

# Rapport de Projet du module de l’Analyse des données : Régression linéaire mutiple, Classiﬁcation, ACP et ACM

**Jamal Rebii Jihane Karib**

Filière 2IA

#### Ecole Nationale Supérieure d’Informatique et d’Analyse des Systèmes - RABAT

*Année Scolaire 2021/2022*.

# Table des matières

[Introduction](#_bookmark0) 1

[Descriptif du jeu de données](#_bookmark1) 3

* 1. [Variables du jeu de données](#_bookmark2) 3
     1. [Variable Dépendante](#_bookmark3) 3
     2. [Variables Explicatives](#_bookmark4) 3
  2. [Visualisation du jeu des données](#_bookmark5) 4
     1. [Aperçu au jeu des données](#_bookmark6) 4
     2. [Exploration du jeu de données](#_bookmark8) 5
     3. [Nuage de points](#_bookmark10) 6

1. [Régression Linéaire Multiple](#_bookmark12) 7
   1. [Modèle de régression linéaire multiple incluant toutes les variables expli-](#_bookmark13) [catives](#_bookmark13) 7
      1. [Existe-t-il des variables explicatives non significatives?](#_bookmark16) 8
      2. [Le coefficient de détermination](#_bookmark18) 9
      3. [Le test de Fisher](#_bookmark20) 9
   2. [Amélioration du modèle initial par la procédure step](#_bookmark22) 9
      1. [Les tests de validation pour le modèle amélioré de la procédure](#_bookmark28)

[de step](#_bookmark28) 12

* 1. [La méthode pas à pas](#_bookmark33) 14
     1. [Tests de validation du modèle obtenu par cette méthode](#_bookmark52) 31
     2. [Le critère AIC du modèle obtenu par la méthode de pas à pas](#_bookmark57) 33
  2. [Comparaison et Conclusion](#_bookmark59) 33

Iii

iv Table des matières

1. [Classification](#_bookmark60) 34
   1. [Préparation du jeu des données](#_bookmark61) 34
   2. [La méthode des centres mobiles K-means](#_bookmark64) 35
      1. [Taux d’inertie](#_bookmark66) 36
      2. [Détermination de nombre de classe Nc](#_bookmark68) 36
   3. [La Classification Ascendante Hiérarchique CAH](#_bookmark72) 38
      1. [Les variables quantitatives les plus corrélées avec la variable clas-](#_bookmark76) [sification](#_bookmark76) 39
      2. [Description des classes retenues par variables](#_bookmark78) 40
      3. [Calcul des taux d’inertie](#_bookmark82) 41
   4. [Comparaison](#_bookmark84) 41
2. [Analyse en Composante Principale](#_bookmark85) 42
   1. [ACP](#_bookmark88) 43
      1. [Application de ACP normé](#_bookmark89) 43
      2. [La justification du centrage et le réduction](#_bookmark91) 43
      3. [Calcul des indices KMO et MSAI](#_bookmark92) 43
      4. [Valeurs propres](#_bookmark94) 44
      5. [Graphiques des valeurs propres](#_bookmark96) 45
      6. [La dimension du sous espace](#_bookmark98) 45
   2. [Nuage des variables](#_bookmark100) 45
      1. [Calcul du cos2 des variables sur le sous espace](#_bookmark101) 45
      2. [Distinction des variables](#_bookmark103) 46
      3. [La contribution des variables dans chaque axe du sous espace](#_bookmark104) 47
      4. [Application de la CAH au tableau des contributions des variables](#_bookmark106)

[aux axes du sous espace](#_bookmark106) 47

* + 1. [le nuage des variables projeté sur les 2 premiers axes](#_bookmark111) 48
    2. [Indication les variables qui sont relativement bien corrélées (posi-](#_bookmark112) [tivement et négativement) avec les axes du 1er plan factoriel](#_bookmark112) 49
  1. [Nuage des individus](#_bookmark114) 50
     1. [Le cos2 des individus sur le sous espace](#_bookmark115) 50
     2. [Distinction des individus bien représentées, moyennement repré-](#_bookmark117) [sentées et faiblement représentées sur le sous espace](#_bookmark117) 51
     3. [La contribution des individus dans chaque axe du sous espace](#_bookmark119) 52
     4. [Application de la CAH au tableau des contributions des individus](#_bookmark121)

[aux axes du sous espace](#_bookmark121) 53

* + 1. [Le nuage des individus projeté sur les 2 premiers axes](#_bookmark126) 54

