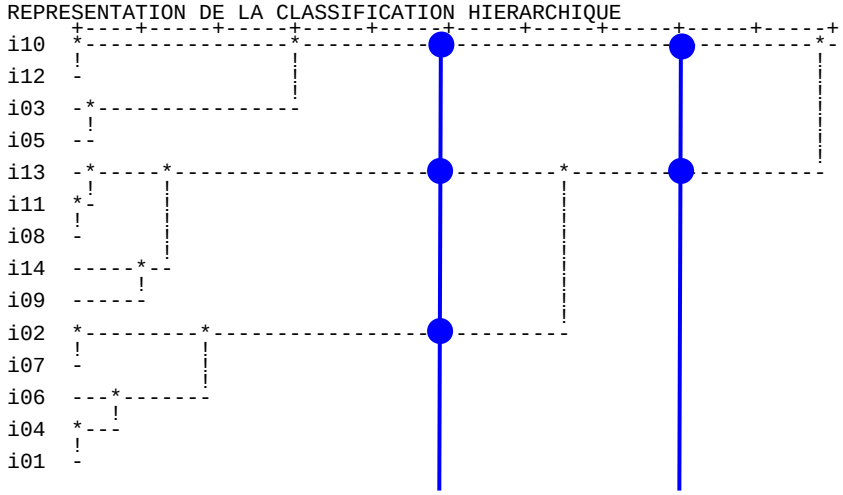


**Figure 10 : CAH – Représentation de l'arbre**

On a une représentation en arbre (dendrogramme) de la classification, du niveau le plus fin (à gauche) au plus grossier (à droite). Les deux traits verticaux ont été ajoutés pour indiquer les principales coupures de l'arbre en 2 et 3 classes. Ce graphique montre que les deux principales classes sont constituées de :
- classe 1 : i10, i12, i03, i05,
- classe 2 : i13, i11…. i01.

Un niveau d'analyse plus fin permet de séparer la classe la plus nombreuse (classe 2) en deux classes (2a et 2b), d'où une nouvelle partition, cette fois en trois classes :
- classe 1 : i10, i12, i03, i05,
- classe 2a : i13, i11, i08, i14, i09,
- classe 2b : i02, i07, i06, i04, i01.

# Analyser les profils des classes

Pour analyser les profils des classes il faut prendre en compte le fait que la classification automatique s'effectue, non pas sur les données de base, mais sur les variables centrées-réduites (scores z). On commencera donc par créer et enregistrer ces nouvelles variables dans la base de données.

## Voir les variables centrées-réduites (scores Z)

SES-Pegase
Menu Analyse – Dériver de nouvelles variables – Variables centrées-réduites (scores Z)

Cliquer sur l'icône  pour enregistrer ces variables dans la base de données.

Dans un deuxième temps on enregistre les classifications dans la base de données

SES-Pegase
Menu Analyse – Classer les INDIVIDU - Classification Ascendante Hiérarchique (Addad) – Enregistrer une partition dans la base de données
Indiquer le nombre de classes de la partition à enregistrer (ici 2)
Une fenêtre affiche l'affectation de chaque individu à l'une des 2 classes.

Cliquer sur l'icône  pour enregistrer cette nouvelle variable (CAH2).

## Analyser les profils moyens de chaque classe

SES-Pegase
Pour chacune des classes :
Menu Données à analyser - Définir un sous-ensemble de données analyser
Sélectionner les 10 variables centrées-réduites comme VD
Dans la fenêtre « sélection des unités », cliquer sur CAH2 puis cocher le numéro de classe à analyser
Menu Analyse – Analyser les tendances centrales – Moyennes

***Tableau 20 : Profils des deux classes principales de la CAH***

| Classes | PROF_Z | TRAN_Z | MENA_Z | ENFA_Z | COUR_Z | TOIL_Z | REPA_Z | SOMM_Z | TELE_Z | LOIS_Z |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -1.48 | -1.31 | 1.40 | 1.50 | 0.68 | 0.04 | 0.70 | 0.35 | 0.27 | 0.59 |
| 2 | 0.59 | 0.52 | -0.56 | -0.60 | -0.27 | -0.02 | -0.28 | -0.14 | -0.11 | -0.24 |

Pour les commentaires, on ne retiendra que les scores Z élevés (en valeur absolue). Par convention, on prendra 0.50 (un demi écart-type) comme valeur repère.

La classe 1 comprend des individus qui, relativement aux autres :
- passent beaucoup de temps en Ménage, Enfants, Courses, Repas, Loisirs,
- passent peu de temps en Profession et Transport.

On peut la qualifier de classe « Activités à domicile ».

