

Rapport de projet du module de l’analyse des données : régression

linéaire mutiple, classification, ACP et ACM

*Réalisé par :*

JNAHI Yousra

CHABLAOUI Hanane

Filière 2IA

Année Scolaire 2021/2022

TABLE DES MATIÈRES

1. [Régression linéaire multiple](#_bookmark3) 12
   1. [Calcule du modèle de régression linéaire multiple](#_bookmark4) 12
      1. [Variables explicatives non significatives](#_bookmark6) 12
      2. [Valeur de R2 et Raju2](#_bookmark7) 13
      3. [Test de Fisher](#_bookmark8) 13
      4. [Conclusion](#_bookmark9) 13
   2. [Amélioration du modèle initiale par la procédure step](#_bookmark10) 13
      1. [Tests de validation](#_bookmark11) 16
      2. [Conclusion](#_bookmark16) 17
   3. [Application de la méthode pas à pas de sélection des variables explicatives](#_bookmark17) 18
      1. [Tests de validation](#_bookmark27) 25
      2. [Critère AIC](#_bookmark32) 26
      3. [Conclusion](#_bookmark33) 26
   4. [Conclusion](#_bookmark34) 27
      1. [Comparaison entre le modèle obtenu par step et celui obtenu par](#_bookmark35)

[cette méthode de sélection](#_bookmark35) 27

* + 1. [Choix entre le modèle obtenu par step et celui obtenu par cette](#_bookmark36) [méthode de sélection](#_bookmark36) 27

1. [Classification](#_bookmark37) 28
   1. [Variables centrées et réduites](#_bookmark38) 28
   2. [Application de k-means au tableau des variables quantitatives](#_bookmark41) 29
   3. [Détermination de](#_bookmark46) *[N](#_bookmark46)[c](#_bookmark46)* 30
   4. [Application de CAH au tableau des variables quantitatives](#_bookmark50) 32
   5. [Variables quantitatives les plus corrélées avec la variable classification](#_bookmark54) 34
   6. [Description des classes retenues par variables](#_bookmark56) 35
   7. [Taux d’inertie](#_bookmark60) 35
   8. [Comparer les classifications faites par k-means et CAH](#_bookmark62) 36
2. [Analyse en Composantes Principales (ACP)](#_bookmark63) 37
   1. [ACP normée sur le tableau des variables quantitatives](#_bookmark64) 37
   2. [Indice KMO et les indices MSAI](#_bookmark66) 38
   3. [Valeurs propres, pourcentage d’inertie et cumul des pourcentages d’inertie](#_bookmark68) . 38
   4. [Graphique des valeurs propres](#_bookmark70) 39

|  |  |  |
| --- | --- | --- |
|  | [3.5 Dimension du sous espace](#_bookmark73) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 40 |
| [3.6 Nuage des variables](#_bookmark75) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 40 |
| [3.6.1 cos2 des variables sur le sous espace](#_bookmark76) . . . . . . . . . . . . . . . . . . | 40 |
| [ment représentées](#_bookmark79) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 41 |
| [3.6.3 Contribution des variables dans chaque axe du sous espace](#_bookmark80) . . . . .  [aux axes du sous espace](#_bookmark82) . . . . . . . . . . . . . . . . . . . . . . . . | 41  42 |
| [3.6.5 Nuage des variables projeté sur les 2 premiers axes](#_bookmark87) . . . . . . . . . | 45 |
| [3.6.6 Variables relativement bien corrélées (positivement et négativement)](#_bookmark89) [avec les axes du 1er plan factoriel](#_bookmark89) . . . . . . . . . . . . . . . . . . . | 45 |
| [3.7 Nuage des individus](#_bookmark90) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  [3.7.1 cos2 des individus sur le sous espace](#_bookmark91) . . . . . . . . . . . . . . . . . | 45  45 |
| [3.7.2 Contribution des individus dans chaque axe du sous espace](#_bookmark94) . . . . . | 46 |
| [axes du sous espace](#_bookmark97) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 48 |
| [3.7.4 Nuage des individus projeté sur les 2 premiers axes](#_bookmark102) . . . . . . . . . | 50 |
| [3.8 Conclusion](#_bookmark104) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 50 |
| **[4](#_bookmark105)** | **[Analyse des Correspondances Multiples (ACM)](#_bookmark105)**  [4.1 Transformation des variables quantitatives](#_bookmark106) . . . . . . . . . . . . . . . . . .  [riables qualitatives et le taux de discrétisation de chaque transformation](#_bookmark108) . . | **51**  51  51 |
|  | [4.3 Tableau disjonctif complet](#_bookmark110) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 52 |
|  | [4.4 Fréquence de chaque modalité](#_bookmark112) . . . . . . . . . . . . . . . . . . . . . . . . . | 52 |
|  | [4.5 Application de l’ACM au tableau disjonctif complet](#_bookmark114) . . . . . . . . . . . . . | 53 |
|  | [4.6 Valeurs propres, pourcentage d’inertie de chaque valeur propre et cumul](#_bookmark117) [des pourcentages d’inertie](#_bookmark117) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 53 |
|  | [4.7 Graphique des valeurs propres](#_bookmark119) . . . . . . . . . . . . . . . . . . . . . . . . . | 54 |
|  | [4.8 Dimension du sous espace](#_bookmark121) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 55 |
|  | [4.9 Nuage de Modalités](#_bookmark124) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 55 |
|  | [4.9.1 cos2 des modalités sur le sous espace](#_bookmark125) . . . . . . . . . . . . . . . . . | 55 |
|  | [ment représentées sur le sous espace](#_bookmark129) . . . . . . . . . . . . . . . . . . | 58 |
|  | [4.9.3 Contribution des modalités dans chaque axe du sous espace](#_bookmark131) . . . . | 59 |
|  | [4.9.4 Application de la CAH au tableau des contributions des modalités](#_bookmark134) [aux axes du sous espace](#_bookmark134) . . . . . . . . . . . . . . . . . . . . . . . . | 62 |
|  | [4.9.5 Nuage des modalités projeté sur les 2 premiers axes](#_bookmark139) . . . . . . . . . | 64 |
|  | [4.10 Nuage des individus](#_bookmark141) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 64 |
|  | [4.10.1 cos2 des individus sur le sous espace](#_bookmark142) . . . . . . . . . . . . . . . . . | 64 |
|  | [4.10.2 Individus bien représentés, moyennement représentés et faiblement](#_bookmark149) [représentés sur le sous espace](#_bookmark149) . . . . . . . . . . . . . . . . . . . . . | 67 |
|  | [4.10.3 Contribution des individus dans chaque axe du sous espace](#_bookmark152) . . . . . | 68 |
|  | [aux axes du sous espace](#_bookmark157) . . . . . . . . . . . . . . . . . . . . . . . . | 71 |
|  | [4.11 Nuage des variables](#_bookmark162) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 73 |
|  | [4.11.1 Coefficients de corrélation des variables avec les projections sur les](#_bookmark163) [axes du sous espace](#_bookmark163) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 73 |

[3.6.2 Variables bien représentées, moyennement représentées et faible-](#_bookmark79)

[3.6.4 Application de la CAH au tableau des contributions des variables](#_bookmark82)

[3.7.3 Application la CAH au tableau des contributions des individus aux](#_bookmark97)

[4.2 Regroupement dans un fichier Excel, les variables quantitatives et les va-](#_bookmark108)

[4.9.2 Modalités bien représentées, moyennement représentées et faible-](#_bookmark129)

[4.10.4 Application de la CAH au tableau des contributions des individus](#_bookmark157)

[4.11.2 Graphique des coefficients de corrélation des variables avec les pro-](#_bookmark165) [jections sur le 1er plan factoriel](#_bookmark165) 73

[4.12 Conclusion](#_bookmark167) 74

TABLE DES FIGURES

1. [Table de notre jeu des données](#_bookmark0) 10
2. [Nuage des points](#_bookmark1) 10
3. [Exploration des attributs de notre jeu des données](#_bookmark2) 11
   1. [Résultat du modèle avec toutes les variables explicatives](#_bookmark5) 12
   2. [Test d’homoscédasticité](#_bookmark12) 16
   3. [Test de normalité : shapiro](#_bookmark13) 16
   4. [Test de normalité : ks](#_bookmark14) 16
   5. [Recherche de valeurs aberrantes](#_bookmark15) 17
   6. [La variable la plus significative (](#_bookmark18)*[F](#_bookmark18)* [le plus grand)](#_bookmark18) 18
   7. [Sommaire du modèle 1](#_bookmark19) 18
   8. [Sommaire du modèle 2](#_bookmark20) 19
   9. [Sommaire du modèle 3](#_bookmark21) 20
   10. [Sommaire du modèle 4](#_bookmark22) 21
   11. [Sommaire du modèle 5](#_bookmark23) 21
   12. [Sommaire du modèle 6](#_bookmark24) 22
   13. [Sommaire du modèle 7](#_bookmark25) 23
   14. [Sommaire du modèle 6](#_bookmark26) 24
   15. [Test d’homoscédasticité](#_bookmark28) 25
   16. [Test de normalité : shapiro](#_bookmark29) 25
   17. [Test de normalité : ks](#_bookmark30) 25
   18. [Recherche de valeurs aberrantes](#_bookmark31) 26
   19. [Tableau de jeu des données centrées et réduites](#_bookmark39) 28
   20. [Exploration des attributs](#_bookmark40) 29
   21. [Affichage des 4 groupes crées par le](#_bookmark42) *[k](#_bookmark42)*[-means sur R](#_bookmark42) 29
   22. [Taux d’inertie expliquée pour chaque classe pour](#_bookmark43) *[N](#_bookmark43)* [= 36](#_bookmark43) 30
   23. [Taux d’inertie expliquée pour chaque classe pour](#_bookmark44) *[N](#_bookmark44)* [= 37](#_bookmark44) 30
   24. [Taux d’inertie expliquée pour chaque classe pour](#_bookmark45) *[N](#_bookmark45)* [= 38](#_bookmark45) 30
   25. [Graphique de l’évolution de la proportion d’inertie expliquée par la parti-](#_bookmark47)

[tion](#_bookmark47) 31

* 1. [Détermination de](#_bookmark48) *[N](#_bookmark48)[c](#_bookmark48)* 31
  2. [Affichage des 13 groupes crées par le](#_bookmark49) *[k](#_bookmark49)*[-means sur R](#_bookmark49) 32
  3. [Résultats après application de CAH](#_bookmark51) 33
  4. [Résultats après application de CAH](#_bookmark52) 33
  5. [Dendrogramme de la CAH](#_bookmark53) 34
  6. [Valeur p-value de chaque variable](#_bookmark55) 34
  7. [Description de la classe 1](#_bookmark57) 35
  8. [Description de la classe 2](#_bookmark58) 35
  9. [Description de la classe 3](#_bookmark59) 35
  10. [Résultats des taux d’inertie/ nertie Inter/Inertie total, avant et après la](#_bookmark61) [consolidation de la CAH](#_bookmark61) 35
  11. [Résultat ACP normée sur le tableau des variables quantitatives](#_bookmark65) 37
  12. [Indice KMO et indices MSAI](#_bookmark67) 38
  13. [Résultat valeurs propres, pourcentage d’inertie de et cumul des pourcen-](#_bookmark69)

[tages d’inertie](#_bookmark69) 39

* 1. [Graphique des valeurs propres](#_bookmark71) 39
  2. [Graphique des valeurs propres](#_bookmark72) 40
  3. [Dimension du sous espace](#_bookmark74) 40
  4. [cos2 des variables sur le sous espace](#_bookmark77) 41
  5. [cos2 cumulé des variables sur le sous espace](#_bookmark78) 41
  6. [Contribution des variables dans chaque axe du sous espace](#_bookmark81) 42
  7. [Dendrogramme de la CAH](#_bookmark83) 42
  8. [Distribution des groupes](#_bookmark84) 43
  9. [Carte des facteurs](#_bookmark85) 43
  10. [Description des axes retenue par classes](#_bookmark86) 44
  11. [Nuage des variables projeté sur les 2 premiers axes](#_bookmark88) 45
  12. [cos2 des individus sur le sous espace](#_bookmark92) 46
  13. [cos2 des individus sur le sous espace](#_bookmark93) 46
  14. [Contribution des individus dans chaque axe du sous espace](#_bookmark95) 47
  15. [Contribution des individus dans chaque axe du sous espace (suite)](#_bookmark96) 47
  16. [Dendrogramme de la CAH](#_bookmark98) 48
  17. [Distribution des groupes](#_bookmark99) 48
  18. [Carte des facteurs](#_bookmark100) 49
  19. [Description des axes retenus par classes](#_bookmark101) 49
  20. [Nuage des individus projeté sur les 2 premiers axes](#_bookmark103) 50
  21. [Taux de discrétisation de chaque variable](#_bookmark107) 51
  22. [Feuilles du fichier Excel nommé projectdata ACM](#_bookmark109) 51
  23. [Tableau disjonctif complet](#_bookmark111) 52
  24. [Fréquence de chaque modalité](#_bookmark113) 52
  25. [ACM sur le tableau disjonctif complet](#_bookmark115) 53
  26. [Nombre d’individus, nombre de variables et nombre des modalités](#_bookmark116) 53
  27. [Valeurs propres, pourcentage d’inertie de chaque valeur propre et cumul](#_bookmark118)

[des pourcentages d’inertie](#_bookmark118) 54

* 1. [Graphique des valeurs propres](#_bookmark120) 54
  2. [Dimension du sous espace](#_bookmark122) 55
  3. [ACM sur le tableau disjonctif complet (dimension = 10)](#_bookmark123) 55
  4. [cos2 des modalités sur le sous espace](#_bookmark126) 56
  5. [cos2 des modalités sur le sous espace (suite)](#_bookmark127) 57
  6. [cos2 des modalités sur le sous espace](#_bookmark128) 58
  7. [Modalités bien représentées, moyennement représentées et faiblement re-](#_bookmark130) [présentées sur le sous espace](#_bookmark130) 59
  8. [Contribution des modalités dans chaque axe du sous espace](#_bookmark132) 60
  9. [Contribution des modalités dans chaque axe du sous espace (suite)](#_bookmark133) 61
  10. [Dendogramme de la CAH](#_bookmark135) 62
  11. [Carte des facteurs](#_bookmark136) 62
  12. [Distribution des groupes](#_bookmark137) 63
  13. [Description des modalités par classes](#_bookmark138) 63
  14. [Nuage des modalités projeté sur les 2 premiers axes](#_bookmark140) 64
  15. [cos2 des individus sur le sous espace](#_bookmark143) 65
  16. [cos2 des individus sur le sous espace (suite)](#_bookmark144) 65
  17. [cos2 des individus sur le sous espace (suite)](#_bookmark145) 66
  18. [cos2 des individus sur le sous espace (suite)](#_bookmark146) 66
  19. [cos2 des individus sur le sous espace (suite)](#_bookmark147) 67
  20. [cos2 des individus sur le sous espace (suite)](#_bookmark148) 67
  21. [Individus bien représentés, moyennement représentés et faiblement repré-](#_bookmark150) [sentés sur le sous espace](#_bookmark150) 68
  22. [Individus bien représentés, moyennement représentés et faiblement repré-](#_bookmark151) [sentés sur le sous espace (suite)](#_bookmark151) 68
  23. [Contribution des individus dans chaque axe du sous espace](#_bookmark153) 69
  24. [Contribution des individus dans chaque axe du sous espace (suite)](#_bookmark154) 69
  25. [Contribution des individus dans chaque axe du sous espace (suite)](#_bookmark155) 70
  26. [Contribution des individus dans chaque axe du sous espace (suite)](#_bookmark156) 70
  27. [Dendogramme de la CAH](#_bookmark158) 71
  28. [Carte des facteurs](#_bookmark159) 71
  29. [Distribution des groupes](#_bookmark160) 72
  30. [Description des modalités par classes](#_bookmark161) 72
  31. [Coefficients de corrélation des variables avec les projections sur les axes du](#_bookmark164)

[sous espace](#_bookmark164) 73

* 1. [Graphique des coefficients de corrélation des variables avec les projections](#_bookmark166)

[sur le 1er plan factoriel](#_bookmark166) 73

INTRODUCTION

L’**analyse des données** est un sous domaine des statistiques qui se préoccupe de la description de données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données. Ce projet a pour objectif l’application de ces méthodes sur un jeu de données qui sera étudié en détail dans les chapitres suivants.

Le présent travail s’inscrit dans le cadre du projet du module d’analyse de données du troisième semestre. Dans ce rapport, nous allons effectuer en premier lieu une description du jeu de données choisi en suite les différentes opérations effectuées pour préparer les don- nées. Les chapitres 1,2,3 et 4 sont dédiés à l’application de la régression linéaire multiple, classification, analyse en composantes principales (ACP) et analyse des correspondances multiples (ACM) sur le jeu des données choisi.

### Description du jeu de données

Notre jeux des données contient 17 variables quantitatives et 60 observations.

### Les variables explicatives

*Index* l’indice

*A*1 Précipitations annuelles moyennes en pouces

*A*2 Température moyenne de janvier en degrés Fahrenheit

*A*3 Température moyenne de juillet en degrés Fahrenheit

*A*4 Pourcentage de la population du SMSA de 1960 âgée de 65 ans ou plus

*A*5 Taille des ménages, 1960

*A*6 Scolarisation des personnes de plus de 22 ans

*A*7 Ménage avec cuisine complète

*A*8 Population par mile carré dans les zones urbanisées

*A*9 Pourcentage de la population non blanche

*A*10 Pourcentage d’employés de bureau

*A*11 Familles pauvres (revenu annuel inférieur à 3 000 $)

*A*12 Potentiel de pollution relatif des hydrocarbures *A*13 Potentiel de pollution relatif des oxydes d’azote *A*14 Pollution relative du dioxyde de soufre

*A*15 Pourcentage d’humidité relative, moyenne annuelle à 13 heures.

### La variable dépendante

*B* Taux de mortalité

Le taux de mortalité est représenté en fonction d’autres variables. Soit (*x*1*, x*2*, x*3*, x*4*, x*5*, x*6*, x*7*, x*8*, x*10*, x*11*, x*12*, x*13*, x*14*, x*15) ∈ R15

*B* = *A*1 × *x*1 + *A*2 × *x*2 + *A*3 × *x*3 + *A*4 × *x*4 + *A*5 × *x*5 + *A*6 × *x*6 + *A*7 × *x*7 + *A*8 ×

*x*8 + *A*9 × *x*9 + *A*10 × *x*10 + *A*11 × *x*11 + *A*12 × *x*12 + *A*13 × *x*13 + *A*14 × *x*14 + *A*15 × *x*15

### Préparation des données

Nos données sont sauvegardés dans un fichier *projectdata.csv* chaque colonne est sé- parée par une virgule et chaque ligne est séparée par un retour à la ligne.

### Visualisation du jeu des données

Durant tout ce projet on élimine la variable *Index* de notre jeu des données.