Table des matières v

* 1. [Conclusion](#_bookmark128) 54

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **[4](#_bookmark129)** | **[Analyse des Correspondances Multiples](#_bookmark129)** | | | **55** |
|  | [4.1 Préparation du jeu de données](#_bookmark130) . . . . . . . . . . . . . . . . . . . . . . . . . | | | 55 |
|  | | [4.1.1](#_bookmark131) | [Discrétisation des variables](#_bookmark131) . . . . . . . . . . . . . . . . . . . . . . . | 55 |
|  | | [4.1.2](#_bookmark136) | [Tableau Disjonctif Complet](#_bookmark136) . . . . . . . . . . . . . . . . . . . . . . . | 57 |
|  | | [4.1.3](#_bookmark138) | [La fréquence de chaque modalité](#_bookmark138) . . . . . . . . . . . . . . . . . . . . | 57 |
| [4.2](#_bookmark140) | | [ACM](#_bookmark140) | . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 58 |
|  | | [4.2.1](#_bookmark142) | [Valeurs propres](#_bookmark142) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 58 |
|  | | [4.2.2](#_bookmark144) | [Le graphique des valeurs propres](#_bookmark144) . . . . . . . . . . . . . . . . . . . | 58 |
|  | | [4.2.3](#_bookmark146) | [La dimension du sous espace](#_bookmark146) . . . . . . . . . . . . . . . . . . . . . . | 59 |
| [4.3](#_bookmark148) | | [Nuage](#_bookmark148) | [de Modalités](#_bookmark148) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 60 |
|  | | [4.3.1](#_bookmark149) | [Calcul du cos2 des modalités sur le sous espace](#_bookmark149) . . . . . . . . . . . | 60 |
|  | | [4.3.2](#_bookmark151) | [Distinction des modalités](#_bookmark151) . . . . . . . . . . . . . . . . . . . . . . . . | 61 |
|  | | [4.3.3](#_bookmark153) | [La contribution des modalités dans chaque axe du sous espace](#_bookmark153) . . | 61 |
|  | | [4.3.4](#_bookmark155) | [Application de la CAH au tableau des contributions des modalités](#_bookmark155) |  |
|  | |  | [aux axes du sous espace](#_bookmark155) . . . . . . . . . . . . . . . . . . . . . . . . . | 62 |
|  | | [4.3.5](#_bookmark160) | [Le nuage des modalités projeté sur les 2 premiers axes](#_bookmark160) . . . . . . . | 63 |
| [4.4](#_bookmark162) | | [Nuage](#_bookmark162) | [des individus](#_bookmark162) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 64 |
|  | | [4.4.1](#_bookmark163) | [Calcul du cos2 des individus sur le sous espace](#_bookmark163) . . . . . . . . . . . | 64 |
|  | | [4.4.2](#_bookmark165) | [Distinction des individus](#_bookmark165) . . . . . . . . . . . . . . . . . . . . . . . . | 65 |
|  | | [4.4.3](#_bookmark167) | [La contribution des individus dans chaque axe du sous espace](#_bookmark167) . . | 65 |
|  | | [4.4.4](#_bookmark168) | [Application de la CAH au tableau des contributions des individus](#_bookmark168) |  |
|  | |  | [aux axes du sous espace](#_bookmark168) . . . . . . . . . . . . . . . . . . . . . . . . . | 66 |
| [4.5](#_bookmark174) | | [Nuage](#_bookmark174) | [des variables](#_bookmark174) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 67 |
|  | | [4.5.1 Les coefficients de corrélation des variables avec les projections](#_bookmark175) | |  |
|  | | [sur les axes du sous espace](#_bookmark175) . . . . . . . . . . . . . . . . . . . . . . . | | 67 |
|  | | [4.5.2 Le graphique des coefficients de corrélation des variables avec les](#_bookmark177) | |  |
|  | | [projections sur le 1er plan factoriel](#_bookmark177) . . . . . . . . . . . . . . . . . . . | | 68 |
| [4.6](#_bookmark179) | | [Conclusion](#_bookmark179) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | 68 |

**[Conclusion générale](#_bookmark180) 69**

# Table des ﬁgures

* 1. [Jeu De données des températures moyennes.](#_bookmark7) 5
  2. [Dictionnaire du jeu de données.](#_bookmark9) 5
  3. [Matrice de corrélation du Jeu De données.](#_bookmark11) 6
  4. [Fonction lm pour la réalisation d’un modèle de régression linéaire multiple](#_bookmark14) [incluant toutes les variables](#_bookmark14) 7
  5. [Coefficient du modèle de régression linéaire multiple](#_bookmark15) 8
  6. [Sommaire du modèle de régression linéaire multiple](#_bookmark17) 8
  7. [Coefficient de détermination du modèle de régression linéaire multiple](#_bookmark19) 9
  8. [Test de Fisher](#_bookmark21) 9
  9. [Procédure Step](#_bookmark23) 10
  10. [Methode de Step (suite)](#_bookmark24) 11
  11. [Sommaire du modèle obtenue par la méthode de Step](#_bookmark25) 11
  12. [AIC du modèle initial](#_bookmark26) 12
  13. [AIC du modèle amélioré par la procédure Step](#_bookmark27) 12
  14. [Test d’homoscédasticité du modèle amélioré de la procédure de step](#_bookmark29) 12
  15. [Test de Shapiro-Wilk du modèle amélioré par la procédure Step](#_bookmark30) 13
  16. [Test de Ks du modèle amélioré de la procédure de step](#_bookmark31) 13
  17. [Les valeurs aberrantes du modèle amélioré de la procédure de step](#_bookmark32) 14
  18. [Introduction de la méthode pas à pas](#_bookmark34) 15
  19. [Sommaire de l’étape 0 de la méthode pas à pas](#_bookmark35) 15
  20. [l’étape 1 de la méthode pas à pas](#_bookmark36) 16
  21. [l’étape 2 de la méthode pas à pas](#_bookmark37) 16
  22. [l’étape 3 de la méthode pas à pas](#_bookmark38) 17
  23. [l’étape 4 de la méthode pas à pas](#_bookmark39) 18