La classe 2 comprend des individus qui :
- passent beaucoup de temps en Profession et Transport,
- passent peu de temps en Ménage, Enfants.

On peut la qualifier de classe « Activités à l'extérieur ».

***Tableau 21 : Profils des trois classes de la CAH***

| Classes | PROF_Z | TRAN_Z | MENA_Z | ENFA_Z | COUR_Z | TOIL_Z | REPA_Z | SOMM_Z | TELE_Z | LOIS_Z |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -1.48 | -1.31 | 1.40 | 1.50 | 0.68 | 0.04 | 0.70 | 0.35 | 0.27 | 0.59 |
| 2a | 0.65 | 0.15 | -0.45 | -0.62 | -1.13 | -0.70 | 0.52 | 0.76 | -0.96 | -0.13 |
| 2b | 0.54 | 0.90 | -0.68 | -0.58 | 0.59 | 0.67 | -1.08 | -1.03 | 0.74 | -0.34 |

On a noté 2a et 2b les deux sous-classes de la classe 2 précédente

La classe 1 (« Activités à domicile ») est inchangée. Elle comprend des individus qui :
- passent plus de temps que les autres en Ménage, Enfants, Courses, Repas, Loisirs,
- passent moins de temps que les autres en Profession et Transport.

La classe 2 (« Activités à l'extérieur ») comprenait des individus qui, en moyenne :
- passent plus de temps que les autres en Profession et Transport,
- passent moins de temps que les autres en Ménage, Enfants.

A l'intérieur de cette classe 2, celle des individus passant beaucoup de temps à l'extérieur, on distingue deux sous-classes selon la manière dont ces individus passent leur temps lorsqu'ils sont à la maison.
La classe 2a (i08, i09, i11, i13, i14) comprend les individus qui passent, relativement à l'ensemble du groupe :
- beaucoup de temps en Repas, Sommeil,
- peu de temps en Courses, Toilette, Télévision.

On peut la qualifier de classe (« Domicile tranquille »).
La classe 2b (i02, i07, i06, i04, i01) comprend les individus qui passent, relativement à l'ensemble du groupe :
- peu de temps en Repas, Sommeil,
- beaucoup de temps en Courses, Toilette, Télé.

On peut la qualifier de classe (« Domicile actif »).

REMARQUE : Les noms proposés pour les différentes classes peuvent être discutés, comme les noms proposés pour les variables factorielles. Nous sommes, là aussi, au-delà du strict résultat statistique, dans le domaine de l'interprétation.

L'analyse des résultats des deux méthodes, analyse factorielle et classification automatique, montre combien ces deux méthodes ne s'opposent pas mais sont complémentaires.

## Rédiger le compte-rendu de l'analyse

La classification automatique (Classification Ascendante Hiérarchique ou CAH) des individus à partir des temps passé dans 10 activités montre l'existence de deux classes principales (cf. Tableau 20).
La classe 1, comprend 4 individus (i03, i05, i10, i12) qui, relativement aux autres, passent beaucoup de temps dans des activités principalement centrées sur la maison (Ménage, Enfants, Courses, Repas, Loisirs) et peu de temps en dehors de la maison (Profession et Transport). On peut la nommer « Activités à domicile ».
La classe 2 comprend les 10 autres individus qui, à l'inverse, passent plus de temps que les autres en Profession et Transport, et moins de temps en Ménage, Enfants. On peut la nommer « Activités à domicile ».

**< Insérer Tableau 20 >**

A l'intérieur de la deuxième classe, on distingue deux sous-classes d'individus (cf. Tableau 21) qui se distinguent par leurs activités à la maison.
Les 5 individus de la classe 2a (i08, i09, i11, i13, i14) passent, relativement à l'ensemble du groupe, beaucoup de temps en Repas et Sommeil et peu de temps en Courses, Toilette et Télévision. On peut la nommer « Domicile tranquille ».
Les 5 individus de la classe 2b (i01, i02, i06, i04, i07) passent, à l'inverse, peu de temps en Repas et Sommeil et beaucoup de temps en Courses, Toilette et Télé. On peut la nommer « Domicile actif ».

**< Insérer Tableau 21>**

REMARQUE : Ces résultats suggèrent de mettre ces classes en relation avec le sexe des individus puisque l'on dispose de cette information.