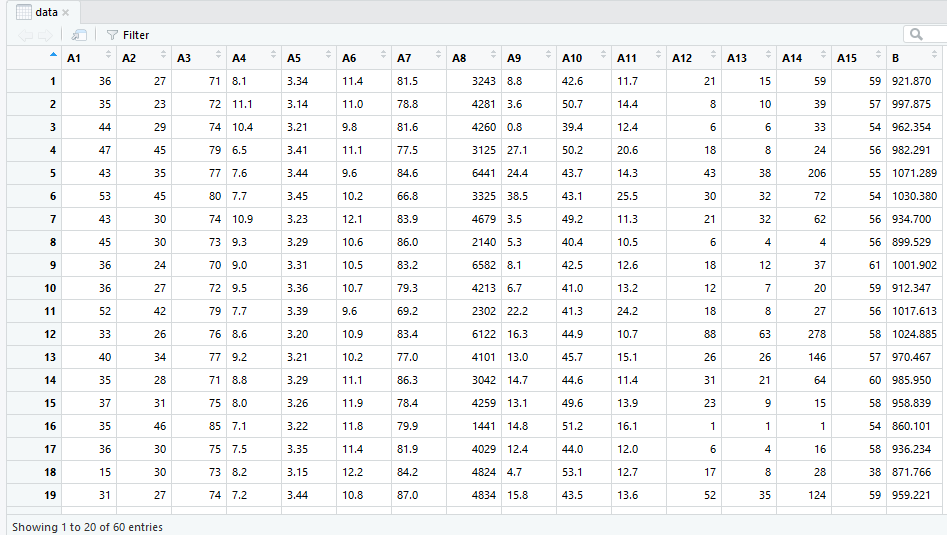


Figure 1: Table de notre jeu des données

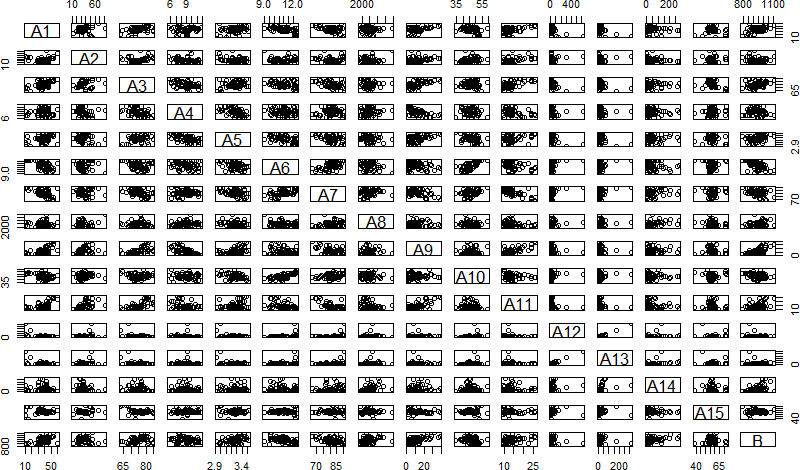


Figure 2: Nuage des points

D’après le nuage des points on déduit que les varibles de notre jeu des données ne sont pas corrélées.

### Correction des types des attributs

Afin de corriger les types des attributs nous avons exploré la structure de nos données.

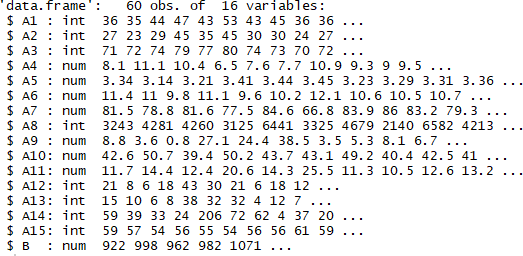


Figure 3: Exploration des attributs de notre jeu des données

On remarque que tout les attributs sont de type correct donc on a pas besoin d’effec- tuer cette étape.

Notre jeu des données (16 variables quantitatives et 60 observations) est maintenant prêt pour les étapes suivantes du projet.

CHAPITRE 1

RÉGRESSION LINÉAIRE MULTIPLE

# Calcule du modèle de régression linéaire multiple

Calculons le modèle de régression linéaire multiple incluant toute les variables expli- catives de notre jeux des données

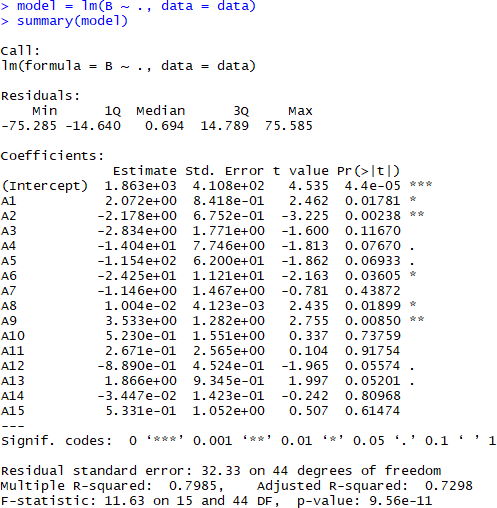


Figure 1.1: Résultat du modèle avec toutes les variables explicatives

## Variables explicatives non significatives

Les variables explicatives non significatives ont une probabilité critique associé au test

*Pr*(*>* |*t*|) *>* 0*.*05, donc d’après le résultat au-dessus on remarque que *A*4, *A*5, *A*7, *A*10,

*A*11, *A*12, *A*13, *A*14 et *A*15 son des variables explicatives non significatives

## Valeur de R2 et Raju2

D’après le résultat de la figure 1.1, on déduit :

*R*2 = 0*.*7985 et *R*2 = 0*.*7298

*ajus*

D’après la valeur de *R*2

*ajus*

## Test de Fisher

on remarque que le modèle explique 73% des données.

D’après le résultat de la figure 1.1, on déduit :

*p* − *value* = 9*.*56*e−*11

La probabilité critique associé au test *p value <* 0*.*05, donc le test de Fisher est significatif.

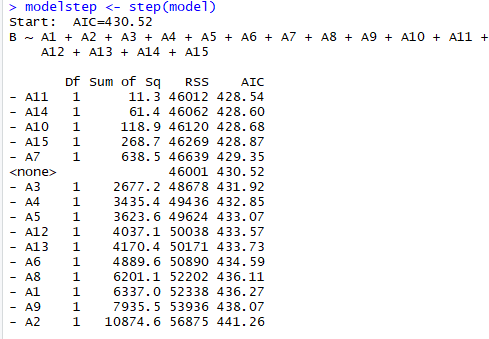
−

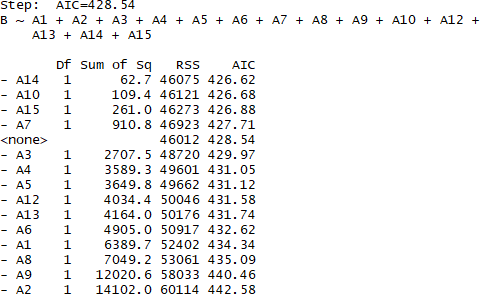
Le test de Fisher signifie qu’il existe au moins une variable significativement non nulle.

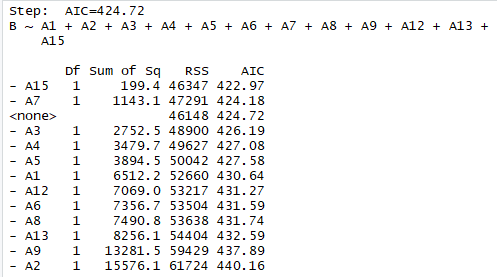
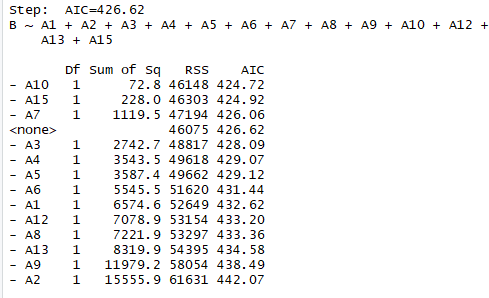
## Conclusion

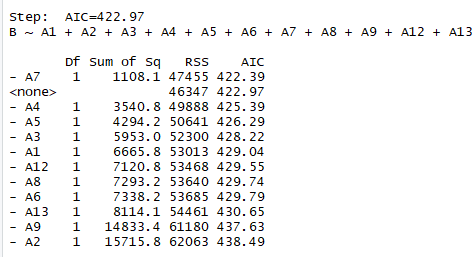
Les résultats montrent que le modèle initiale calculé par toutes les variables a besoin d’une amélioration pour donner des meilleurs résultats.

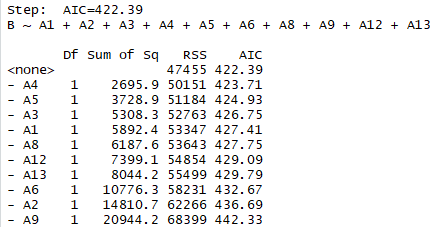
# Amélioration du modèle initiale par la procédure step

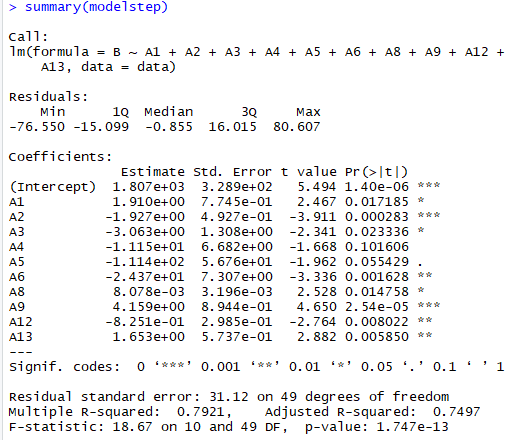












### Remarque

Après avoir améliorer le modèle initiale par la procédure step, nous avons remarqué

que la valeur de *R*2 est passée de 0*.*7298 à 0*.*7497 c’est à dire que le modèle explique

*ajus*

75% des données, donc il y a une amélioration au niveau du modèle.

## Tests de validation

### Test d’homoscédasticité

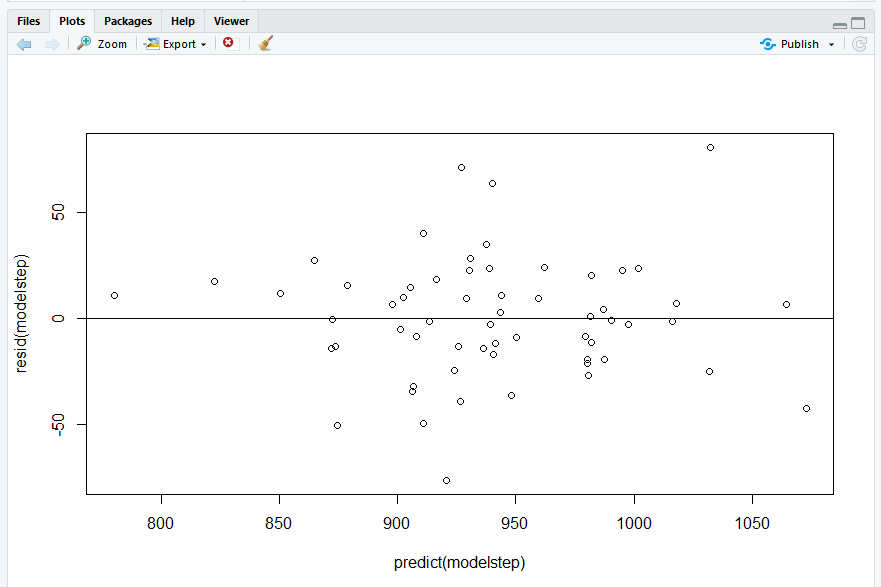


Figure 1.2: Test d’homoscédasticité

D’après la figure 1.2, on remarque que la distribution des points est assez homogène.

### Test de normalité : shapiro

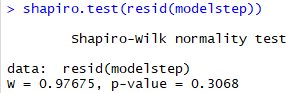


Figure 1.3: Test de normalité : shapiro

D’après la figure 1.3, on remarque que *p value >* 0*.*05, donc la normalité des résidus est acceptée par Shapiro.

−

### Test de normalité : ks

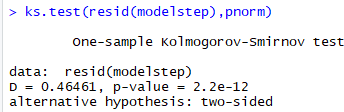


Figure 1.4: Test de normalité : ks

D’après la figure 1.4, on remarque que *p value <* 0*.*05, donc la normalité des résidus n’est pas acceptée par ks.

−

### Recherche de valeurs aberrantes

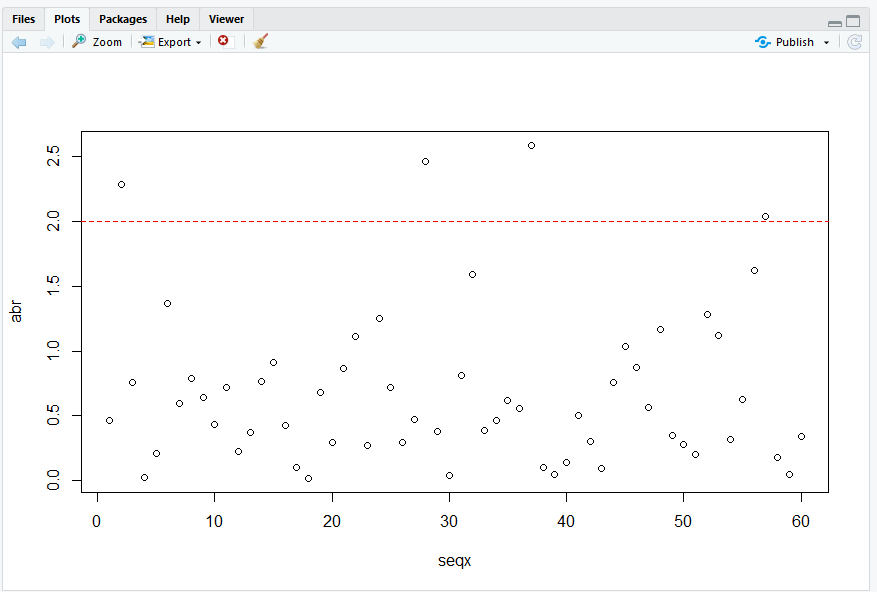


Figure 1.5: Recherche de valeurs aberrantes D’après la figure 1.5, on remarque que 4 valeurs sont aberrantes.

## Conclusion

D’après les résultats précédents on obtient :

* + - * Les variables les plus explicatives de notre modèle obtenu par la procédure step sont :

*A*1, *A*2, *A*3, *A*4, *A*5, *A*6, *A*8, *A*9, *A*12 et *A*13

* + - * 75% des données sont expliqués par le modèle (*R*2 = 0*.*7497)

*ajus*

* + - * Le modèle est relativement homogène (test d’homoscédasticité)
      * La normalité des résidus est accepté par Shapiro (test de normalité : shapiro)
      * Un faible nombre des valeurs aberrante (4 valeurs sont aberrantes)

On conclut donc que le modèle obtenu après l’amélioration par la méthode step est valide.

# Application de la méthode pas à pas de sélection des variables explicatives

On applique la méthode pas à pas de sélection des variables explicatives, basée sur le test de Fisher :

On commence par intégrer la variable la plus significative (*F* le plus grand)

**Début :**

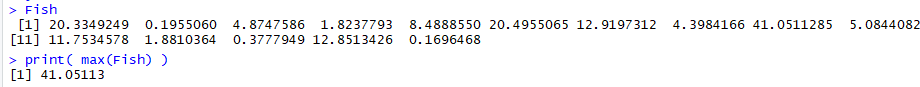
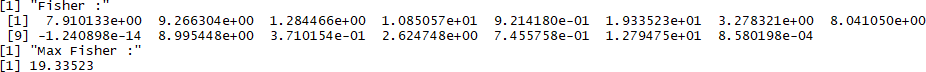


Figure 1.6: La variable la plus significative (*F* le plus grand) Dans notre cas la variable la plus significative est *A*9.

A chaque étape :

* La variable entrante est celle qui présente le plus grand *F* avec *p* − *value <* 10%.
* La variable sortante est celle qui présente le plus petit *F* avec *p* − *value >* 10%.

**Etape** 1 **:**



La variable la plus significative (*F* le plus grand) parmis toutes les variables pour cette étape est *A*6.

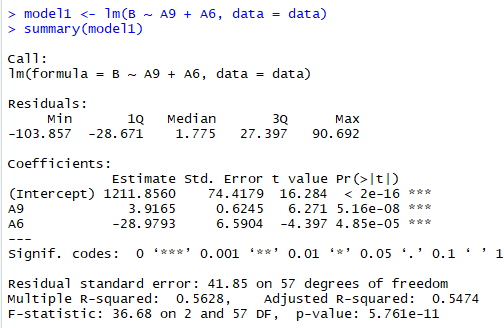


Figure 1.7: Sommaire du modèle 1

*A*6 a une *p* − *value* = 4*.*85*e−*05 *<* 10%

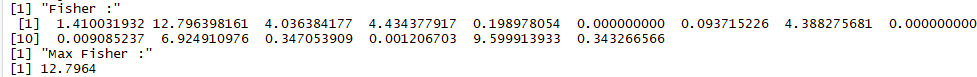
Donc la variable entrante pour l’étape 1 est *A*6.



La variable la moins significative (*F* le plus petit) parmis *A*9 et *A*6 pour cette étape est *A*6.

*A*6 a une *p* − *value* = 4*.*85*e−*05 *<* 10% donc *A*6 n’est pas une variable sortante.

**Etape** 2 **:**



La variable la plus significative (*F* le plus grand) parmis toutes les variables pour cette étape est *A*2.

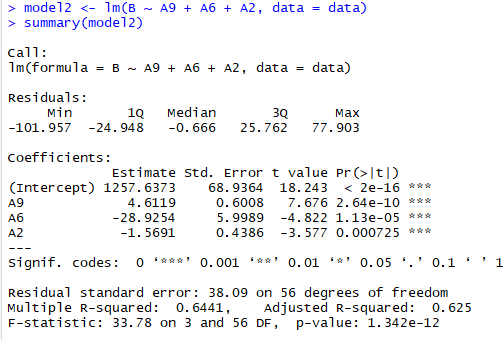


Figure 1.8: Sommaire du modèle 2

*A*2 a une *p* − *value* = 0*.*000725 *<* 10%

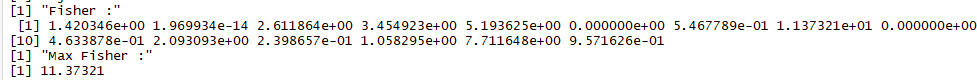
Donc la variable entrante pour l’étape 2 est *A*2.



La variable la moins significative (*F* le plus petit) parmis *A*9 et *A*6 et *A*2 pour cette étape est *A*2.

*A*2 a une *p* − *value* = 0*.*000725 *<* 10% donc *A*2 n’est pas une variable sortante.

**Etape** 3 **:**



La variable la plus significative (*F* le plus grand) parmis toutes les variables pour cette étape est *A*8.

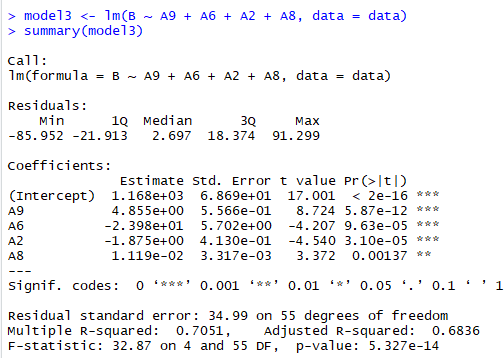
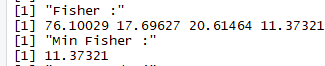


Figure 1.9: Sommaire du modèle 3

*A*8 a une *p* − *value* = 0*.*00137 *<* 10%

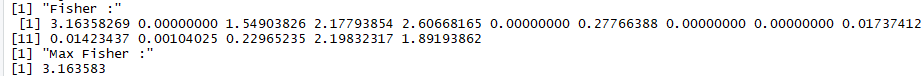
Donc la variable entrante pour l’étape 3 est *A*8.



La variable la moins significative (*F* le plus petit) parmis *A*9, *A*6, *A*2 et *A*8 pour cette étape est *A*8.

*A*8 a une *p* − *value* = 0*.*00137 *<* 10% donc *A*8 n’est pas une variable sortante.

**Etape** 4 **:**



La variable la plus significative (*F* le plus grand) parmis toutes les variables pour cette étape est *A*1.

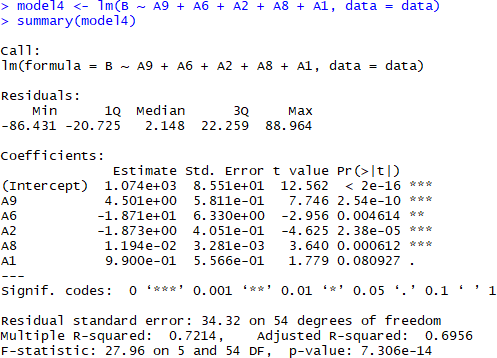


Figure 1.10: Sommaire du modèle 4

*A*1 a une *p* − *value* = 0*.*080927 *<* 10%

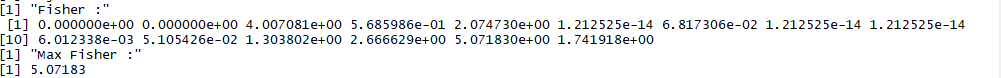
Donc la variable entrante pour l’étape 4 est *A*1.