vi

* 1. [l’étape 5 de la méthode pas à pas](#_bookmark40) 19
  2. [l’étape 6 de la méthode pas à pas](#_bookmark41) 20
  3. [l’étape 7 de la méthode pas à pas](#_bookmark42) 21
  4. [l’étape 8 de la méthode pas à pas](#_bookmark43) 22
  5. [l’étape 9 de la méthode pas à pas](#_bookmark44) 23
  6. [l’étape 10 de la méthode pas à pas](#_bookmark45) 24
  7. [l’étape 11 de la méthode pas à pas](#_bookmark46) 25
  8. [l’étape 12 de la méthode pas à pas](#_bookmark47) 26
  9. [l’étape 13 de la méthode pas à pas](#_bookmark48) 27
  10. [l’étape 14 de la méthode pas à pas](#_bookmark49) 28
  11. [l’étape 15 de la méthode pas à pas](#_bookmark50) 29
  12. [Sommaire du modèle obtenu par la méthode pas à pas](#_bookmark51) 29
  13. [Test d’homoscédasticité du modèle obtenu par la méthode pas à pas](#_bookmark53) 31
  14. [Test de Shapiro-Wilk du modèle obtenu par la méthode pas à pas](#_bookmark54) 31
  15. [Test de Ks du modèle obtenu par la méthode pas à pas](#_bookmark55) 32
  16. [Les valeurs aberrantes du modèle obtenu par la méhode pas à pas](#_bookmark56) 32
  17. [AIC du modèle obtenu par la méthode pas à pas](#_bookmark58) 33
  18. [Aperçu au jeu des données centrées réduites](#_bookmark62) 34
  19. [Dictionnaire jeu des données centrées réduites](#_bookmark63) 35
  20. [Clusters crées par le k-means pour k=4](#_bookmark65) 35
  21. [Recherche du N le plus adéquat](#_bookmark67) 36
  22. [Évolution de l’inertie en fonction du Nb. de groupes](#_bookmark69) 37
  23. [Recherche du Nc le plus adéquat](#_bookmark70) 37
  24. [Clusters crées par le k-means pour k=30](#_bookmark71) 38
  25. [Carte des facteurs](#_bookmark73) 39
  26. [Distribution des groupes](#_bookmark74) 39
  27. [Le dendrogramme de la CAH](#_bookmark75) 39
  28. [Relation entre les variables quantitatives et la variable classification](#_bookmark77) 40
  29. [Description de la classe 1 retenues par variables](#_bookmark79) 40
  30. [Description de la classe 2 retenues par variables](#_bookmark80) 40
  31. [Description de la classe 3 retenues par variables](#_bookmark81) 41
  32. [Calcul des taux d’inertie](#_bookmark83) 41
  33. [Sommaire détaillé du jeu de données](#_bookmark86) 42
  34. [corrélation entre les variables](#_bookmark87) 42
  35. [Graphes de l’ACP des variables](#_bookmark90) 43

viii *Table des figures*

* 1. [Les indices KMO et MSAI](#_bookmark93) 44
  2. [Valeurs propres](#_bookmark95) 44
  3. [Graphiques des valeurs propres](#_bookmark97) 45
  4. [Dimension du sous espace](#_bookmark99) 45
  5. [cos2 des variables sur le sous espace](#_bookmark102) 46
  6. [Contribution des variables dans chaque axe du sous espace](#_bookmark105) 47
  7. [Carte des facteurs](#_bookmark107) 48
  8. [Distribution des groupes](#_bookmark108) 48
  9. [Le dendrogramme de la CAH](#_bookmark109) 48
  10. [Description de la contribution aux axes par groupes](#_bookmark110) 48
  11. [le nuage des variables projeté sur les 2 premiers axes](#_bookmark113) 49
  12. [le cos2 des individus sur le sous espace](#_bookmark116) 50
  13. [Distinction des individus](#_bookmark118) 51
  14. [Contribution des individus](#_bookmark120) 52
  15. [Carte des facteurs](#_bookmark122) 53
  16. [Distribution des groupes](#_bookmark123) 53
  17. [Le dendrogramme de la CAH](#_bookmark124) 53
  18. [Description de la contribution aux axes par groupes](#_bookmark125) 53
  19. [le nuage des individus projeté sur les 2 premiers axes](#_bookmark127) 54
  20. [Aperçu aux variables quantitatives](#_bookmark132) 55
  21. [Taux de discrétisation](#_bookmark133) 56
  22. [Aperçu aux variables qualitatives](#_bookmark134) 56
  23. [Regroupement dans un fichier Excel](#_bookmark135) 56
  24. [Aperçu au tableau disjonctif complet](#_bookmark137) 57
  25. [La fréquence de chaque modalité](#_bookmark139) 57
  26. [Données générales](#_bookmark141) 58
  27. [Valeurs propres en ACM](#_bookmark143) 58
  28. [Le graphique des valeurs propres en ACM](#_bookmark145) 59
  29. [Dimension du sous espace en ACM](#_bookmark147) 59
  30. [cos2 des modalités sur le sous espace](#_bookmark150) 60
  31. [Distinction des modalités](#_bookmark152) 61
  32. [La contribution des modalités dans chaque axe du sous espace en ACM](#_bookmark154) 61
  33. [Carte des facteurs](#_bookmark156) 62
  34. [Distribution des groupes](#_bookmark157) 62
  35. [Le dendrogramme de la CAH](#_bookmark158) 62
  36. [Description des Modalités retenue par classes](#_bookmark159) 62
  37. [Le nuage des modalités projeté sur les 2 premiers axes](#_bookmark161) 63
  38. [Aperçu au cos2 des individus sur le sous espace en ACM](#_bookmark164) 64
  39. [Distinction des individus en ACM](#_bookmark166) 65
  40. [Aperçu à la contribution des individus dans chaque axe du sous espace](#_bookmark169) 66
  41. [Carte des facteurs](#_bookmark170) 66
  42. [Distribution des groupes](#_bookmark171) 66
  43. [Le dendrogramme de la CAH](#_bookmark172) 67
  44. [Description des axes retenue par classes](#_bookmark173) 67
  45. [Les coefficients de corrélation des variables avec les projections sur les axes](#_bookmark176)