# ANNEXES

## Qualité de représentation des variables par les variables factorielles

SES-Pegase
Menu **Résumer par méthodes factorielles – Analyse en Composantes Principales (ACP) –Aides à l'interprétation des facteurs sur les variables – Qualité de représentation des variables par les variables factorielles.**

***Tableau 22 : Qualité de représentation des variables par les variables factorielles (ACP)***

| | VF1 | VF2 | VF3 | VF4 | VF5 | VF6 | VF7 | VF8 | VF9 | VF10 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PROF | 91 % | 8 % | 1 % | 1 % | 0 % | 0 % | 0 % | 0 % | 0 % | 0 % | 100 % |
| TRAN | 97 % | 0 % | 1 % | 0 % | 0 % | 0 % | 1 % | 0 % | 0 % | 0 % | 100 % |
| MENA | 93 % | 1 % | 0 % | 6 % | 0 % | 0 % | 0 % | 0 % | 0 % | 0 % | 100 % |
| ENFA | 82 % | 10 % | 2 % | 3 % | 2 % | 1 % | 0 % | 0 % | 0 % | 0 % | 100 % |
| COUR | 8 % | 89 % | 0 % | 0 % | 0 % | 3 % | 0 % | 0 % | 0 % | 0 % | 100 % |
| TOIL | 0 % | 64 % | 32 % | 0 % | 1 % | 2 % | 0 % | 0 % | 0 % | 0 % | 100 % |
| REPA | 33 % | 58 % | 7 % | 1 % | 1 % | 0 % | 0 % | 1 % | 0 % | 0 % | 100 % |
| SOMM | 29 % | 63 % | 3 % | 2 % | 3 % | 0 % | 0 % | 0 % | 0 % | 0 % | 100 % |
| TELE | 4 % | 48 % | 44 % | 2 % | 2 % | 0 % | 0 % | 0 % | 0 % | 0 % | 100 % |
| LOIS | 17 % | 7 % | 3 % | 73 % | 0 % | 0 % | 0 % | 0 % | 0 % | 0 % | 100 % |

Ce tableau est la base du Tableau 12 des contributions cumulées. Dans ce Tableau 12 la contribution cumulée de LOIS (dernière ligne des tableaux) pour VF1 + VF2 est égale à 17% + 7% (cf. Tableau 22) soit 24% (cf. Tableau 12).

## Coordonnées factorielles des variables initiales

***Tableau 23 : Coordonnées factorielles (ACP) des variables***

| | VF1 | VF2 | VF3 | VF4 | VF5 | VF6 | VF7 | VF8 | VF9 | VF10 |
|---|---|---|---|---|---|---|---|---|---|---|
| PROF | 0.952 | -0.274 | -0.081 | 0.075 | -0.040 | -0.039 | -0.048 | 0.006 | -0.010 | 0.000 |
| TRAN | 0.984 | 0.021 | 0.122 | -0.060 | 0.027 | 0.032 | 0.107 | -0.003 | -0.011 | 0.000 |
| MENA | -0.964 | 0.077 | -0.060 | -0.238 | -0.004 | 0.053 | 0.030 | -0.009 | 0.027 | 0.000 |
| ENFA | -0.903 | 0.315 | 0.149 | -0.174 | -0.138 | -0.104 | 0.022 | 0.038 | -0.025 | 0.000 |
| COUR | -0.276 | 0.943 | -0.042 | 0.009 | 0.035 | 0.175 | -0.023 | -0.013 | -0.025 | 0.000 |
| TOIL | -0.043 | 0.801 | -0.567 | -0.022 | 0.122 | -0.136 | 0.012 | -0.023 | -0.003 | 0.000 |
| REPA | -0.572 | -0.760 | 0.270 | -0.088 | 0.076 | -0.040 | -0.003 | -0.080 | -0.018 | 0.000 |
| SOMM | -0.543 | -0.793 | -0.165 | 0.124 | 0.172 | 0.028 | 0.007 | 0.066 | -0.011 | 0.000 |
| TELE | 0.200 | 0.695 | 0.660 | -0.128 | 0.143 | -0.059 | -0.021 | 0.023 | 0.008 | 0.000 |
| LOIS | -0.415 | 0.262 | 0.165 | 0.854 | -0.016 | -0.023 | 0.021 | -0.010 | 0.004 | 0.000 |

# RÉFÉRENCES

APA (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
Escoffier, B., & Pagès, J. (1998). *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation* (3ème ed.). Paris: Dunod.
Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests ?* Mahwahn, NJ: Erlbaum.
Lebart, L., Piron, M., & Morineau, A. (2006). *Statistique exploratoire multidimensionnelle - Visualisation et inférence en fouilles de données* (4 ed.). Paris: Dunod.
Rouanet, Le Roux (1993). *Analyse des données multidimensionnelles*, Paris: Dunod.

# LISTE DES TABLEAUX



# LISTE DES FIGURES

# SOMMAIRE