La variable la moins significative (*F* le plus petit) parmis *A*9, *A*6, *A*2, *A*8 et *A*1 pour cette étape est *A*1.

*A*1 a une *p* − *value* = 0*.*080927 *<* 10% donc *A*1 n’est pas une variable sortante.

**Etape** 5 **:**



La variable la plus significative (*F* le plus grand) parmis toutes les variables pour cette étape est *A*14.

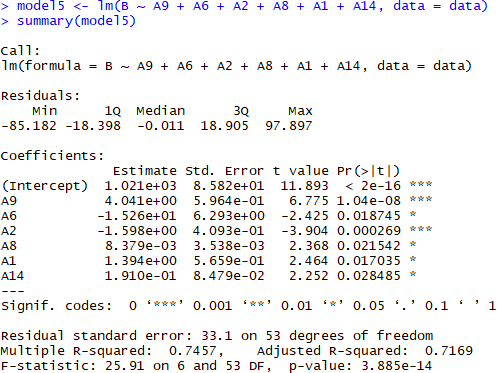


Figure 1.11: Sommaire du modèle 5

*A*14 a une *p* − *value* = 0*.*028485 *<* 10%

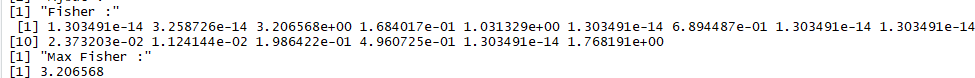
Donc la variable entrante pour l’étape 5 est *A*14.



La variable la moins significative (*F* le plus petit) parmis *A*9, *A*6, *A*2, *A*8, *A*1 et *A*14 pour cette étape est *A*14.

*A*14 a une *p* − *value* = 0*.*028485 *<* 10% donc *A*14 n’est pas une variable sortante.

**Etape** 6 **:**



La variable la plus significative (*F* le plus grand) parmis toutes les variables pour cette étape est *A*3.

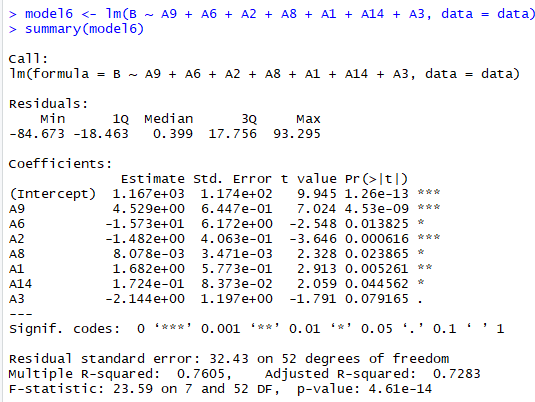


Figure 1.12: Sommaire du modèle 6

*A*3 a une *p* − *value* = 0*.*079165 *<* 10%

Donc la variable entrante pour l’étape 6 est *A*3.

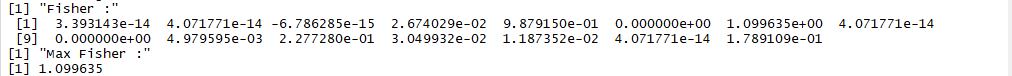


La variable la moins significative (*F* le plus petit) parmis *A*9, *A*6, *A*2, *A*8, *A*1, *A*14 et

*A*3 pour cette étape est *A*14.

*A*3 a une *p* − *value* = 0*.*079165 *<* 10% donc *A*3 n’est pas une variable sortante.

**Etape** 7 **:**



La variable la plus significative (*F* le plus grand) parmis toutes les variables pour cette étape est *A*7.

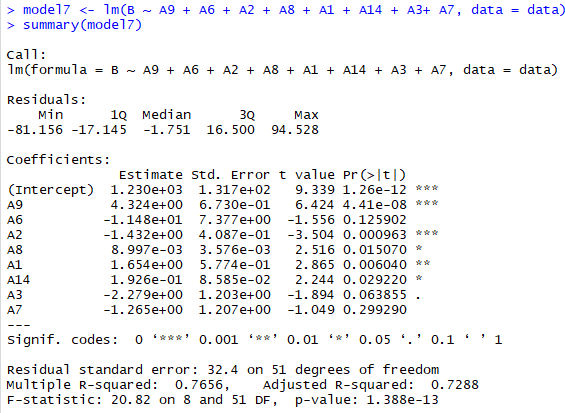


Figure 1.13: Sommaire du modèle 7

*A*7 a une *p* − *value* = 0*.*299290 *>* 10%

Donc la variable *A*7 n’est pas entrante pour l’étape 7.

Donc le modèle contiendra les variables suivantes *A*9, *A*6, *A*2, *A*8, *A*1, *A*14, *A*3.

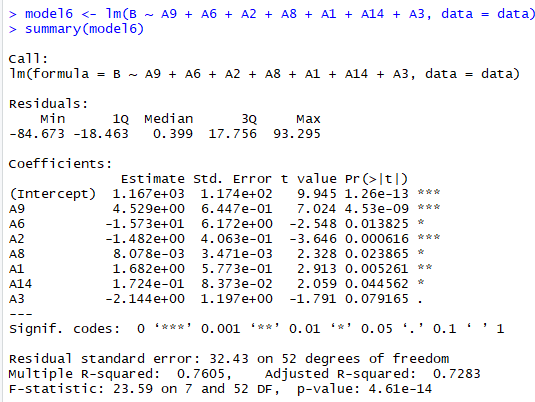


Figure 1.14: Sommaire du modèle 6



La variable la moins significative (*F* le plus petit) parmis *A*9, *A*6, *A*2, *A*8, *A*1, *A*14, *A*3 pour cette étape est *A*3.

*A*3 a une *p* − *value* = 0*.*079165 *<* 10% donc *A*3 n’est pas une variable sortante.

**Arrêt :**

Si les varibales entrantes ont *p* − *values >* 10% et les variables sortantes ont des

*p* − *values <* 10%

La condition d’arrêt est atteinte donc on s’arrête.

Le tableau suivant résume toutes les étapes précédentes :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Etape | Var entrée | Var sortie | Rajus ^2 | F | p-value |
| 1 | A6 | - | 0.5474 | 36.68 | 5.761e-11 |
| 2 | A2 | - | 0.625 | 33.78 | 1.342e-12 |
| 3 | A8 | - | 0.6836 | 32.87 | 5.327e-14 |
| 4 | A1 | - | 0.6956 | 27.96 | 7.306e-14 |
| 5 | A14 | - | 0.7169 | 25.91 | 3.885e-14 |
| 6 | A3 | - | 0.7283 | 23.59 | 4.61e-14 |

## Tests de validation

### Test d’homoscédasticité

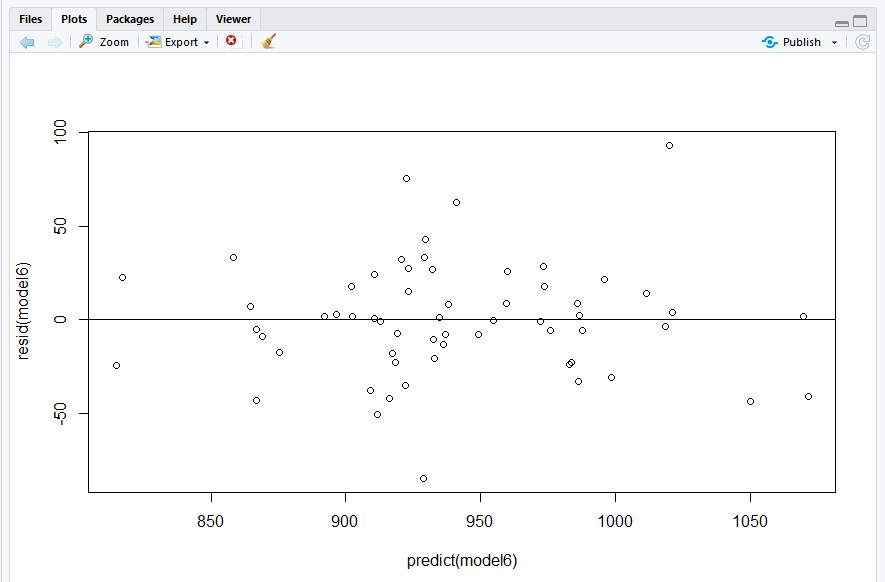


Figure 1.15: Test d’homoscédasticité

D’après la figure 1.15, on remarque que la distribution des points est assez homogène.

### Test de normalité : shapiro

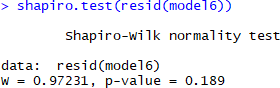


Figure 1.16: Test de normalité : shapiro

D’après la figure 1.16, on remarque que *p value* = 0*.*189 *>* 0*.*05, donc la normalité des résidus est acceptée par Shapiro.

−

### Test de normalité : ks

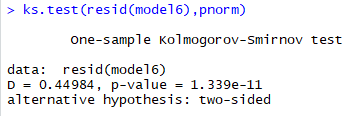


Figure 1.17: Test de normalité : ks

D’après la figure 1.17, on remarque que *p value* = 1*.*339*e−*11 *<* 0*.*05, donc la normalité des résidus n’est pas acceptée par ks.

−

### Recherche de valeurs aberrantes

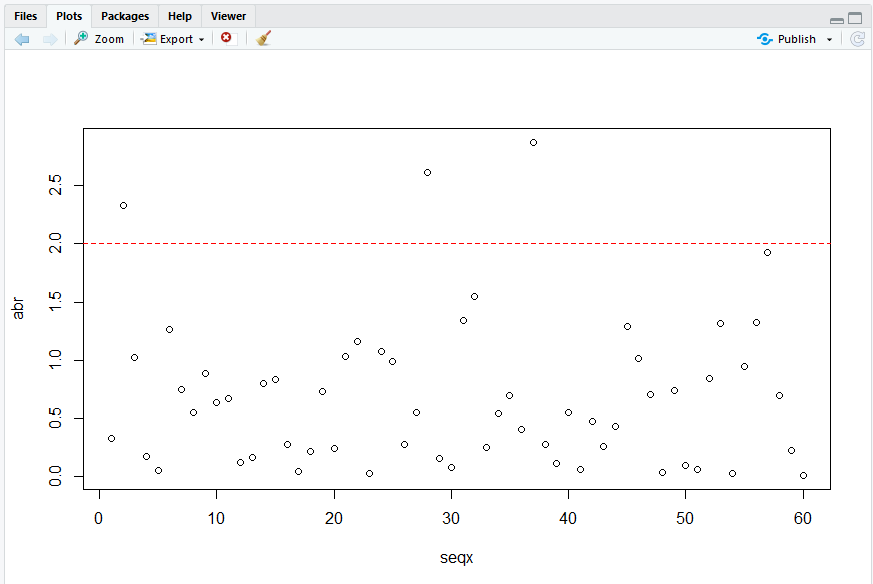


Figure 1.18: Recherche de valeurs aberrantes D’après la figure 1.18, on remarque que 3 valeurs sont aberrantes.

## Critère AIC

La valeur AIC pour le modèle obtenu par la méthode pas à pas de sélection des variables explicatives est :



Pour comparer cette valeur on calcule la valeur AIC du modèle obtenu par la procédure step :



On remarque que la valeur de AIC a augmenté.

## Conclusion

D’après les résultats précédents on obtient :

* + - * Les variables les plus explicatives de notre modèle obtenu par la procédure pas à pas sont : *A*9, *A*6, *A*2, *A*8, *A*1, *A*14 et *A*3
      * 73% des données sont expliqués par le modèle (*R*2 = 0*.*7283)

*ajus*

* + - * Le modèle est relativement homogène (test d’homoscédasticité)
      * La normalité des résidus est accepté par Shapiro (test de normalité : shapiro)
      * Un faible nombre des valeurs aberrante (3 valeurs sont aberrantes)
      * AIC du modèle pas à pas est supérieur de l’AIC du modèle step.

On conclut donc que le modèle obtenu après l’amélioration par la méthode pas à pas est valide.

# Conclusion

## Comparaison entre le modèle obtenu par step et celui ob- tenu par cette méthode de sélection

D’après ce qui précède on a :

* + - * AIC du modèle pas à pas est supérieur de l’AIC du modèle step.
      * 73% des données est expliqué par le modèle obtenu par la procédure pas à pas or pour 75% pour le modèle obtenu par la procédure step.
      * 3 valeurs sont aberrantes pour le modèle obtenu par la procédure pas à pas or 4 valeurs sont aberrantes pour le modèle obtenu par la procédure step.

## Choix entre le modèle obtenu par step et celui obtenu par cette méthode de sélection

D’après la comparasion précèdente on chosit le modèle step vu qu’il est plus efficace.

CHAPITRE 2

CLASSIFICATION

# Variables centrées et réduites

Les varibales centrées et réduites sont sauvgardées dans un fichier ”*projectdata*\_*cr.csv*”.

Les données s’affichent comme suit :

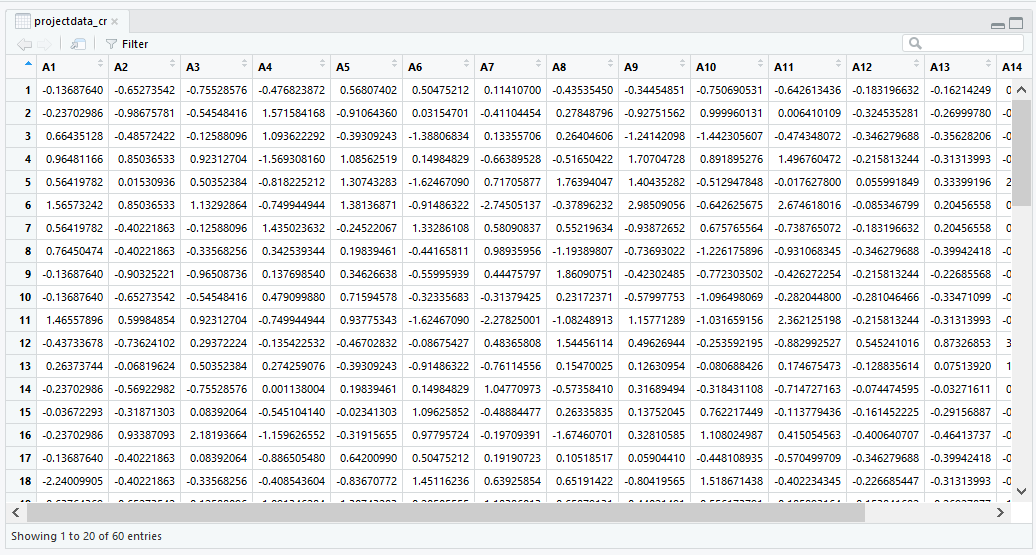


Figure 2.1: Tableau de jeu des données centrées et réduites

Exploration des attributs de notre jeu des données avec variables centrées et réduites.

Les résultats s’affichent comme suit :

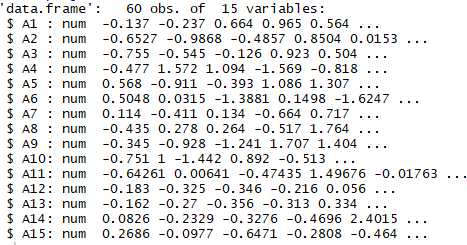


Figure 2.2: Exploration des attributs

On remarque que toute les varibales sont de type correct donc on a pas besoin de corriger le type de ces dernières.

Nos données avec variables centrées et réduites sont prêts pour l’application des algo- rithmes de classification.

# Application de k-means au tableau des variables quantitatives

Soit *k* = 4, on applique le *k*-means au tableau des varibles quantitatives centrées et réduites. On affiche les groupes crées par le *k*-means :

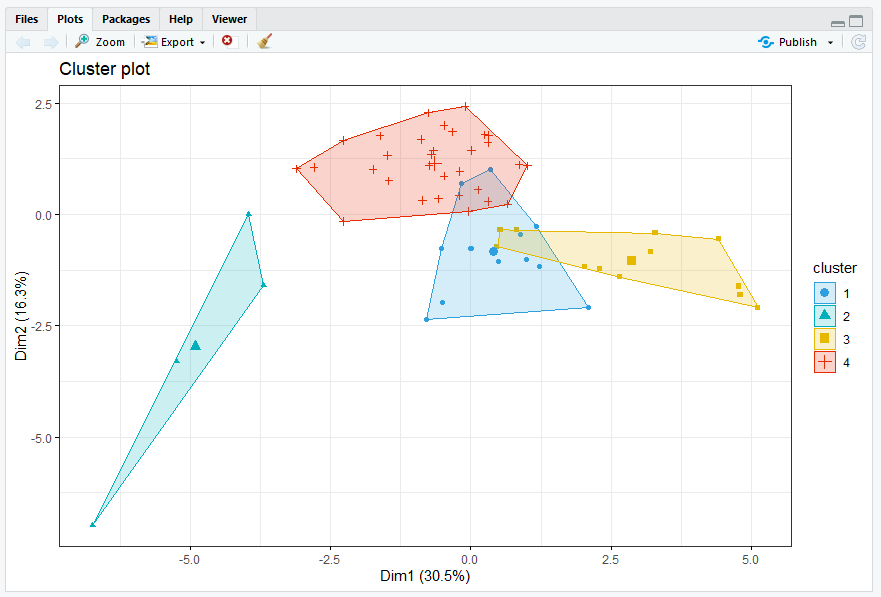


Figure 2.3: Affichage des 4 groupes crées par le *k*-means sur R

Soit *N* = 36, calculons maintenant le taux d’inertie expliquée de la classification pour chaque classe.

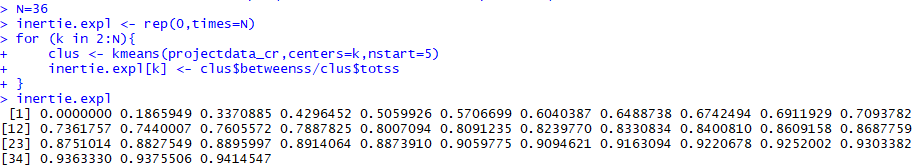


Figure 2.4: Taux d’inertie expliquée pour chaque classe pour *N* = 36

Puisque *N* étant le plus petit entier tel que le taux d’inertie expliquée de la classifica- tion à N classes est supérieur à 0.95 et

0*.*9414547 *<* 0*.*95 donc on pose *N* = 37.

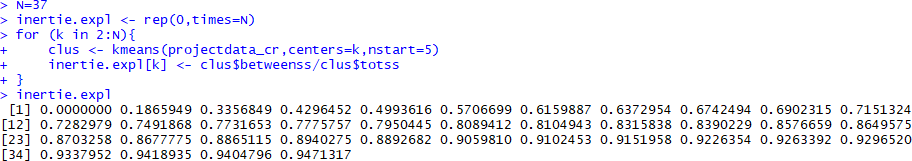


Figure 2.5: Taux d’inertie expliquée pour chaque classe pour *N* = 37 On a 0*.*9471317 *<* 0*.*95 donc *N* = 38

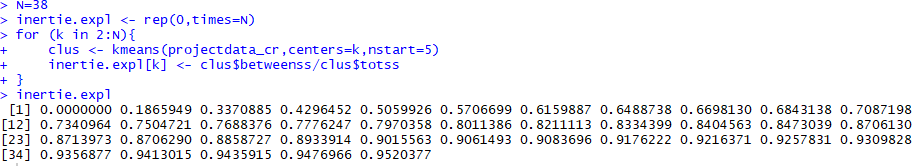


Figure 2.6: Taux d’inertie expliquée pour chaque classe pour *N* = 38 On a 0*.*9520377 *>* 0*.*95 donc *N* = 38

# Détermination de *Nc*

*Nc* est le nombre de classes à retenir en utilisant la méthode de la diapositive 22 du cours. On utilise le critère de taux d’inertie expliquée par la partition. On va surveiller l’évolution de la proportion d’inertie expliquée par la partition, on va chercher le « coude

» dans le graphique donnée.