[du sous espace](#_bookmark176) 67

* 1. [Le graphique des coefficients de corrélation des variables avec les projections](#_bookmark178)

[sur le 1er plan factoriel](#_bookmark178) 68

# Introduction

Ce rapport présente le travail que nous avons réalisé dans le cadre du projet du mo- dule d’analyse de données, ayant comme objectif l’étude de la température moyenne en considérant les différents facteurs qui peuvent l’affecter.

Notre démarche consisterait en l’intervention à toutes les phases du projet. En com- mençant par un descriptif du jeu de données choisi, puis en procédant à la régression linéaire multiple pour concevoir un modèle optimal de la prédiction. Ensuite, on va aborder la classification, et enfin on va s’étaler les différentes méthodes de l’analyse fac- torielles à savoir ACP et ACM.

# Descriptif du jeu de données

Notre travail a été basé sur un jeu de données dans le domaine météorologique ayant pour but l’analyse de la température moyenne.

Ce jeu de données est composé de 18 variables quantitatives et 197 observations. Il a pour but de corriger le biais des prévisions de températures de l’air moyennes du jour suivant du modèle LDAPS exploité par l’Administration météorologique coréenne au- dessus de Séoul, en Corée du Sud. Ces données sont constituées de données estivales de 2013 à 2017. Les données d’entrée sont en grande partie composées des données de prévision du jour suivant du modèle LDAPS.

## | Variables du jeu de données

### | Variable Dépendante

* + - * **Next\_Tmoy** : La température moyenne de l’air du lendemain .

### | Variables Explicatives

* + - * **Slope** : Pente(Â°) : 0.1 to 5.2.
      * **Solar radiation** : Rayonnement solaire entrant quotidien (wh/m2) : 4329,5 à 5992,9.
      * **Present\_Tmax** : Température maximale de l’air entre 0 et 21 h aujourd’hui (Â°C) : 20 à 37,6.
      * **Present\_Tmin** : Température minimale de l’air entre 0 et 21 h aujourd’hui (Â°C) : 11,3 à 29,9.
* **RHmin** : Prévision du modèle LDAPS de l’humidité relative minimale du jour suivant (%) : 19,8 à 98,5.
* **RHmax** : Prévision du modèle LDAPS de l’humidité relative maximale du jour suivant (%) : 58,9 à 100.
* **Tmax\_lapse** : Prévision du modèle LDAPS de la température maximale de l’air appliquée le jour suivant (°C) : 17,6 à 38,5.
* **Tmin\_lapse** : Prévision du modèle LDAPS de la température minimale de l’air appliquée le jour suivant (°C) : 14,3 à 29,6.
* **WS** : Prévision du modèle LDAPS de la vitesse moyenne du vent le jour suivant (m/s) : 2,9 à 21,9.
* **LH** : Prévision du modèle LDAPS du flux de chaleur latente moyen du jour sui- vant (W/m2) : -13,6 à 213,4.
* **CC1** : Prévision du modèle LDAPS de la couverture nuageuse moyenne fraction- née du premier jour sur 6 heures (0-5 h) (%) : 0 à 0,97.
* **CC2** : Prévision du modèle LDAPS de la couverture nuageuse moyenne répartie sur 6 heures le jour suivant (6-11 h) (%) : 0 à 0,97.
* **CC3** : Prévision du modèle LDAPS de la couverture nuageuse moyenne fraction- née sur 6 heures du 3e jour suivant (12-17 h) (%) : 0 à 0,98.
* **CC4** : Prévision du modèle LDAPS de la couverture nuageuse moyenne fraction- née sur 6 heures du 4e jour suivant (18-23 h) (%) : 0 à 0,97.

\* **lat** : Latitude (Â°) : 37.456 a 37.645.

\* **lon** : Longitude (Â°) : 126.826 a 127.135.

\* **DEM** : Dénivelé (m) : 12.4 a 212.3.

## | Visualisation du jeu des données

### | Aperçu au jeu des données

Table des ﬁgures 0.2. Visualisation du jeu des données

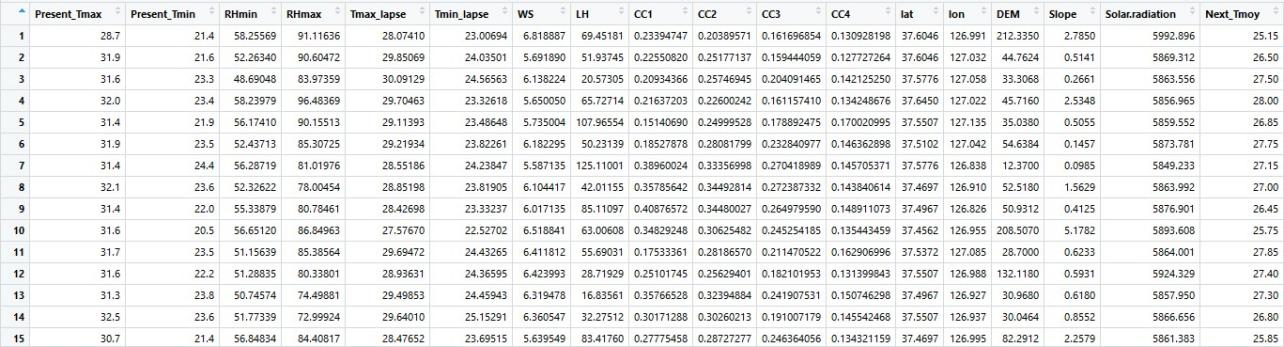


FIGURE 0.1 – Jeu De données des températures moyennes.

### | Exploration du jeu de données

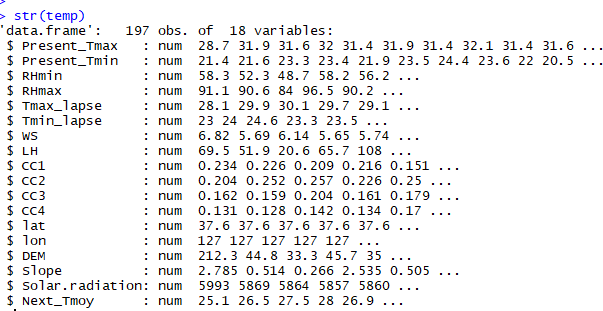


FIGURE 0.2 – Dictionnaire du jeu de données.

D’après la figure ci-dessus, on a toutes les variables sont numériques.

### | Nuage de points

Afin de visualiser et d’estimer des relations possibles entre chaque couple de va- riable, on va utilisé le graphe du nuage de points.

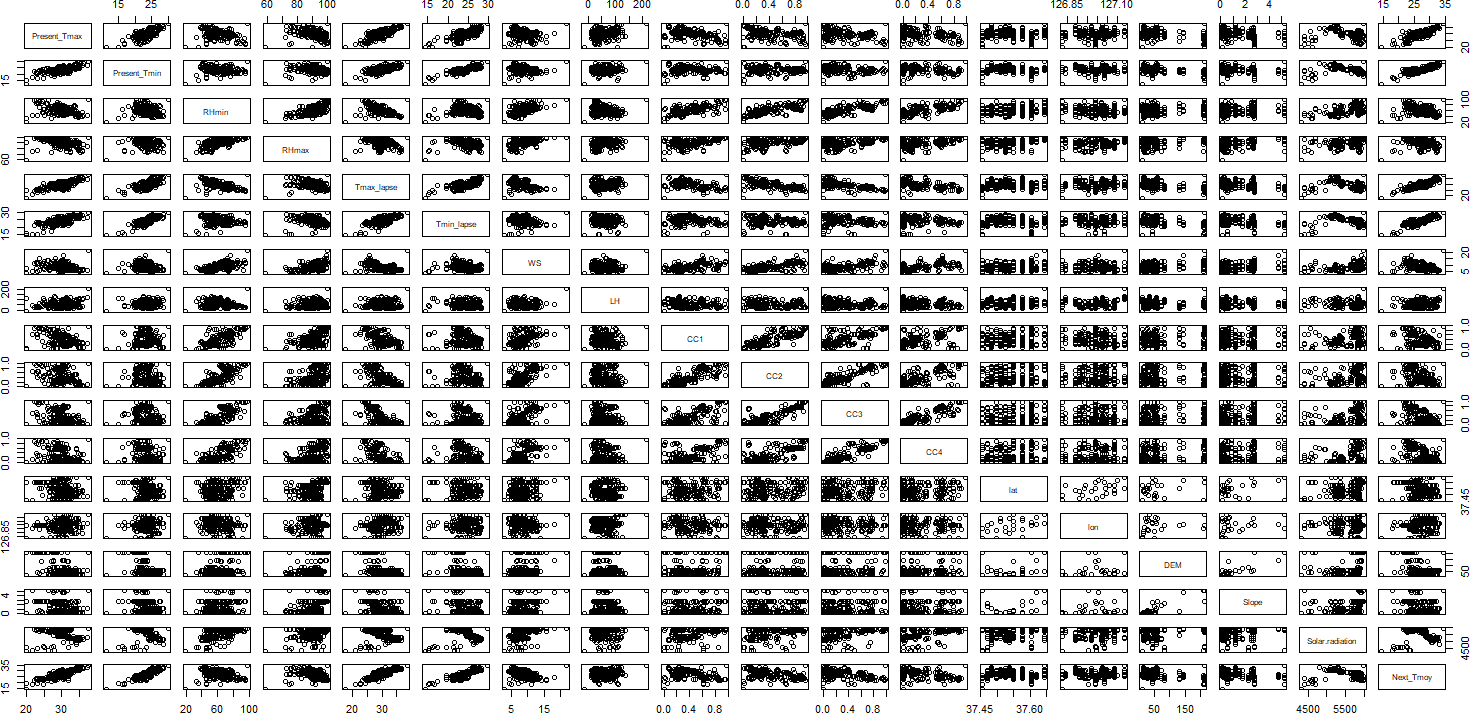


FIGURE 0.3 – Matrice de corrélation du Jeu De données.