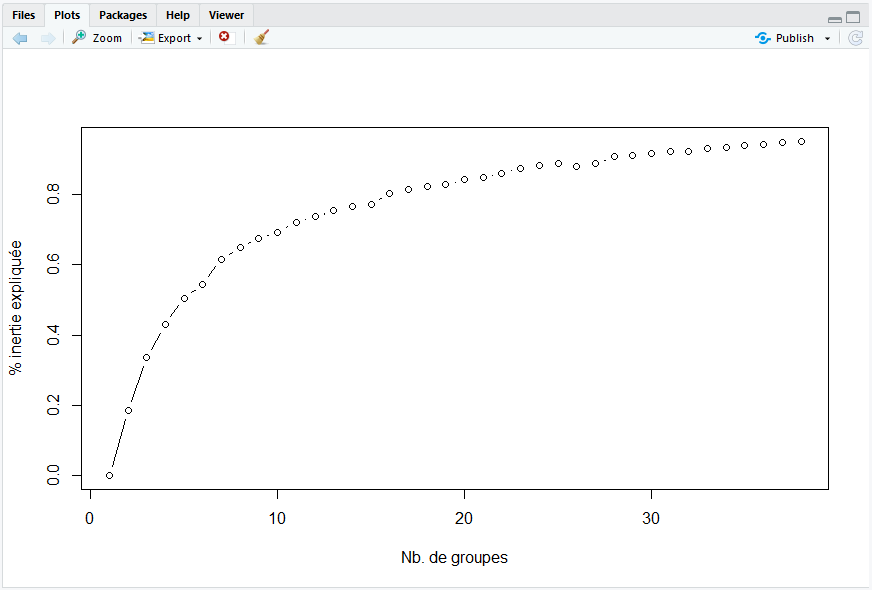


Figure 2.7: Graphique de l’évolution de la proportion d’inertie expliquée par la partition

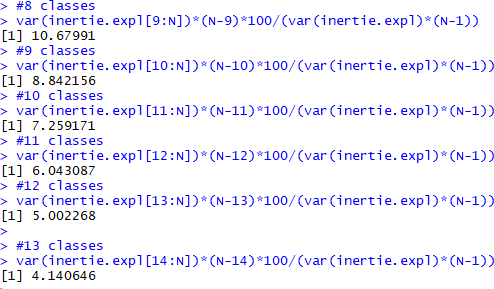


Figure 2.8: Détermination de *Nc*

Donc le nombre à retenir est *Nc* = 13 car 4*.*140646 *<* 5. En appliquant le *k*-means avec *k* = *Nc* = 13 on obtient les résultats suivants :

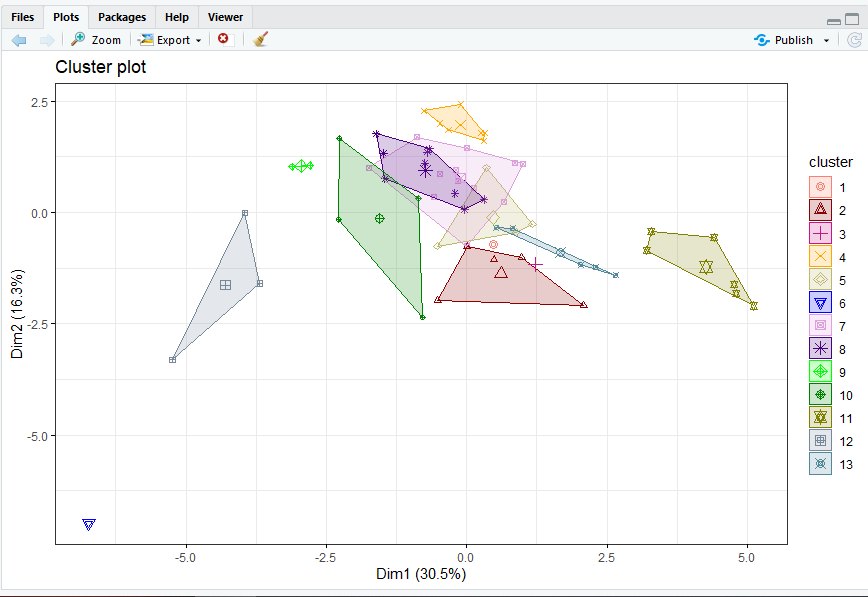


Figure 2.9: Affichage des 13 groupes crées par le *k*-means sur R

# Application de CAH au tableau des variables quan- titatives

On applique le CAH au tableau des varibles quantitatives centrées et réduites. Les résultats se présentent comme suit :

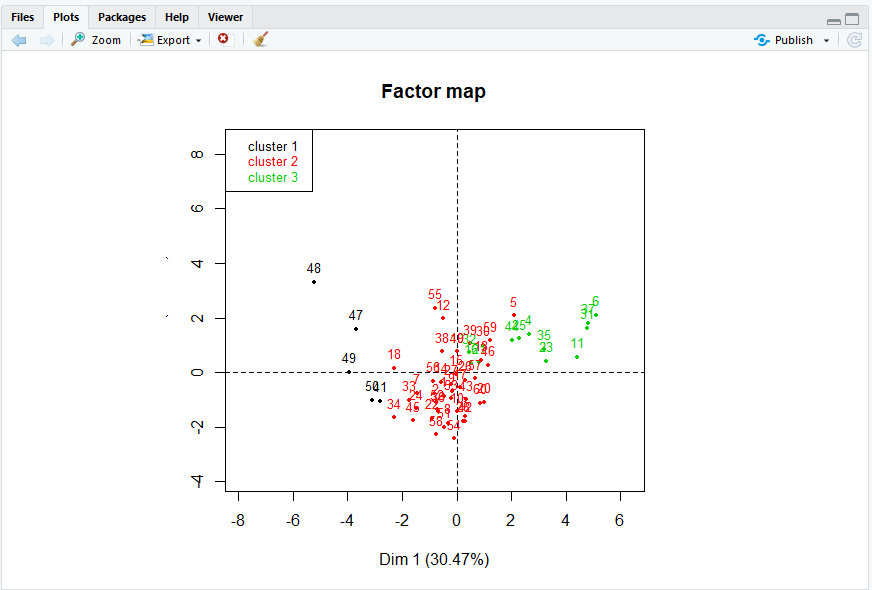


Figure 2.10: Résultats après application de CAH

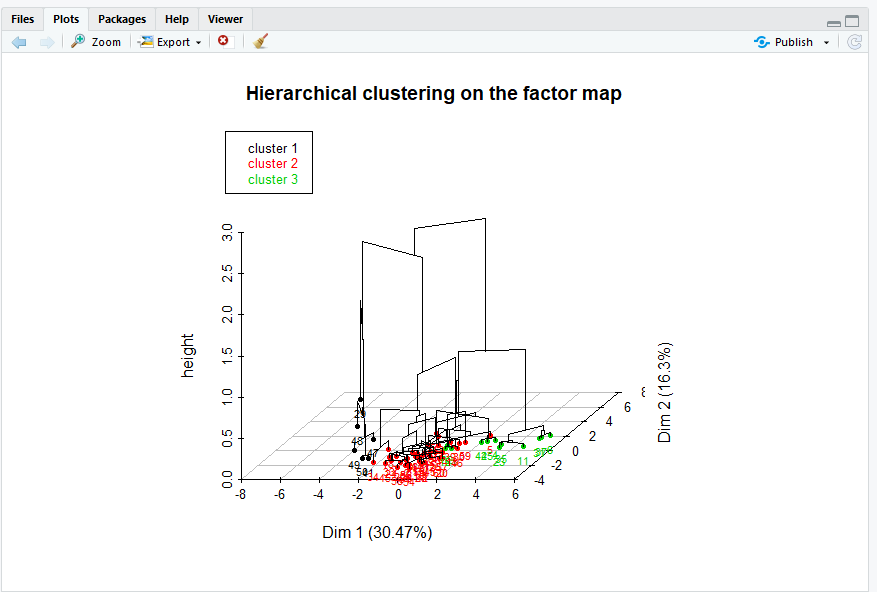


Figure 2.11: Résultats après application de CAH

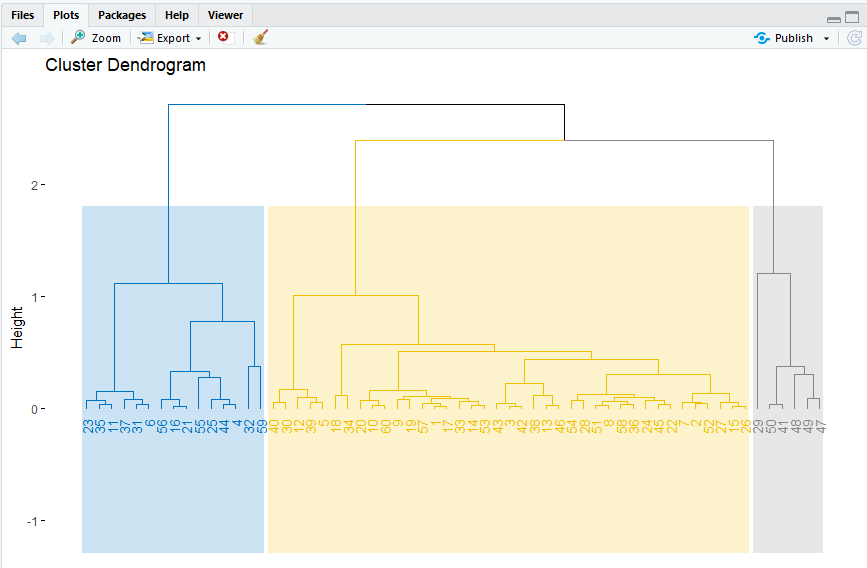


Figure 2.12: Dendrogramme de la CAH Le dendrogramme suggère un découpage en 3 groupes.

# Variables quantitatives les plus corrélées avec la variable classification

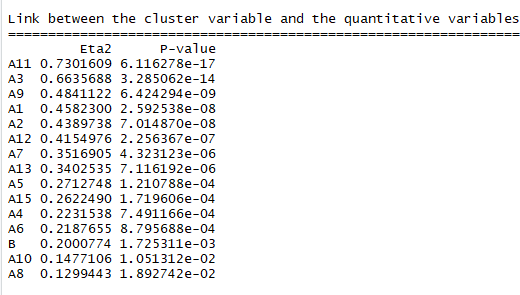


Figure 2.13: Valeur p-value de chaque variable

On conclut que les variables quantitatives les plus corrélées avec la variable classifica- tion sont A11, A3 et A9.

# Description des classes retenues par variables

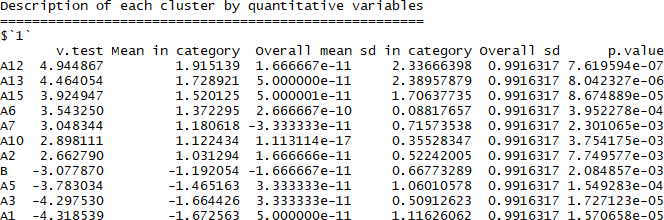


Figure 2.14: Description de la classe 1

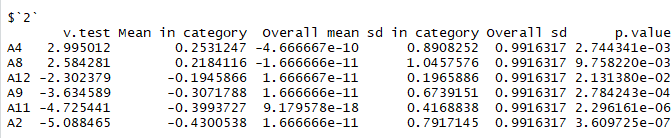


Figure 2.15: Description de la classe 2

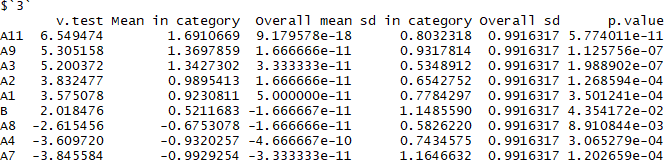


Figure 2.16: Description de la classe 3

# Taux d’inertie

Pour calculer les taux d’inertie : Inertie Inter/Inertie total, avant et après la consoli- dation de la CAH, on utilise les commandes suivantes :

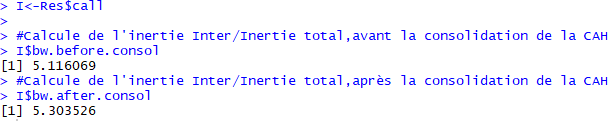


Figure 2.17: Résultats des taux d’inertie/ nertie Inter/Inertie total, avant et après la consolidation de la CAH

# Comparer les classifications faites par k-means et CAH

En appliquant k-means on a obtenu une classification de 13 groupes or CAH a pro- posé de façon optimal une classification de 3 groupes de plus CAH nous laissé la posibilité de choisir le nombre de classes souhaité. Or pour k-means il fallait tout d’abord choisir de façon arbitraire un nombre de groupe qui peut être non optimal, pour cela il fallait trouver le *Nc* en calculant le taux d’intertie.

•

k-means possède un taux d’inertie (95%) plus élevée par rapport à la classification CAH (51%).

•

CHAPITRE 3

ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

# ACP normée sur le tableau des variables quanti- tatives

La normalisation des données se fait automatiquement par la fonction PCA() qui se trouve dans la bibilothèque FactoMineR.

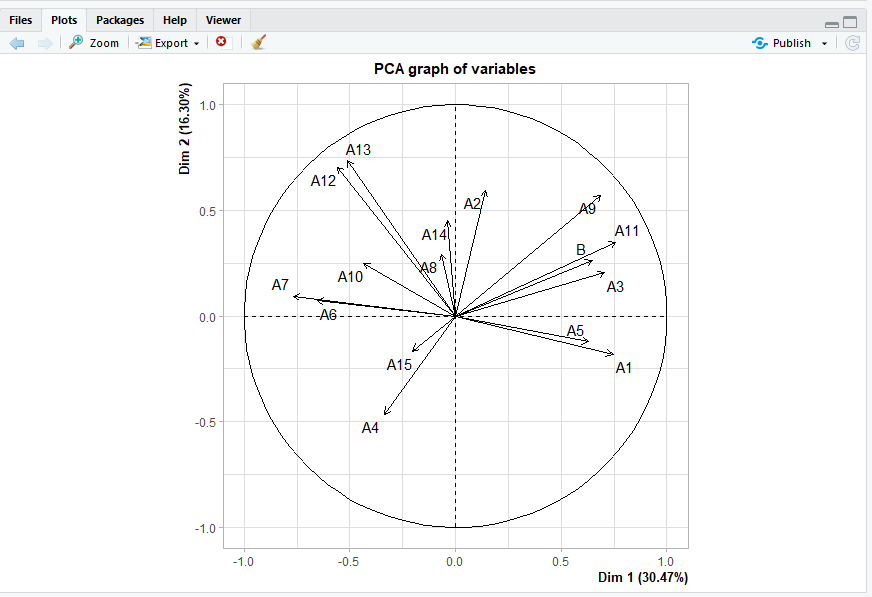


Figure 3.1: Résultat ACP normée sur le tableau des variables quantitatives

### Justification du centrage et réduction des variables

Les raisons pour centrer et réduire les variables :

* + - Rendre les variables qui ont des unités et des niveaux très différents comparables.
    - Rendre le centre du repère au centre de gravité du nuage.
    - Pouvoir traiter les variables de la même manière indépendamment de leurs unités.

Pour éviter le problème des variables à variance importante. Ces variables auront plus de poids que des variables à variance faible.

•

# Indice KMO et les indices MSAI

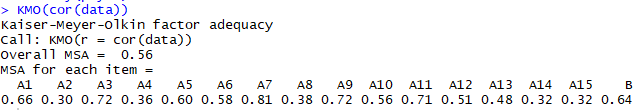


Figure 3.2: Indice KMO et indices MSAI

### Conclusion

D’après la figure 3.2, on conclut que notre échantillon est plutôt « médiocre » (l’in- dice KMO est égale à 0*.*56) ce qui signifie qu’il y a une faible corrélation entre les variables.

Pour les insices MSAI on conclut que :

* + - A2, A4, A8, A13, A14 et A15 sont mal représentés car *MSA <* 0*.*5.
    - A10 et A12 sont moyennement représentés 0*.*5 ≤ *MSA <* 0*.*6.
    - A1,A5,A6 et B sont assez bien représentés car 0*.*6 ≤ *MSA <* 0*.*7.
    - A3, A7, A9 et A11 sont bien représentés car *MSA* ≥ 0*.*7.

# Valeurs propres, pourcentage d’inertie et cumul des pourcentages d’inertie

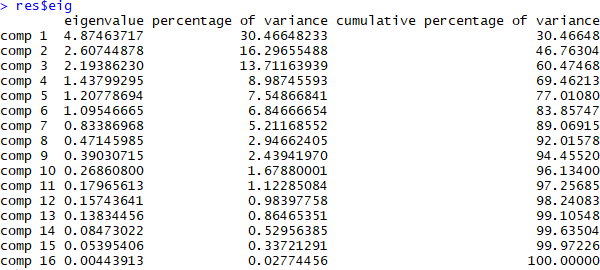


Figure 3.3: Résultat valeurs propres, pourcentage d’inertie de et cumul des pourcentages d’inertie

Les valeurs propres de chaque composante sont calculées dans la première colonne. Le pourcentage d’inertie de chaque valeur propre est calculé dans la deuxième colonne. Le cumul des pourcentages d’inertie est calculé dans la troisième colonne.

# Graphique des valeurs propres

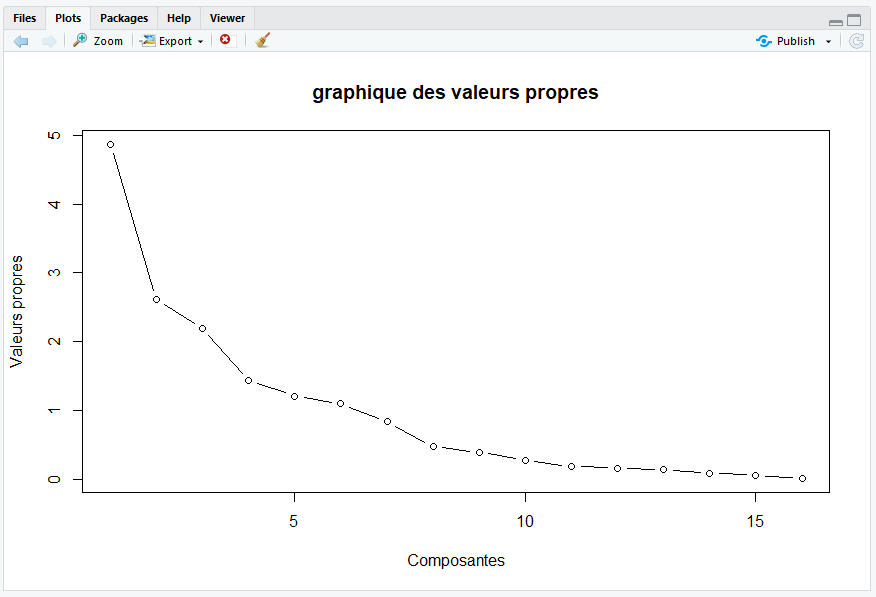


Figure 3.4: Graphique des valeurs propres

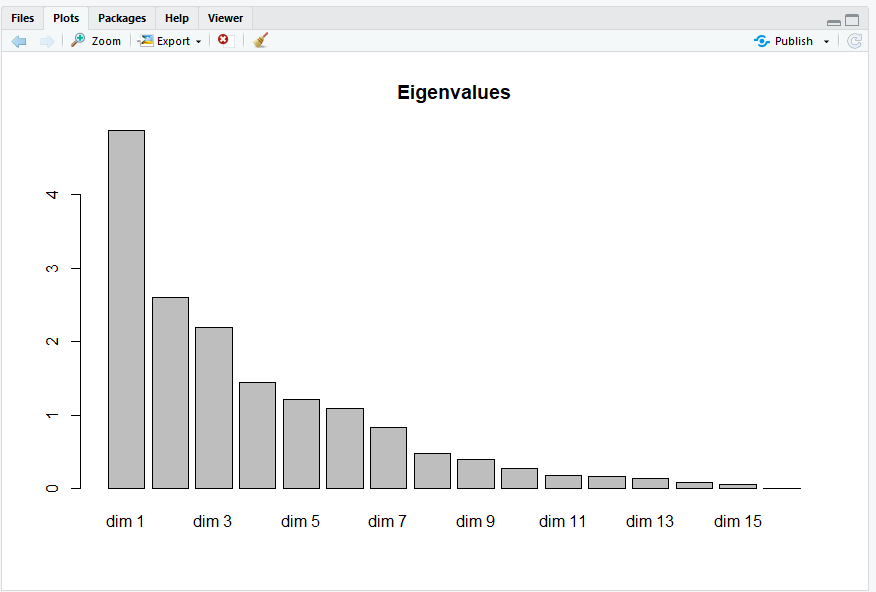


Figure 3.5: Graphique des valeurs propres

D’après les figures 3.4 et 3.5 montrent l’évolution des valeurs propres. On remarque qu’après la 7è*me* composante les valeurs propres deviennent inférieur à 1.