D’après le nuage de points, on remarque l’existence de la corrélation entre certaines variables.

**1**

# Régression Linéaire Multiple

## | Modèle de régression linéaire multiple incluant toutes les variables explicatives

Le modèle de régression linéaire multiple est de la forme suivante :

*p*

*y* = *β*0 + ∑ *βjxj* + *ϵ*

*j*=1

où :

* y est la variable expliquée dans notre jeu de données c’est Next\_Tmoy.
* *x*1, *x*2, , *xp* sont les variables explicatives.
* *ϵ* est le terme d’erreur aléatoire du modèle.
* *β*0, *β*1, , *βp* Sont les paramètres à estimer.

Dans R, la fonction prédéfinie **lm()** nous permet d’appliquer la régression linéaire mul- tiple assez aisément.



FIGURE 1.1 – Fonction lm pour la réalisation d’un modèle de régression linéaire multiple incluant toutes les variables



FIGURE 1.2 – Coefficient du modèle de régression linéaire multiple

D’après la figure ci dessus,le modèle de régression linéaire multiple est de la forme suivante :

*Next*\_*Tmoy* = −2.18*e*02 + 9.90*e*−02 ∗ *Present*\_*Tmax* + + 4.58*e*−05 ∗ *Solar*.*radiation*

### | Existe-t-il des variables explicatives non signiﬁcatives ?

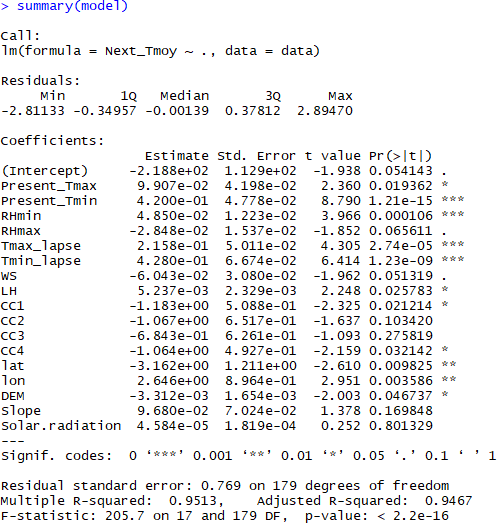


FIGURE 1.3 – Sommaire du modèle de régression linéaire multiple

D’après la figure au-dessus, on constate que les variables explicatives **Present\_Tmax**

, **RHmax**, **Slope**, **Solar.radiation**, **WS**, **CC2**, **CC3** ne sont pas significatives comme leur Pr(>|t|)≥ 0.05

### | Le coeﬃcient de détermination

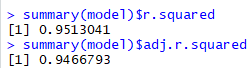


FIGURE 1.4 – Coefficient de détermination du modèle de régression linéaire multiple D’après la figure au-dessus on a le coefficient de détermination est le suivant :

*R*2 = 0.9513 et *R*2

*ajus*

= 0.9467.

Ces valeurs sont assez proches de 1 donc l’adéquation entre le modèle et les données observées est assez forte.

### | Le test de Fisher

Le test de Fisher permet de tester la significativité globale du modèle.

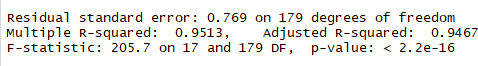


FIGURE 1.5 – Test de Fisher

Comme **p-value < 2.2e-16 < 0.05** alors le test de Fisher est significatif, ie il existe au moins une variable significativement non nulle.

## | Amélioration du modèle initial par la procédure step

La procédure du step permet d’améliorer notre modèle en eliminant les variables non significatives et en minimisant le AIC.

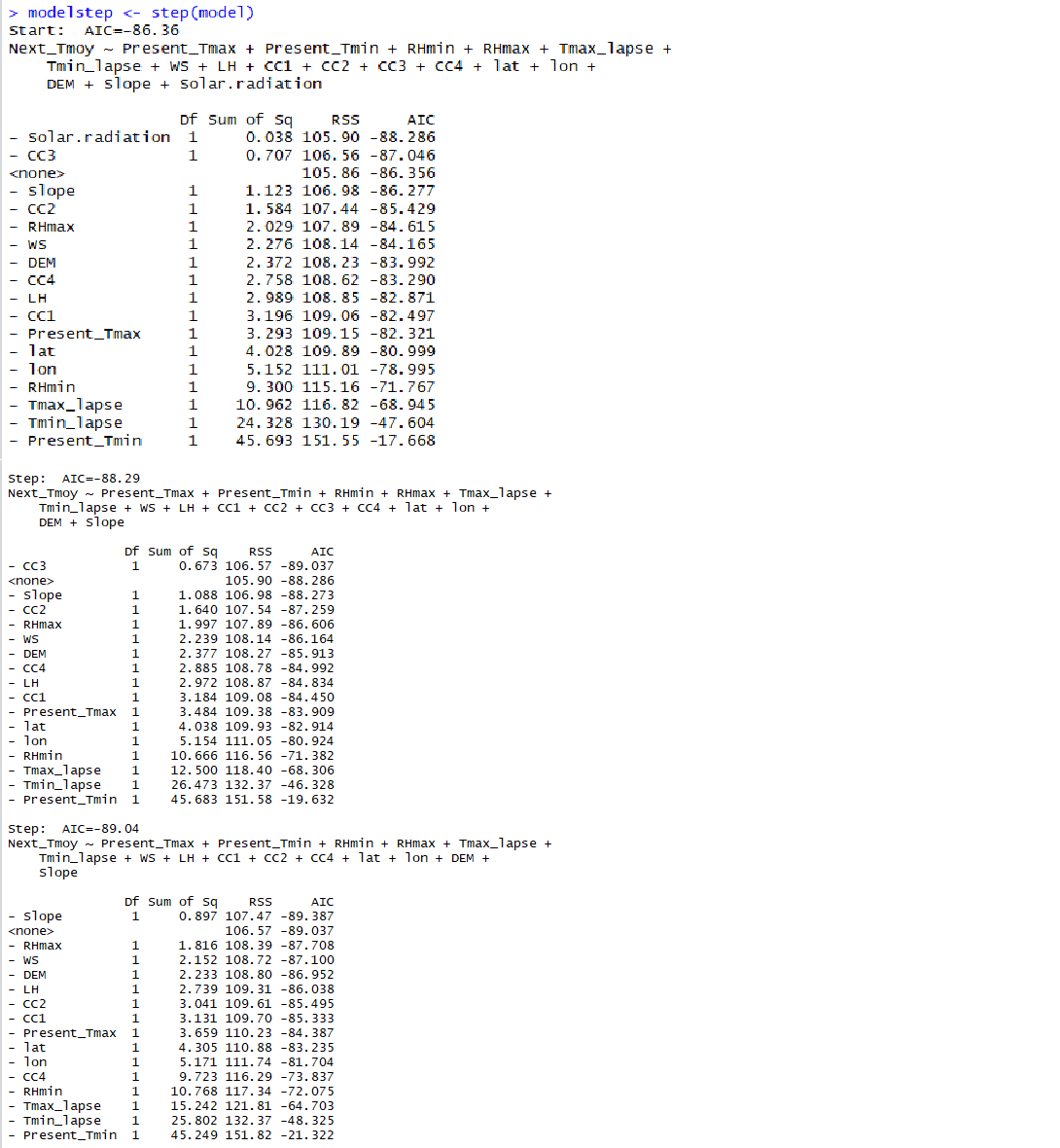


FIGURE 1.6 – Procédure Step

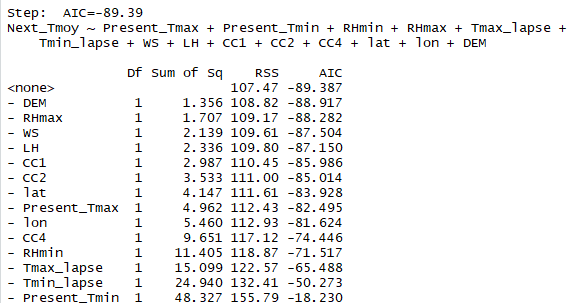


FIGURE 1.7 – Methode de Step (suite)

Affichons maintenant le modèle obtenue par la méthode de Step.

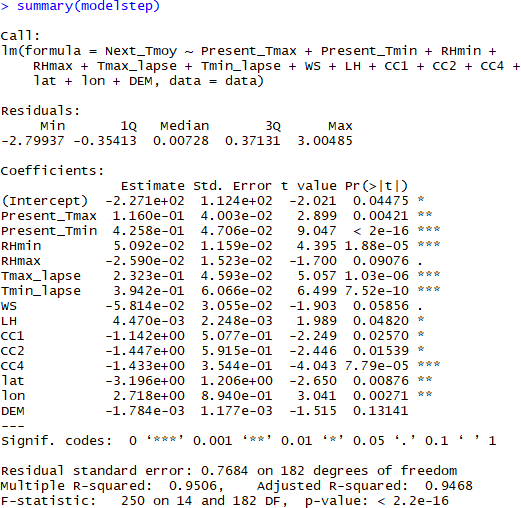


FIGURE 1.8 – Sommaire du modèle obtenue par la méthode de Step Camparons ensuites les AICs du modèle initiale et celui amélioré par la procédure

Step.



FIGURE 1.9 – AIC du modèle initial

FIGURE 1.10 – AIC du modèle amélioré par la procédure Step

D’après les figures au-dessus, On remarque que nous avons pu diminuer le AIC

de 474.7056 jusqu’à 471.6749, et augmenter *R*2

*adj*

de 0.9467 à 0.9468 gràce à un modèle

amélioré de variables : **Present\_Tmax**, **Present\_Tmin**, **RHmin**, **RHmax**, **Tmax\_lapse**, **Tmin\_lapse**, **WS**, **LH**, **CC1**, **CC2**, **CC4**, **lat**, **lon**, **DEM** .

### | Les tests de validation pour le modèle amélioré de la procé- dure de step

#### | Test d’homoscédasticité

C’est une représentation graphique où on trace le résidu par rapport à les valeurs prédites.