# Dimension du sous espace

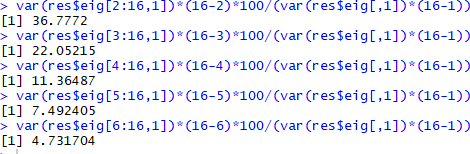


Figure 3.6: Dimension du sous espace 4.731704 < 5 donc on conclut que la dimension du sous espace est 5.

# Nuage des variables

## cos2 des variables sur le sous espace

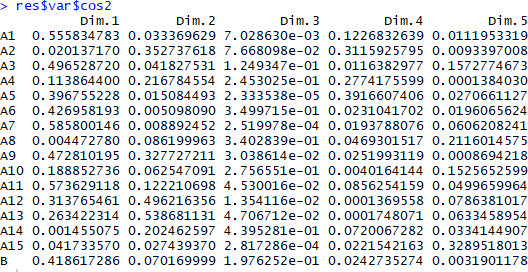


Figure 3.7: cos2 des variables sur le sous espace

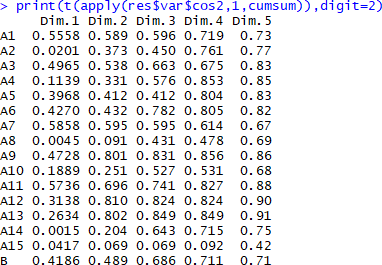


Figure 3.8: cos2 cumulé des variables sur le sous espace

## Variables bien représentées, moyennement représentées et faiblement représentées

Les variables bien représentées sont : A3, A4, A5, A6, A9, A11, A12 et A13. Les variables moyennement représentées sont : A1, A2, A7, A8, A14 et B. Les variables faiblement représentées sont : A15.

## Contribution des variables dans chaque axe du sous espace

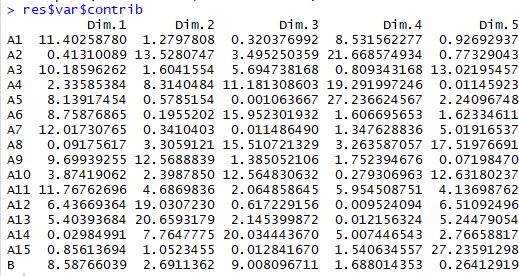


Figure 3.9: Contribution des variables dans chaque axe du sous espace

## Application de la CAH au tableau des contributions des variables aux axes du sous espace

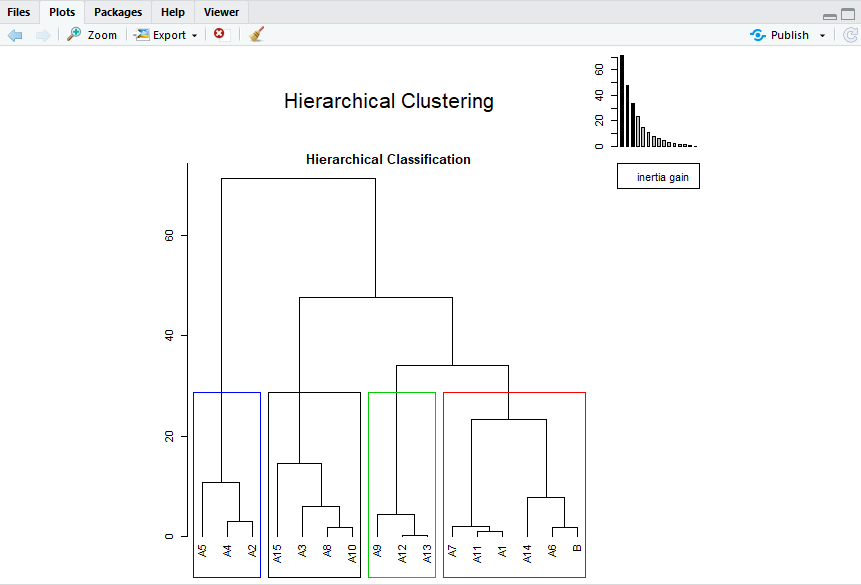


Figure 3.10: Dendrogramme de la CAH

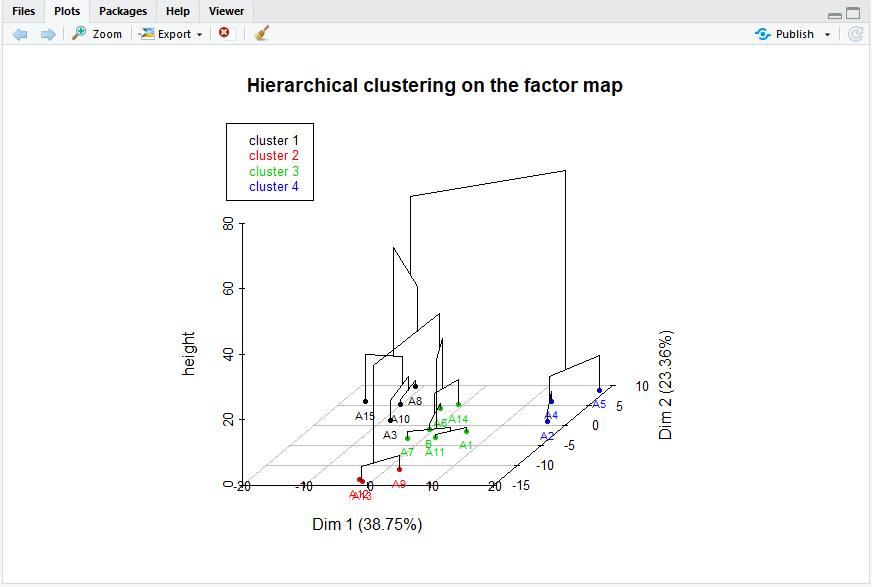


Figure 3.11: Distribution des groupes

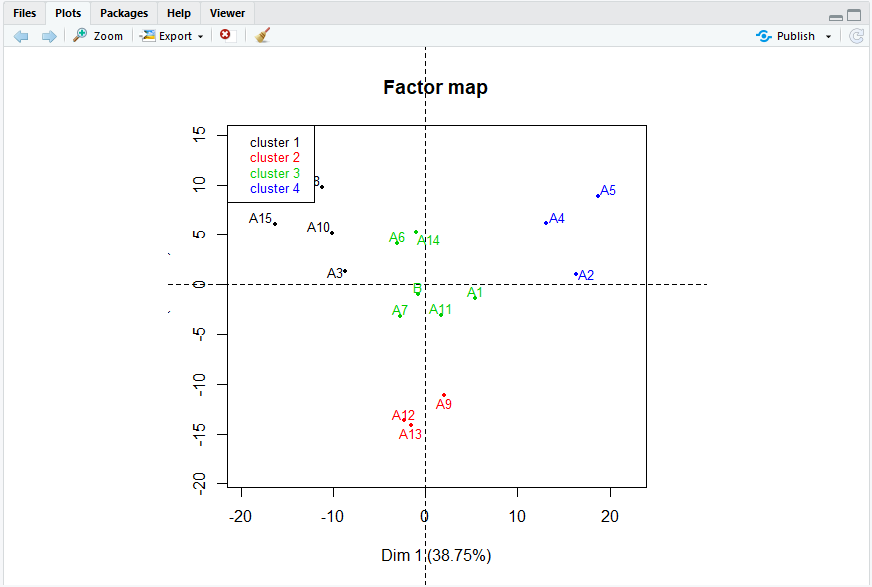


Figure 3.12: Carte des facteurs

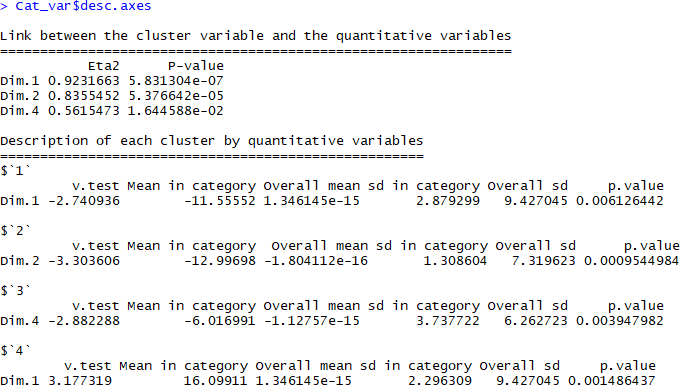


Figure 3.13: Description des axes retenue par classes

### Interprétation des résultats

LA CAH suggère un découpage en 4 groupes. Les variables sont regroupées en 4 classes selon leur contribution aux axes principaux.

Les variables dans le groupes 4 contribuent a l’axe 1.

Les variables dans le groupes 1 ne contribuent pas a l’axe 1. Les variables dans le groupes 2 ne contribuent pas a l’axe 2. Les variables dans le groupes 3 ne contribuent pas a l’axe 4.

## Nuage des variables projeté sur les 2 premiers axes

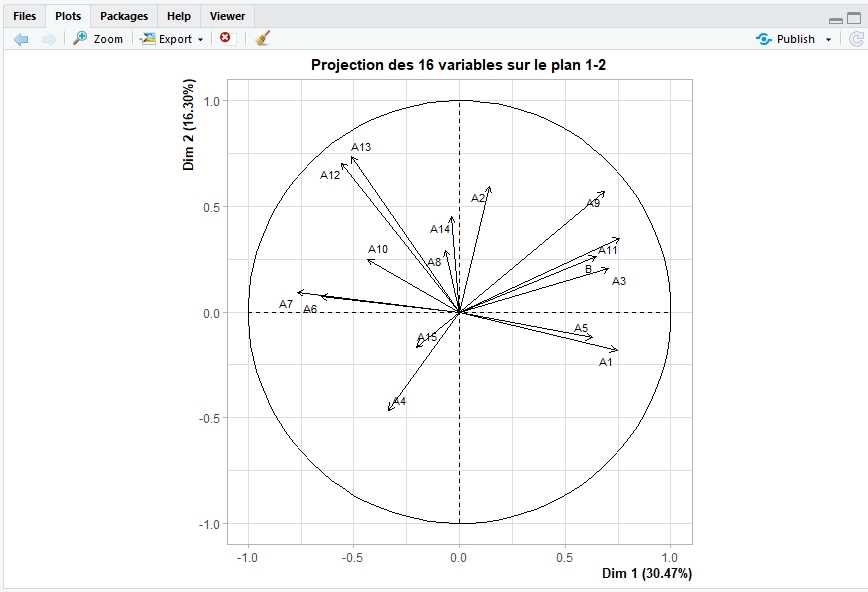


Figure 3.14: Nuage des variables projeté sur les 2 premiers axes

## Variables relativement bien corrélées (positivement et né- gativement) avec les axes du 1er plan factoriel

Les variables relativement bien corrélées positivement avec l’axe 1 : A2, A9, A11, B, A3, A5 et A1.

Les variables relativement bien corrélées négativement avec l’axe 1 : A14, A8, A13, A12, A10, A7, A6, A15 et A4.

Les variables relativement bien corrélées positivement avec l’axe 2 : A3, B, A11, A9, A2, A14, A8, A13, A12, A10, A6 et A7.

Les variables relativement bien corrélées négativement avec l’axe 2 : A15, A4, A1 et A5.

# Nuage des individus

## cos2 des individus sur le sous espace

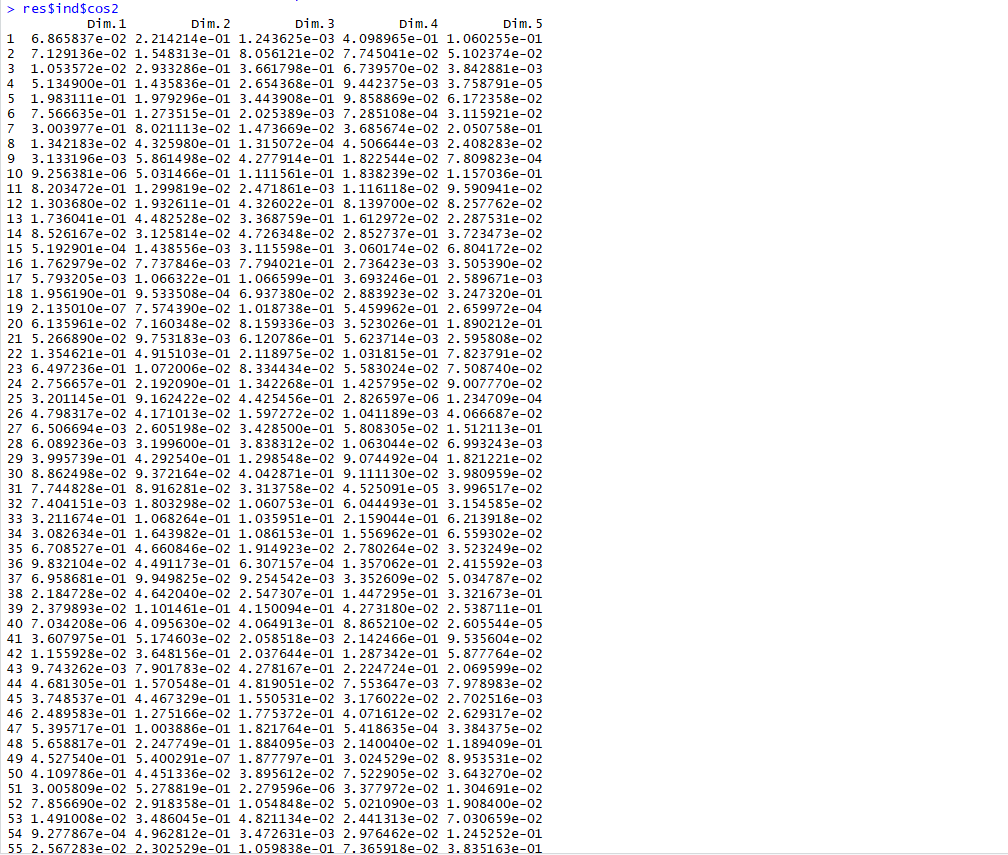


Figure 3.15: cos2 des individus sur le sous espace

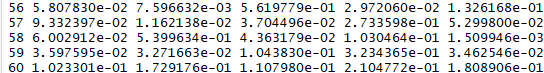


Figure 3.16: cos2 des individus sur le sous espace

## Contribution des individus dans chaque axe du sous espace

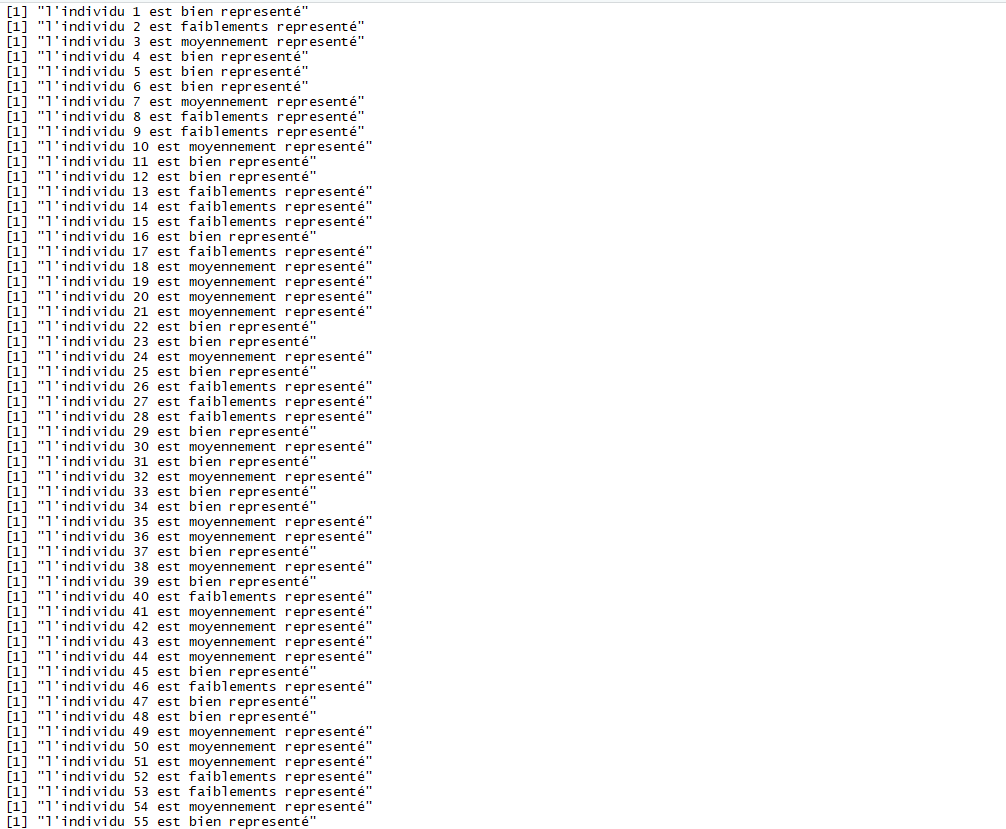


Figure 3.17: Contribution des individus dans chaque axe du sous espace

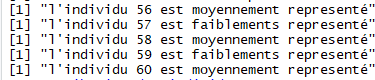


Figure 3.18: Contribution des individus dans chaque axe du sous espace (suite)

## Application la CAH au tableau des contributions des in- dividus aux axes du sous espace

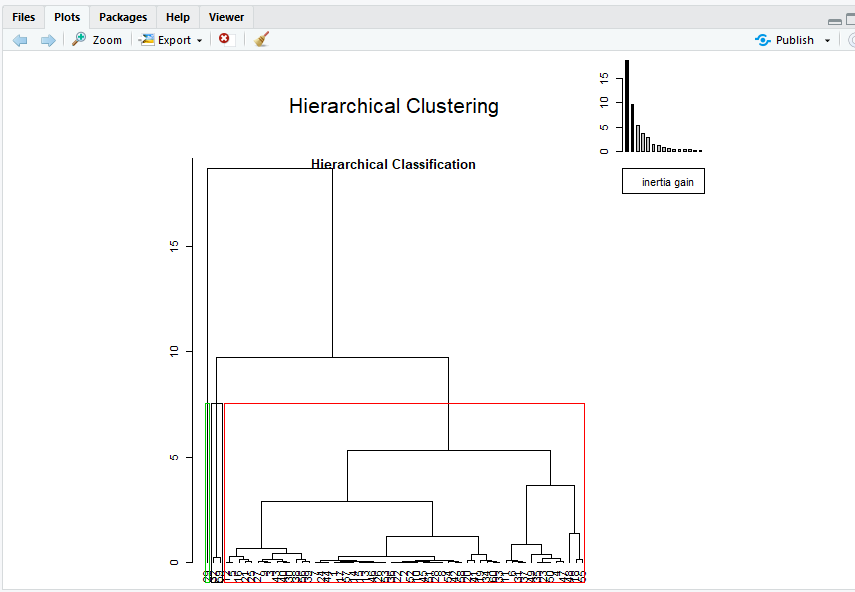


Figure 3.19: Dendrogramme de la CAH

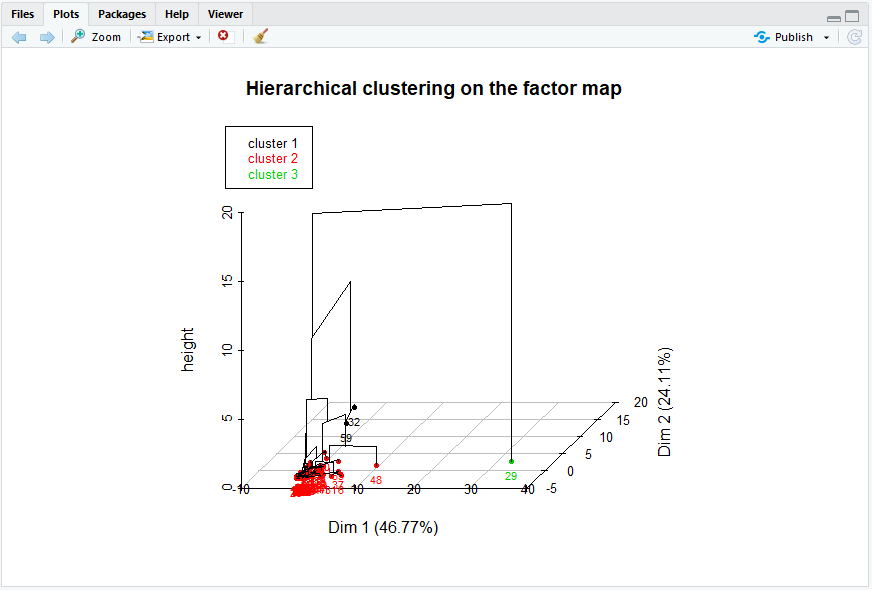


Figure 3.20: Distribution des groupes

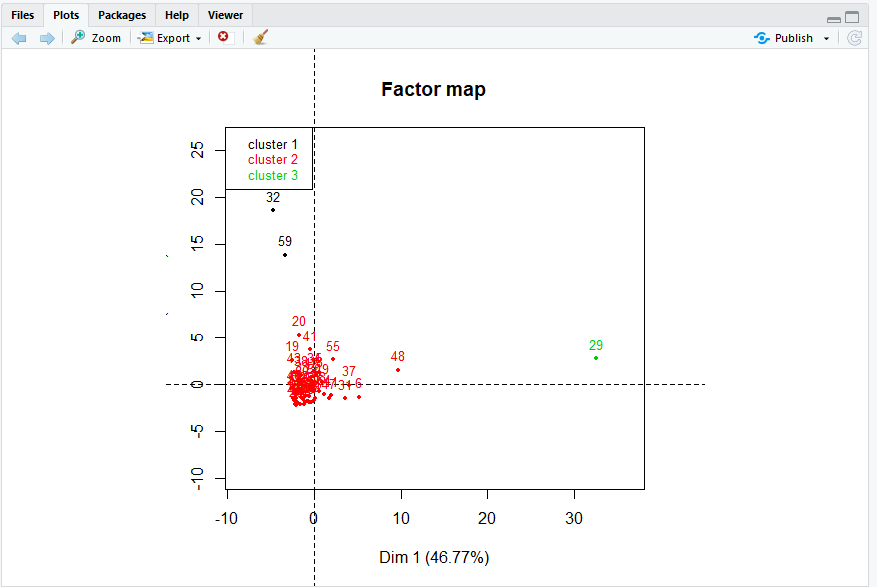


Figure 3.21: Carte des facteurs

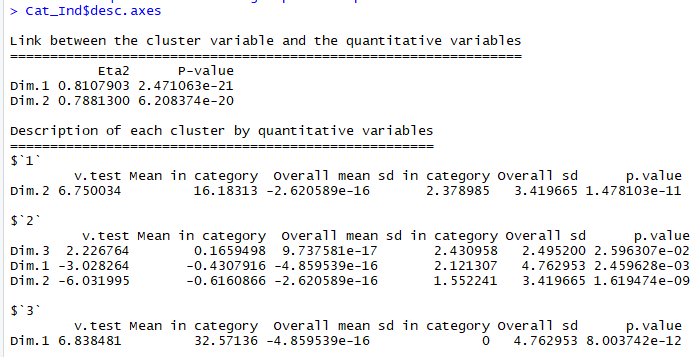


Figure 3.22: Description des axes retenus par classes

### Interprétation des résultats

LA CAH suggère un découpage en 3 groupes. Les individus sont regroupées en 3 classes selon leur contribution aux axes principaux.

Les individus dans le groupes 1 contribuent a l’axe 2.

Les individus dans le groupes 2 ne contribuent pas a l’axe 2 et 1 et contribuent à l’axe 3.

Les individus dans le groupes 3 contribuent a l’axe 1.

## Nuage des individus projeté sur les 2 premiers axes

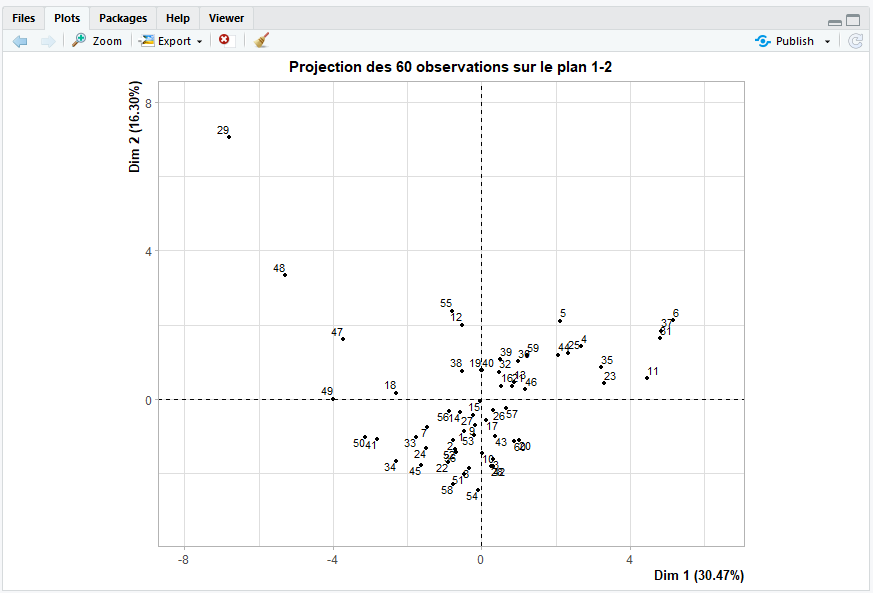


Figure 3.23: Nuage des individus projeté sur les 2 premiers axes

# Conclusion

L’ACP nous a permis de réduire les 16 variables de notre jeu des données en seulement 5 composantes principales. Ceci nous a donné l’opportunité de voir les différentes relations globales existantes entre les variables et les individus de notre jeu des données.

CHAPITRE 4

ANALYSE DES CORRESPONDANCES MULTIPLES (ACM)

# Transformation des variables quantitatives

Si taux de discrétisation est supérieur à 0.5 donc on transforme chaque variable quantitative en 3 modalités.

•

Si le taux de discrétisation est inférieur à 0.5 pour une certaine variable donc on transforme chaque variable quantitative en 4 modalités.

•



Figure 4.1: Taux de discrétisation de chaque variable

On remarque que toutes les variables de notre jeu des données ont un taux de discré- tisation est supérieur à 0.5, donc on transforme chacune en 3 modalités.

# Regroupement dans un fichier Excel, les variables quantitatives et les variables qualitatives et le taux de discrétisation de chaque transformation

Le fichier Excel nommé projectdata ACM regroupe les variables quantitatives et les variables qualitatives et le taux de discrétisation de chaque transformation.

* + - Les variables quantitatives dans la feuille nommée data quanti
    - Les variables qualitatives dans la feuille nommée data quali
    - Le taux de discrétisation de chaque transformation dans la feuille nommée TAUX



Figure 4.2: Feuilles du fichier Excel nommé projectdata ACM

# Tableau disjonctif complet

Les 20 premières valeurs du tableau disjonctif complet :

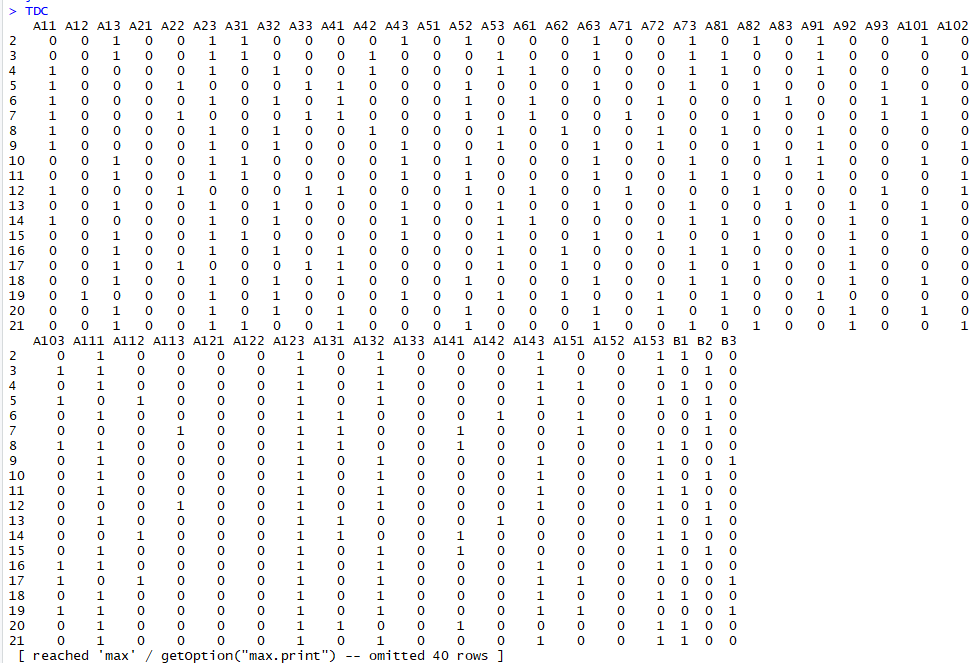


Figure 4.3: Tableau disjonctif complet Le fichier Excel TDC contient toutes les valeurs.

# Fréquence de chaque modalité

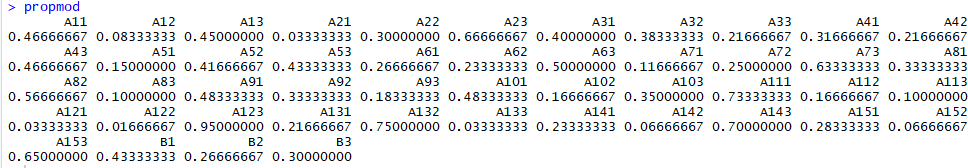


Figure 4.4: Fréquence de chaque modalité

On remarque qu’il y a absence d’une modalité rare (fréquence < 0.01) et donc on ne va rien éliminer.

# Application de l’ACM au tableau disjonctif com- plet

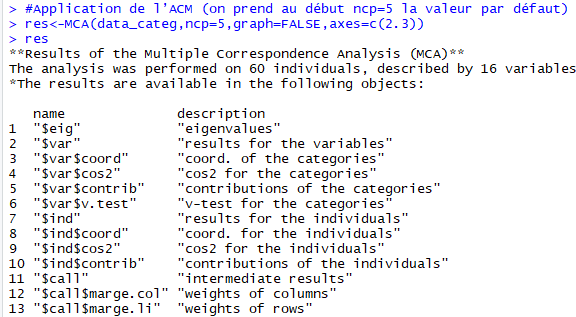


Figure 4.5: ACM sur le tableau disjonctif complet



Figure 4.6: Nombre d’individus, nombre de variables et nombre des modalités

* + - Le nombre d’individus est 60
    - Le nombre de variables 16
    - Le nombre des modalités 48

# Valeurs propres, pourcentage d’inertie de chaque valeur propre et cumul des pourcentages d’inertie

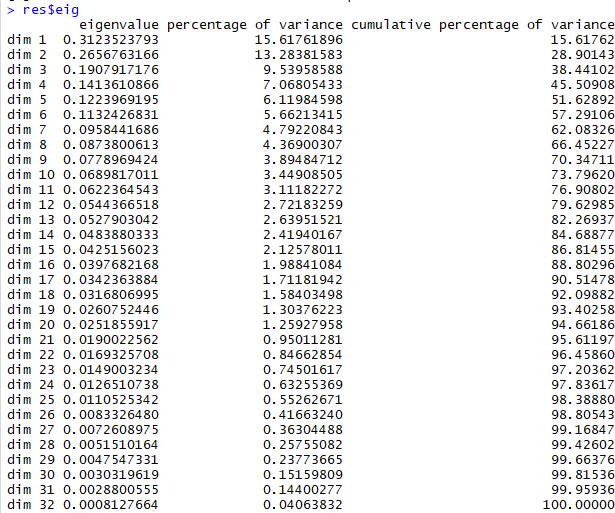


Figure 4.7: Valeurs propres, pourcentage d’inertie de chaque valeur propre et cumul des pourcentages d’inertie

# Graphique des valeurs propres

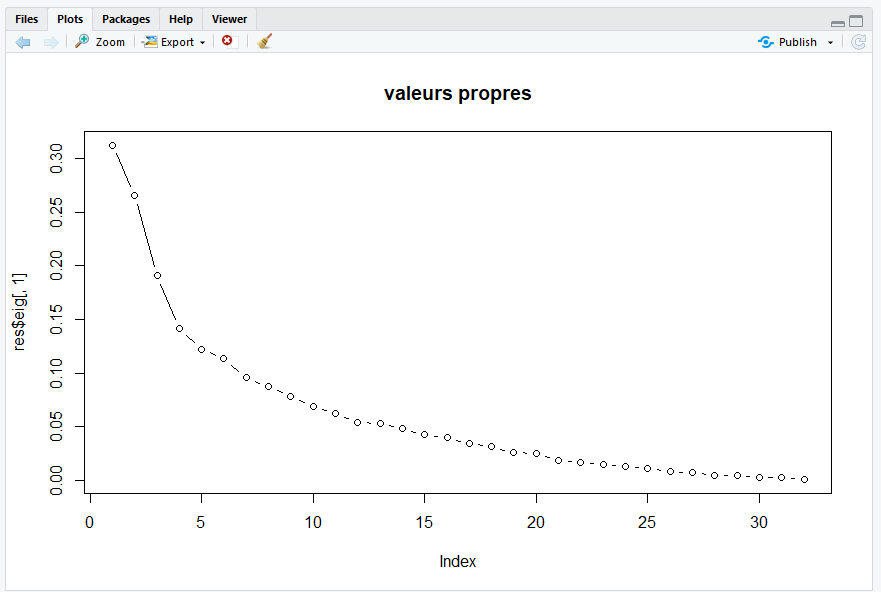


Figure 4.8: Graphique des valeurs propres

# Dimension du sous espace

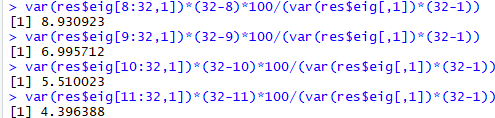


Figure 4.9: Dimension du sous espace

4.396388 < 5 donc on conclut donc que la dimension du sous espace est 10.

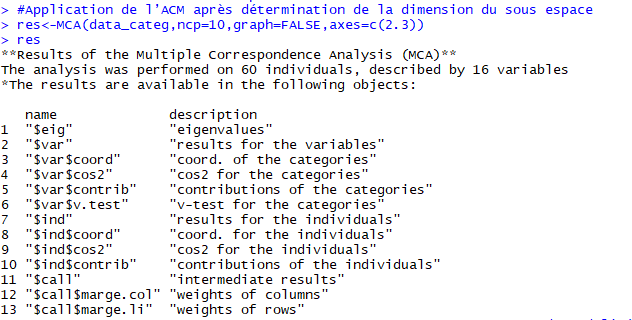


Figure 4.10: ACM sur le tableau disjonctif complet (dimension = 10)

# Nuage de Modalités

## cos2 des modalités sur le sous espace

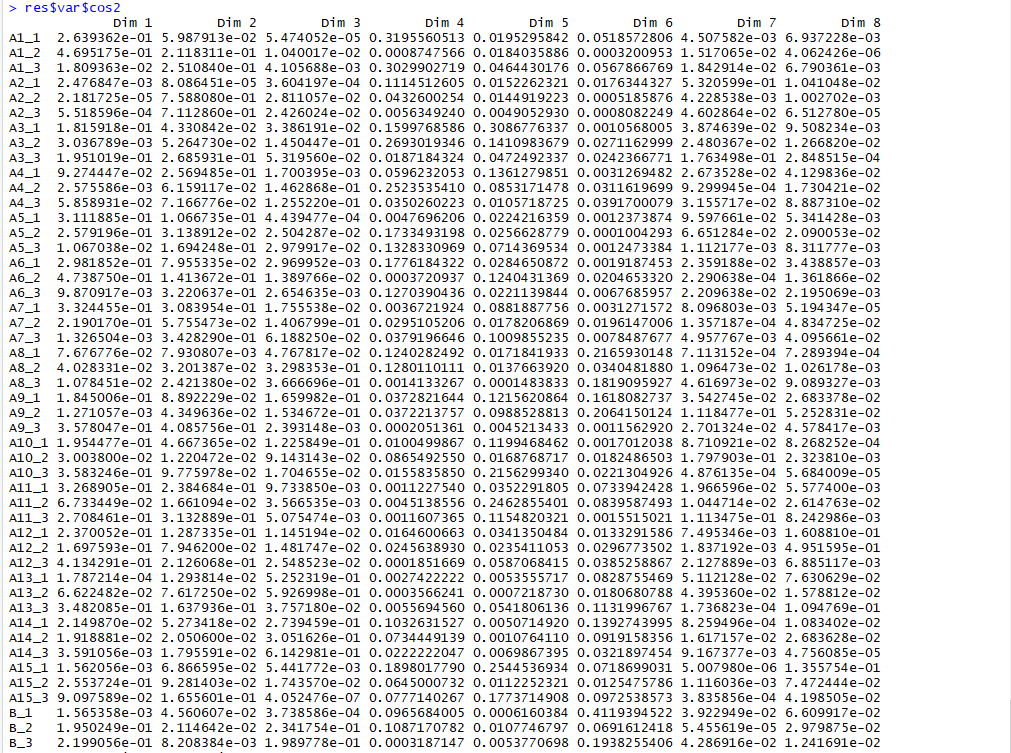


Figure 4.11: cos2 des modalités sur le sous espace

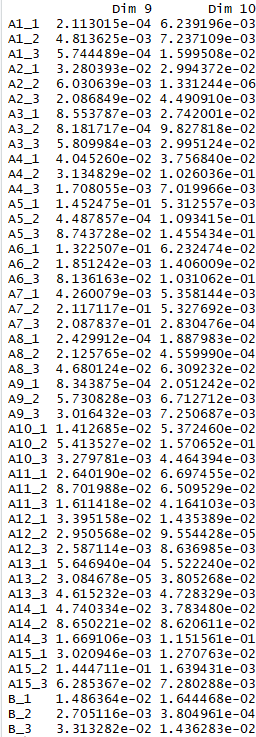


Figure 4.12: cos2 des modalités sur le sous espace (suite)

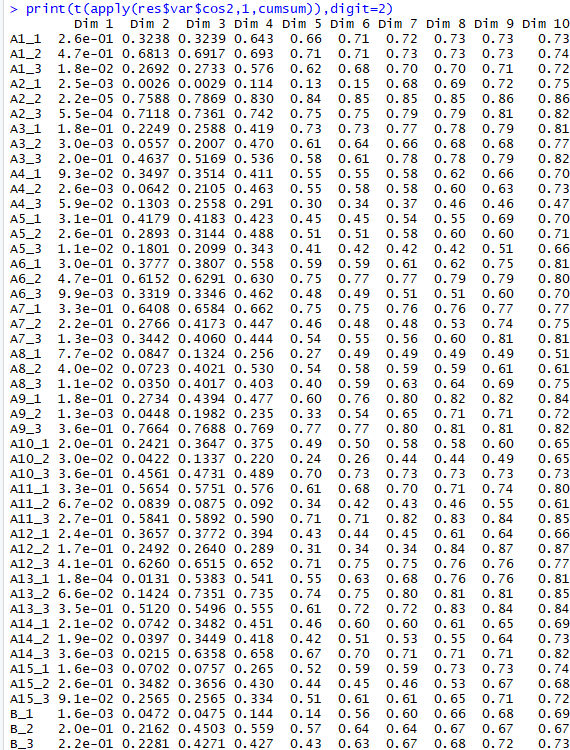


Figure 4.13: cos2 des modalités sur le sous espace

## Modalités bien représentées, moyennement représentées et faiblement représentées sur le sous espace

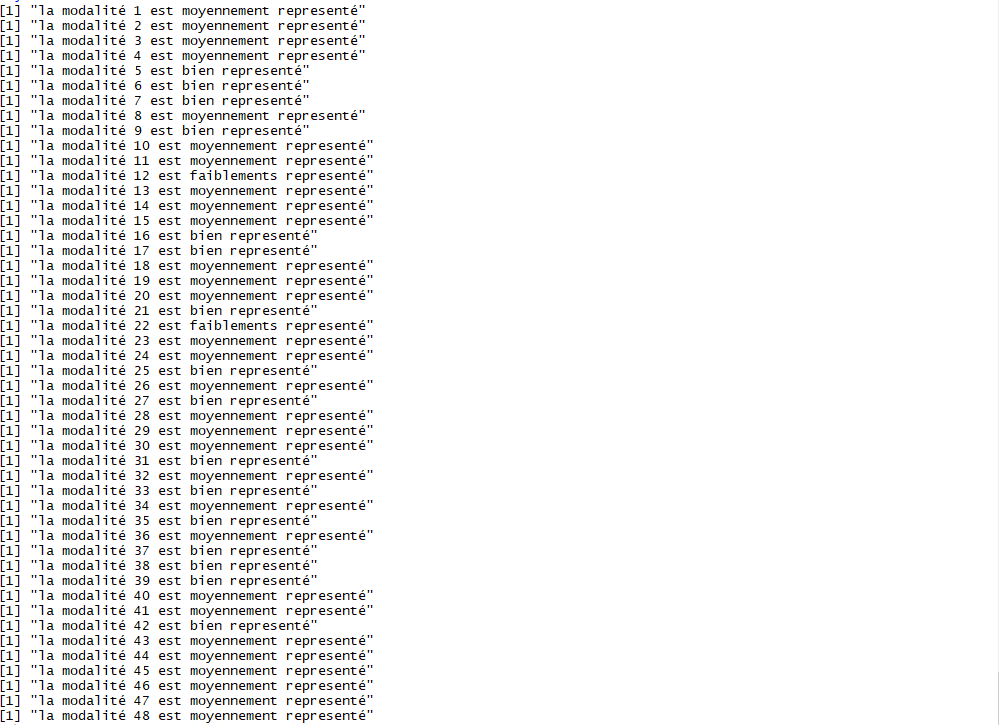


Figure 4.14: Modalités bien représentées, moyennement représentées et faiblement re- présentées sur le sous espace

## Contribution des modalités dans chaque axe du sous es- pace

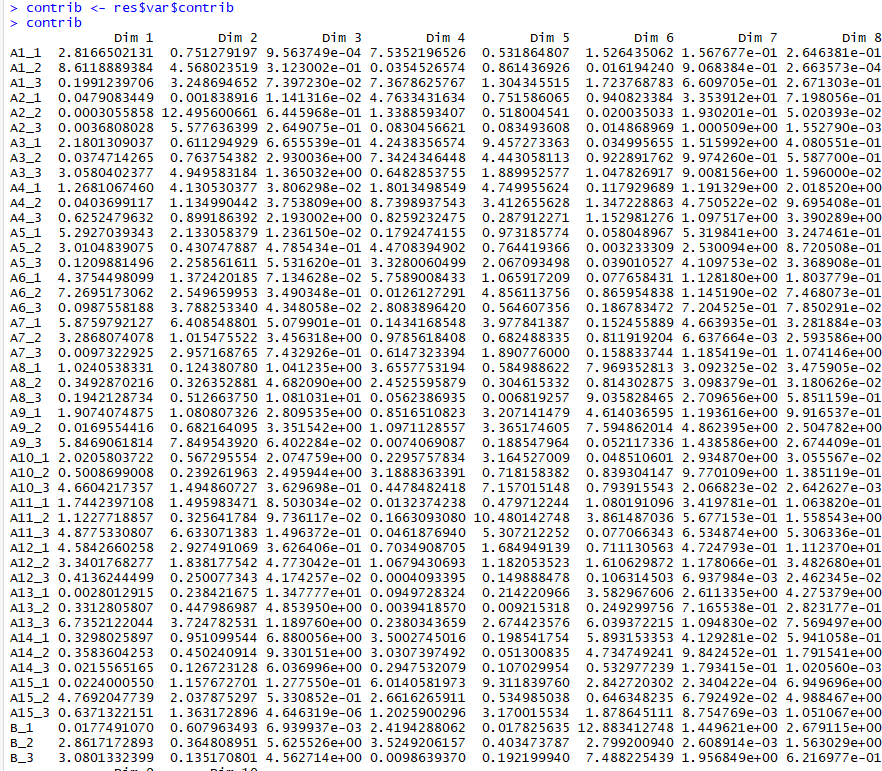


Figure 4.15: Contribution des modalités dans chaque axe du sous espace

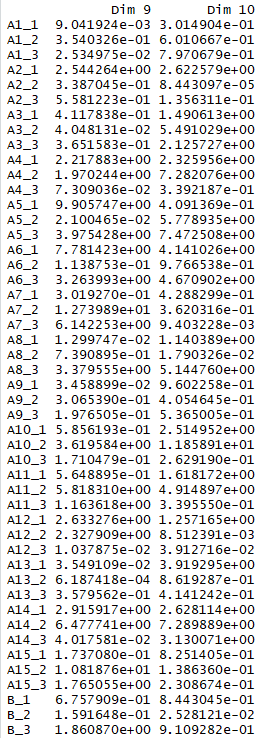


Figure 4.16: Contribution des modalités dans chaque axe du sous espace (suite)

## Application de la CAH au tableau des contributions des modalités aux axes du sous espace

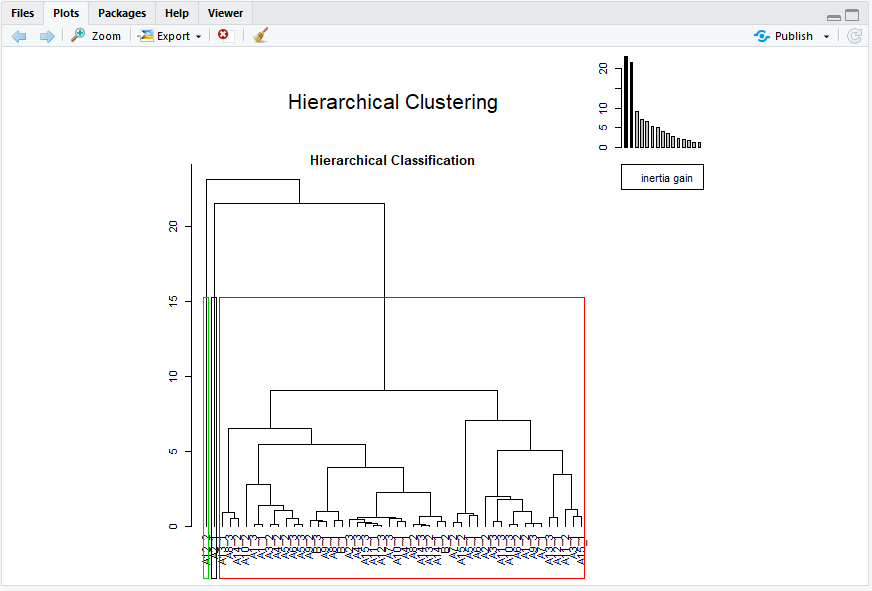


Figure 4.17: Dendogramme de la CAH

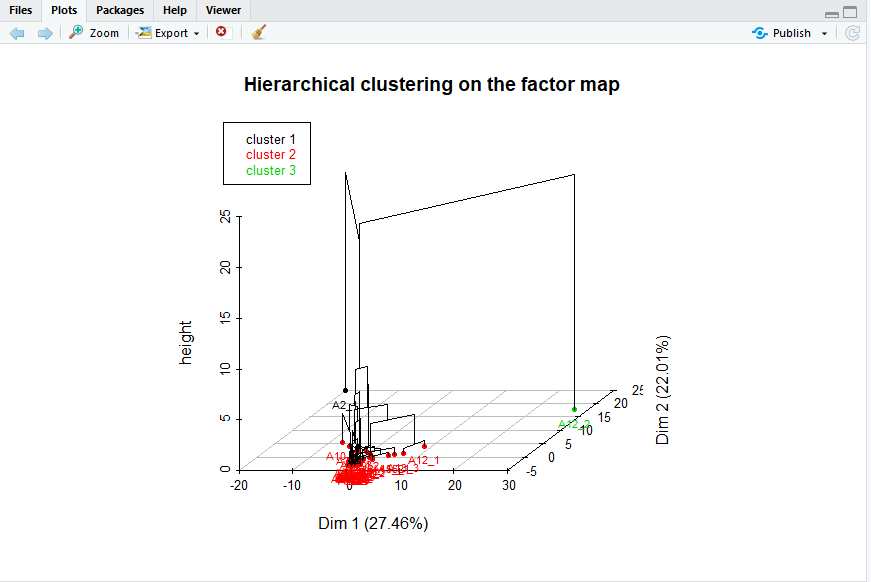


Figure 4.18: Carte des facteurs

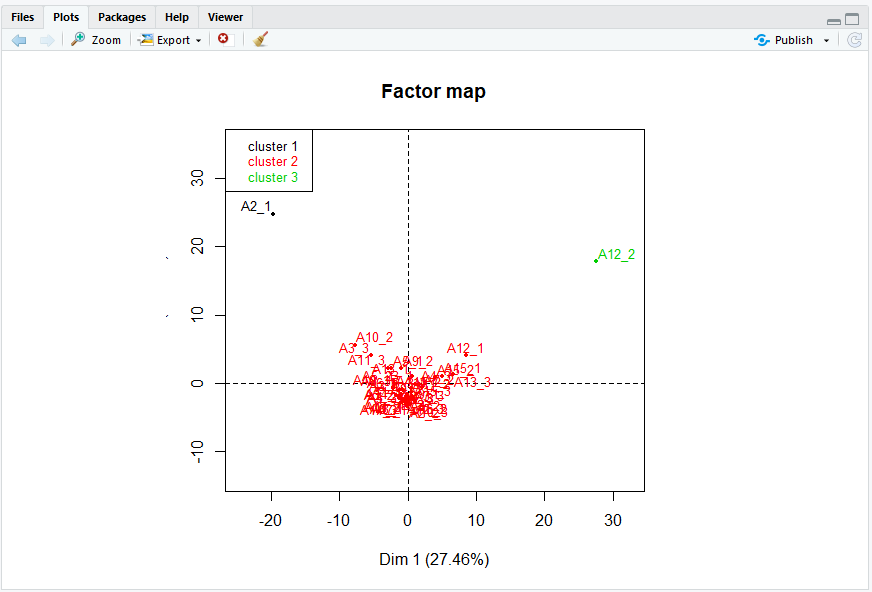


Figure 4.19: Distribution des groupes

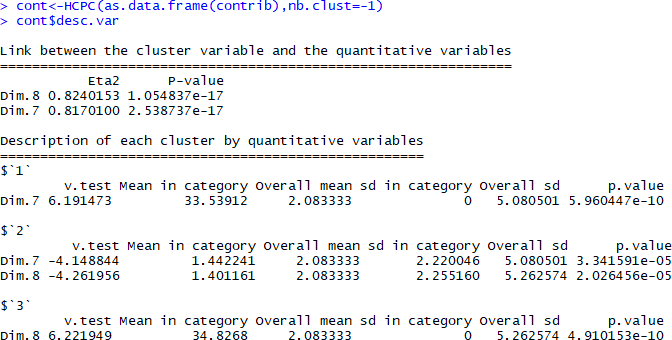


Figure 4.20: Description des modalités par classes

### Interprétation des résultats

Le dendogramme suggère 3 groupes.

Les modalités du premier cluster contribuent à l’axe 7.

Les modalités du deuxieme cluster ne contribuent pas aux axes 7 et 8. Les modalités du troisieme cluster contribuent à l’axe 8.

## Nuage des modalités projeté sur les 2 premiers axes

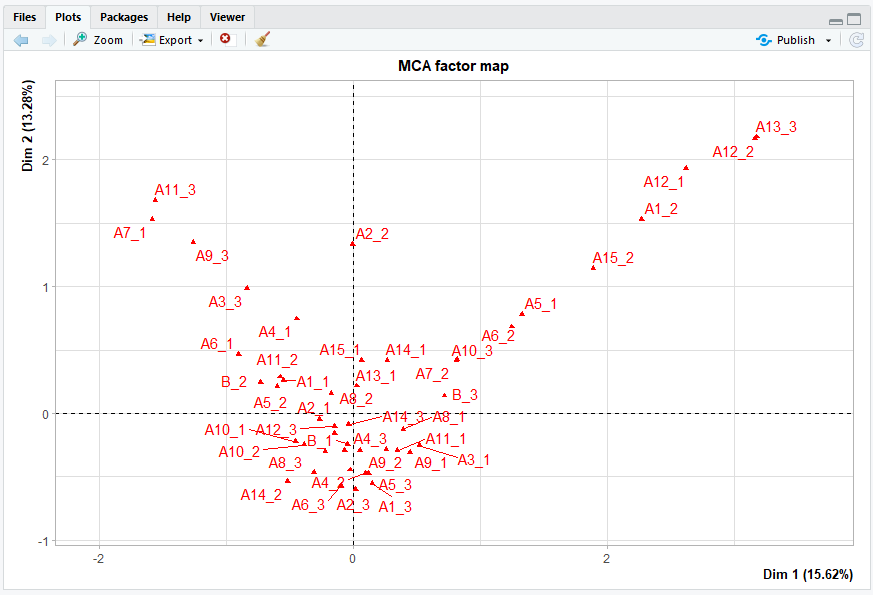


Figure 4.21: Nuage des modalités projeté sur les 2 premiers axes

# Nuage des individus

## cos2 des individus sur le sous espace

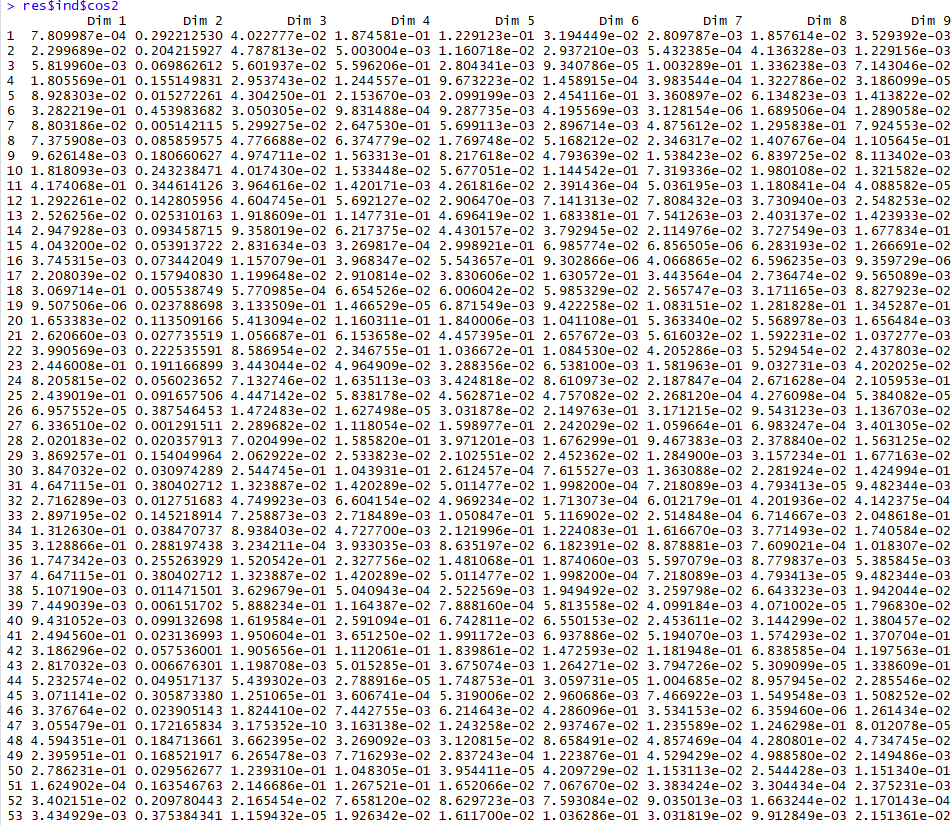


Figure 4.22: cos2 des individus sur le sous espace

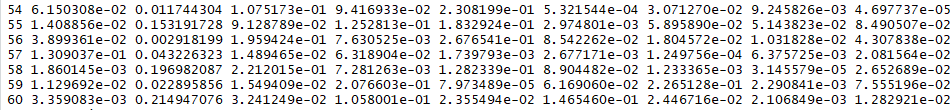


Figure 4.23: cos2 des individus sur le sous espace (suite)

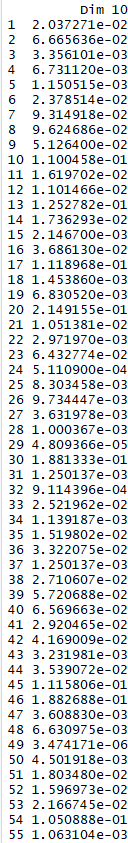


Figure 4.24: cos2 des individus sur le sous espace (suite)

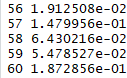


Figure 4.25: cos2 des individus sur le sous espace (suite)

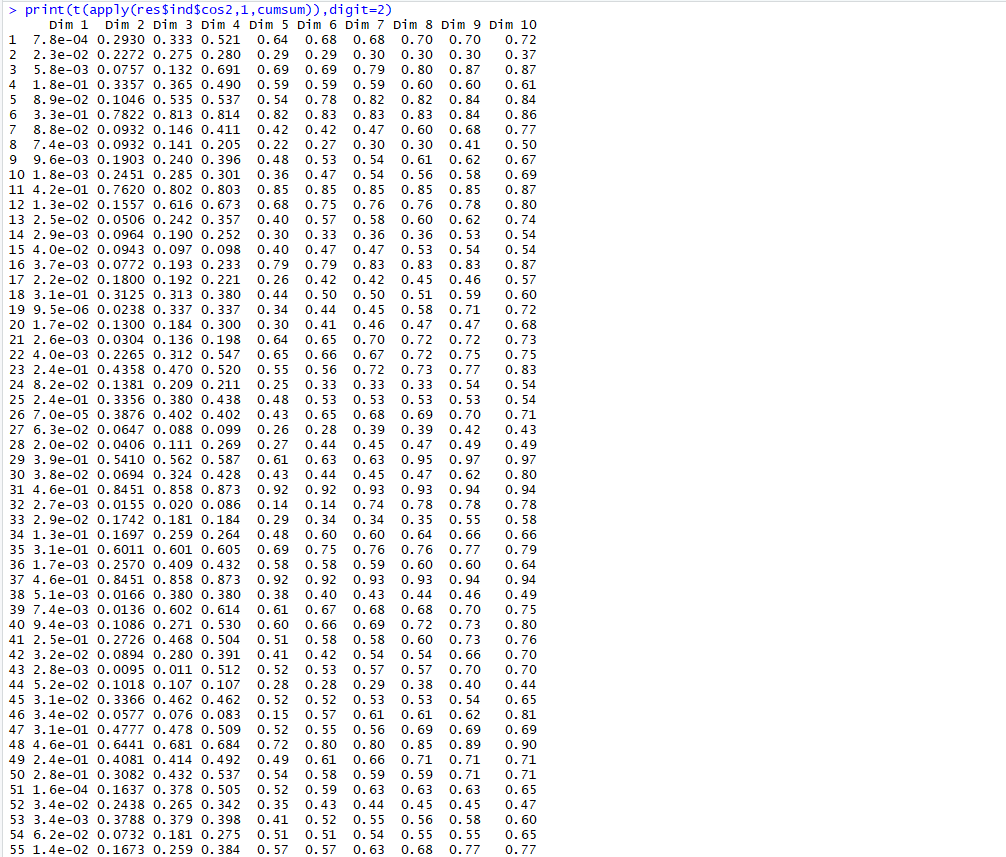


Figure 4.26: cos2 des individus sur le sous espace (suite)

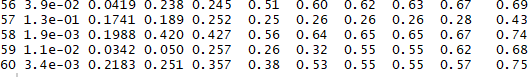


Figure 4.27: cos2 des individus sur le sous espace (suite)

## Individus bien représentés, moyennement représentés et faiblement représentés sur le sous espace

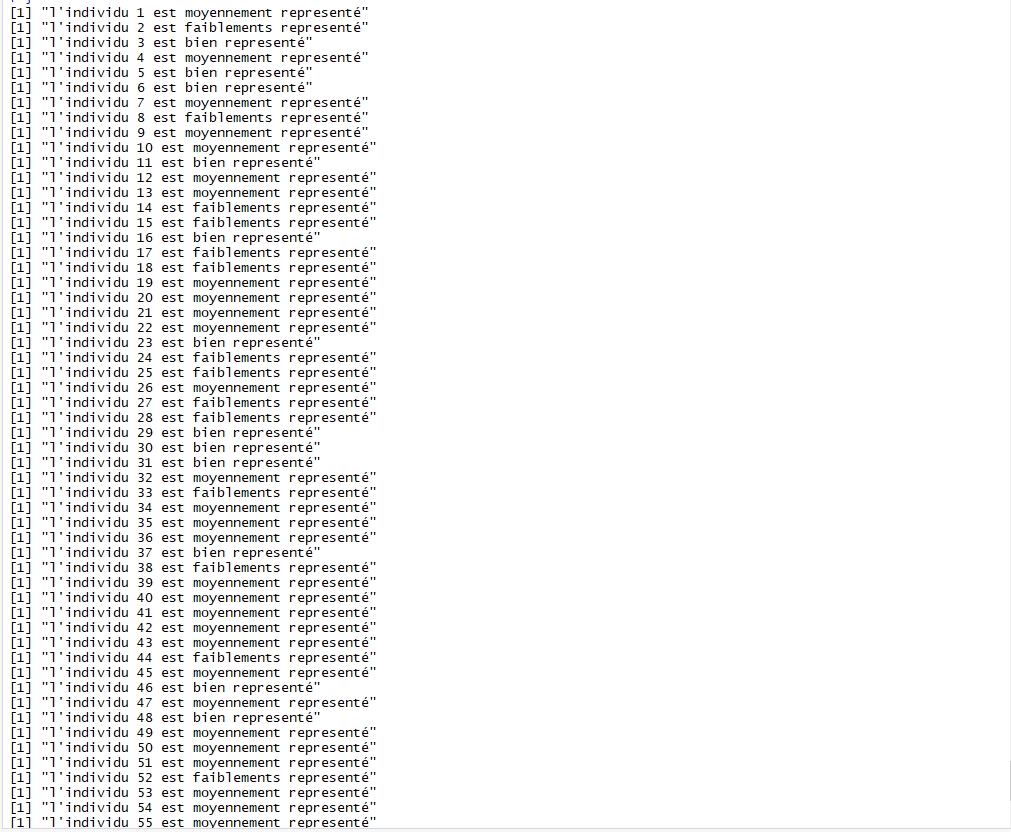


Figure 4.28: Individus bien représentés, moyennement représentés et faiblement repré- sentés sur le sous espace

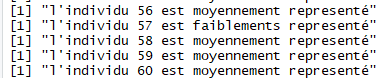


Figure 4.29: Individus bien représentés, moyennement représentés et faiblement repré- sentés sur le sous espace (suite)

## Contribution des individus dans chaque axe du sous es- pace

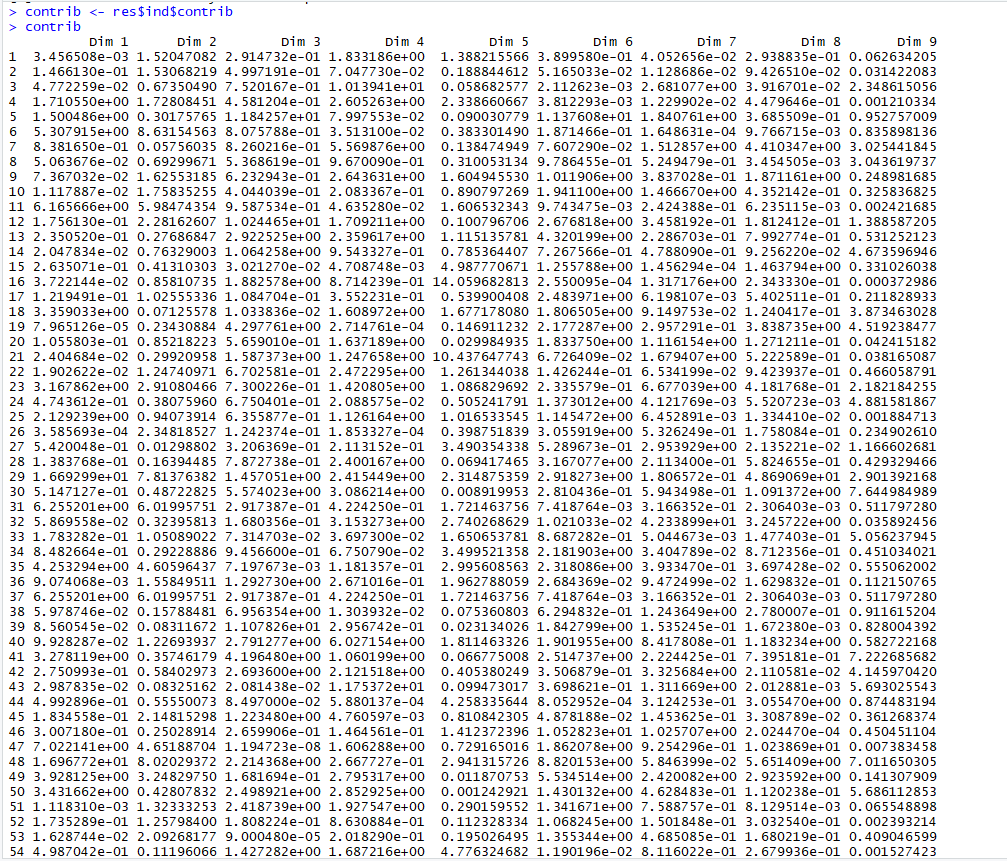


Figure 4.30: Contribution des individus dans chaque axe du sous espace

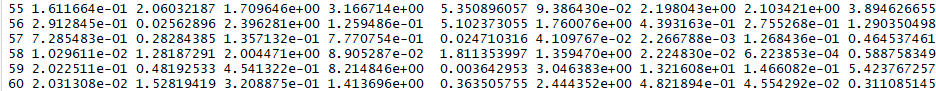


Figure 4.31: Contribution des individus dans chaque axe du sous espace (suite)

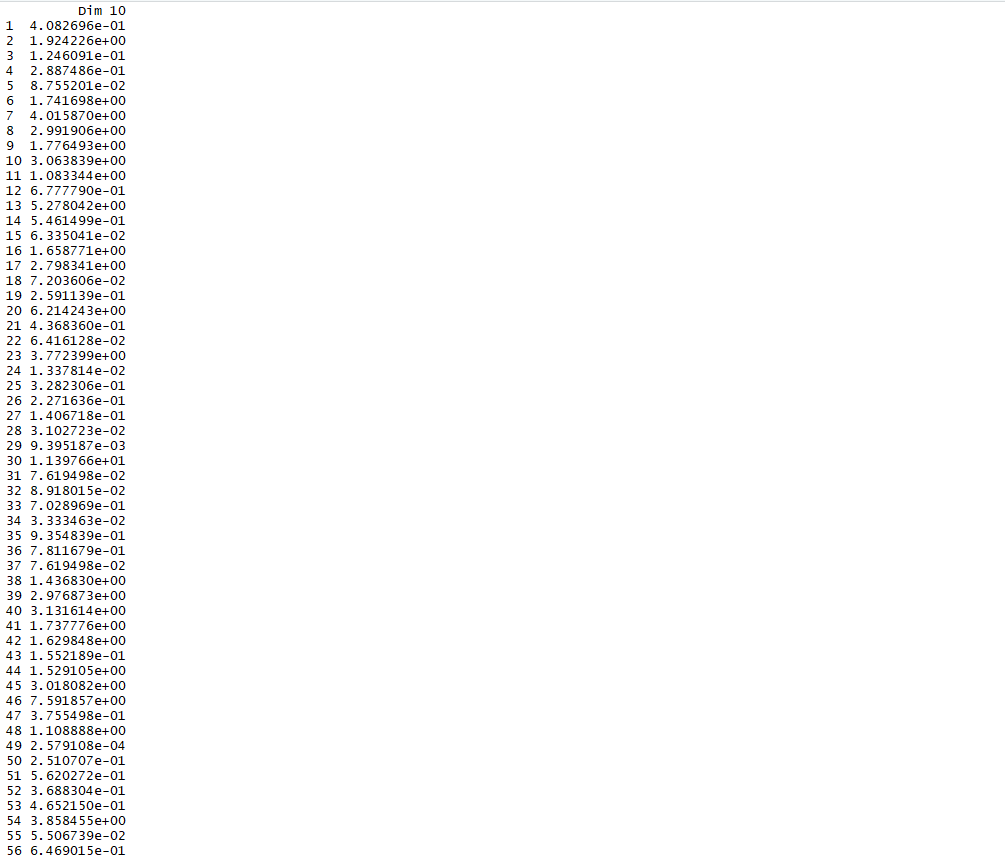


Figure 4.32: Contribution des individus dans chaque axe du sous espace (suite)

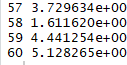


Figure 4.33: Contribution des individus dans chaque axe du sous espace (suite)

## Application de la CAH au tableau des contributions des individus aux axes du sous espace

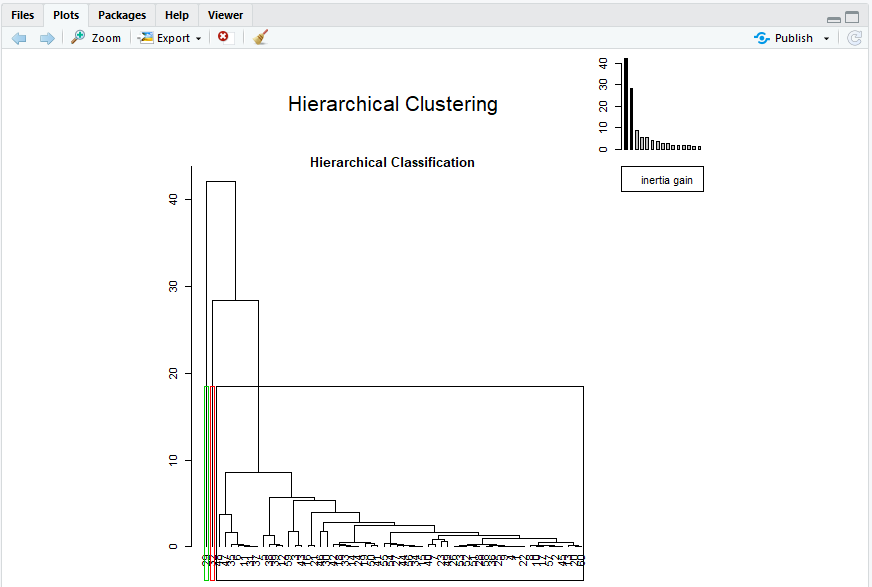


Figure 4.34: Dendogramme de la CAH

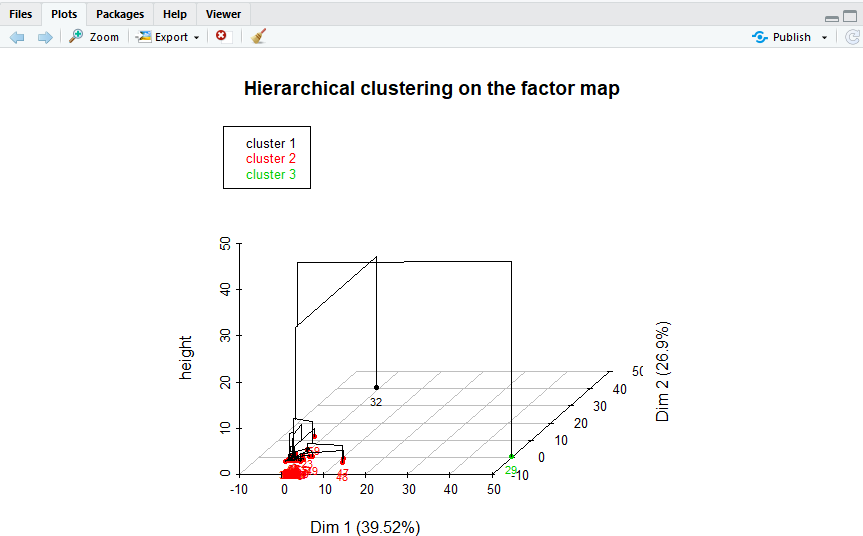


Figure 4.35: Carte des facteurs

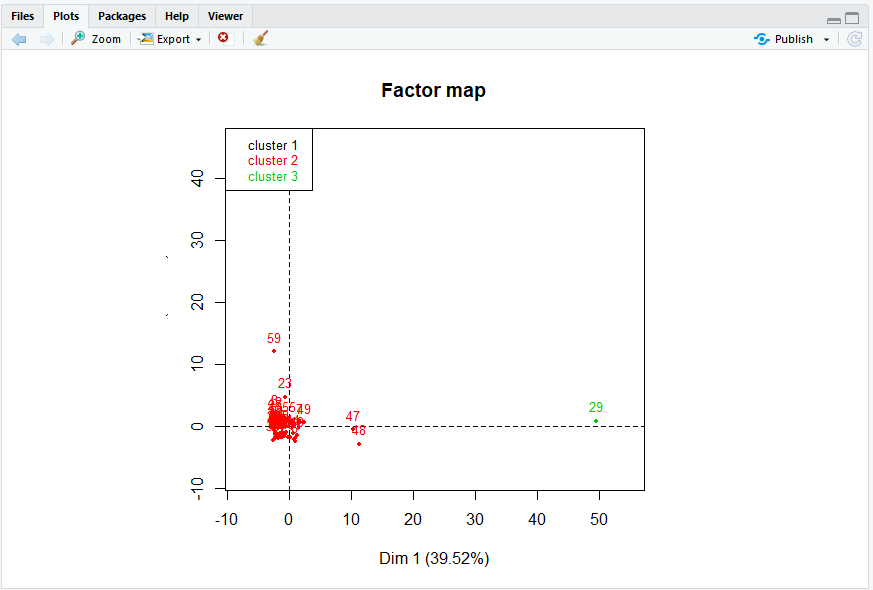


Figure 4.36: Distribution des groupes

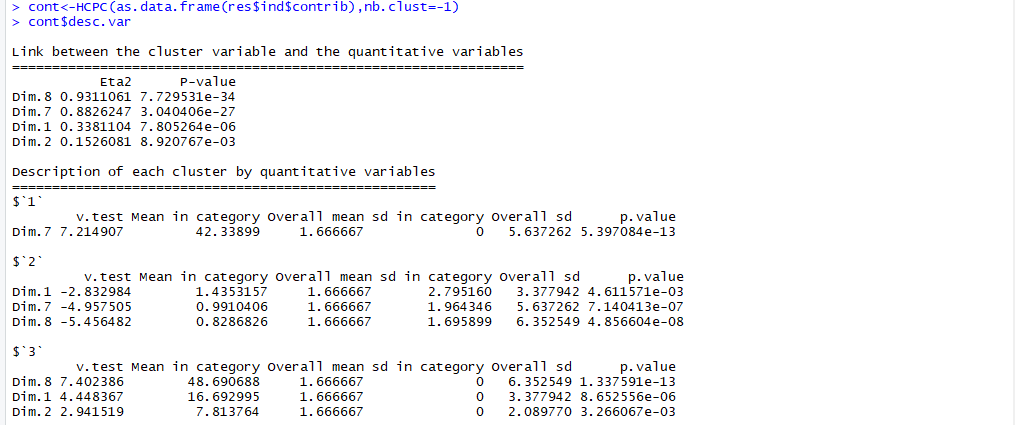


Figure 4.37: Description des modalités par classes

### Interprétation des résultats

CAH suggère un découpage en 3 groupes. Les individus sont regroupés en 3 classes selon leur contribution aux axes principaux.

Les individus du premier cluster contribuent à l’axe 7.

Les individus du deuxieme cluster ne contribuent pas aux axes 1, 8 et 7.

Les individus du troisieme cluster contribuent aux axes 1,2 et 8.

# Nuage des variables

## Coefficients de corrélation des variables avec les projec- tions sur les axes du sous espace

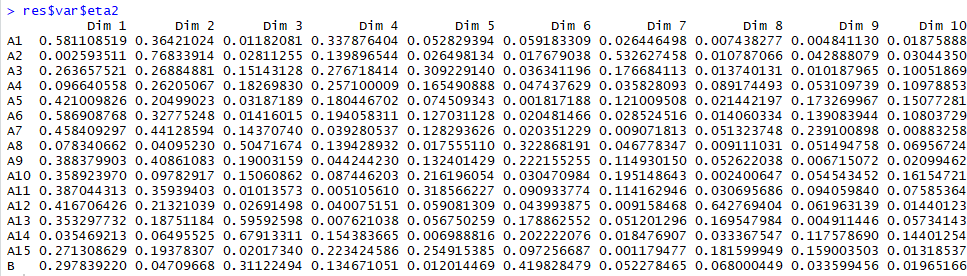


Figure 4.38: Coefficients de corrélation des variables avec les projections sur les axes du sous espace

## Graphique des coefficients de corrélation des variables avec les projections sur le 1er plan factoriel

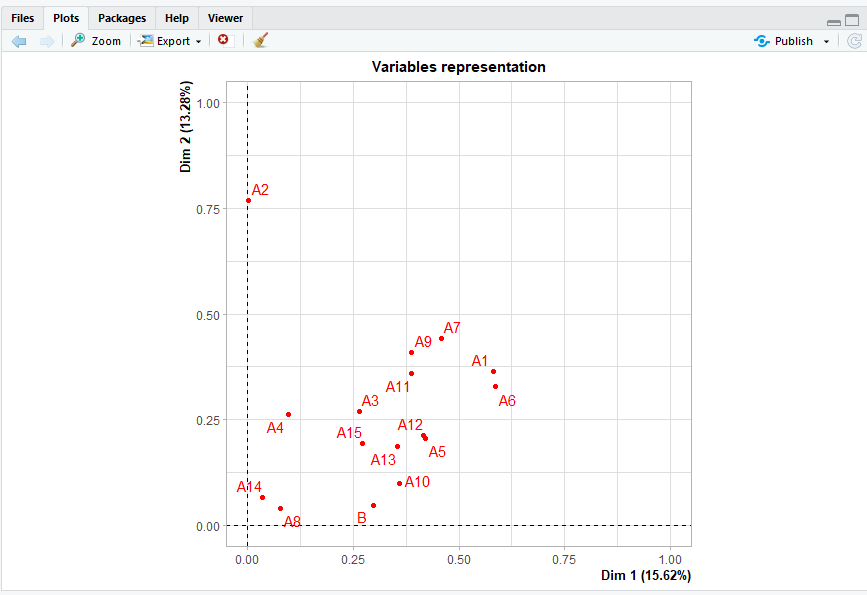


Figure 4.39: Graphique des coefficients de corrélation des variables avec les projections sur le 1er plan factoriel

# Conclusion

L’ACP nous a permis de réduire les 16 variables de notre jeu des données en 10 composantes principales. Ceci nous a donné l’opportunité de voir les différentes relations globales existantes entre les variables et les individus de notre jeu des données.

CONCLUSION GÉNÉRALE

Ce projet nous a permis d’appliquer les concepts théoriques vu dans le cours d’ana- lyse des données. Grâce à cette application nous avons pu non seulement visualiser les diffrents résultats des méthodes d’analyse des données (la méthode de régression linéaire multiple, les méthodes de classification, l’analyse des composantes principales et l’analyse des Correspondances Multiples), mais aussi améliorer notre esprit d’interprétation.

De plus, ce projet était une occasion pour se familiariser avec le language R et maîtriser ces différentes librairie.