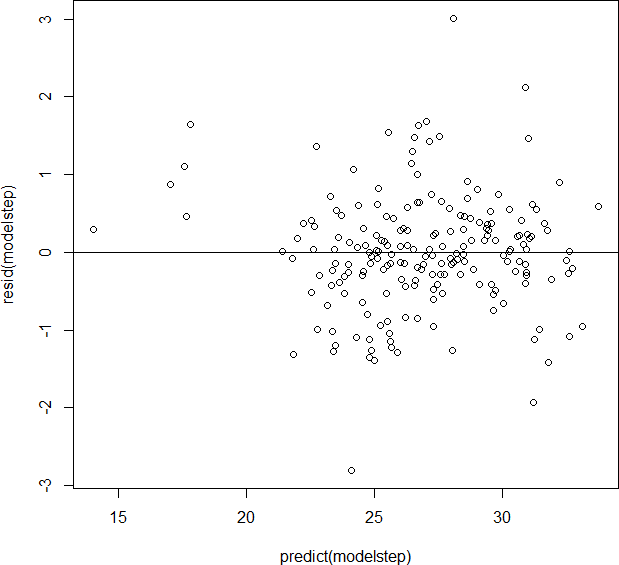


FIGURE 1.11 – Test d’homoscédasticité du modèle amélioré de la procédure de step

D’après la figure au-dessus, la distribution des points est assez homogène.

#### | Tests de normalité

Ils permettent de savoir si une série de données suit une loi normale.

##### Test de shapiro-Wilk :

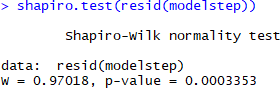


FIGURE 1.12 – Test de Shapiro-Wilk du modèle amélioré par la procédure Step

On a **p-value = 0.0003353** ≤ **0.05** ie la p-value est significative. L’échantillon ne suit donc pas une loi normale.

##### Test de Kolmogorov-Smirnov :

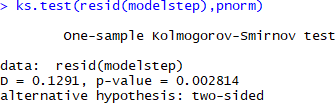
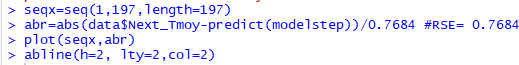


FIGURE 1.13 – Test de Ks du modèle amélioré de la procédure de step

On a **p-value = 0.002814** ≤ **0.05** ie la p-value est significative. L’échantillon ne suit donc pas une loi normale.

#### | Recherche des valeurs aberrantes



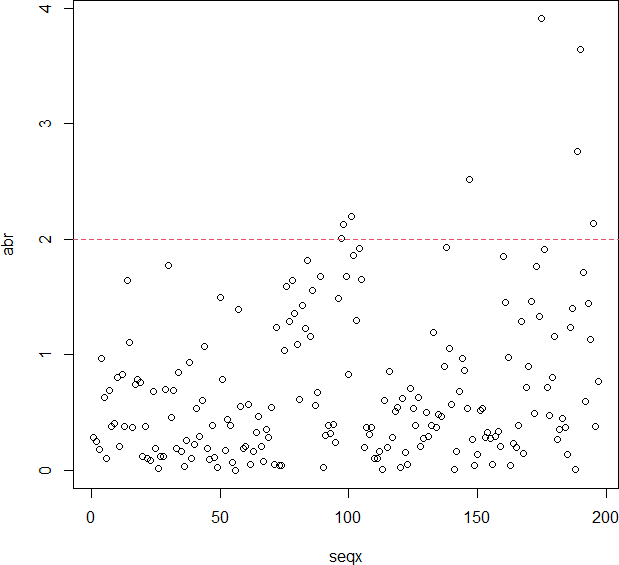


FIGURE 1.14 – Les valeurs aberrantes du modèle amélioré de la procédure de step

On remarque qu’il y a 7 valeurs aberrantes.

## | La méthode pas à pas

Il s’agit d’une méthode de sélection des variables explicatives basée sur le test de Fisher.

D’abord, On commence par intégrer la variable la plus significative ayant F le plus grand.

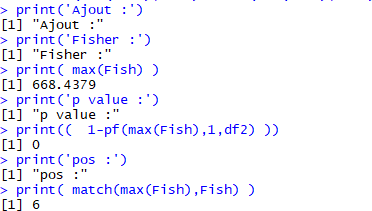


FIGURE 1.15 – Introduction de la méthode pas à pas

Donc d’après la figure au dessus, cette variable est celle ayant comme position 6 c’est à dire **Tmin\_Lapse**.

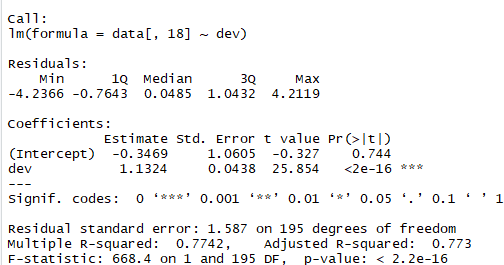


FIGURE 1.16 – Sommaire de l’étape 0 de la méthode pas à pas

**\*Etape 1 :** D’aprés la figure au dessous, On a :

* La variable entrante est celle ayant comme position 5 ie **Tmax\_lapse** car *p* −

*value <* 10%.

* On a pas de variable sortante car même si **Tmax\_lapse** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait au cours de cette étape.

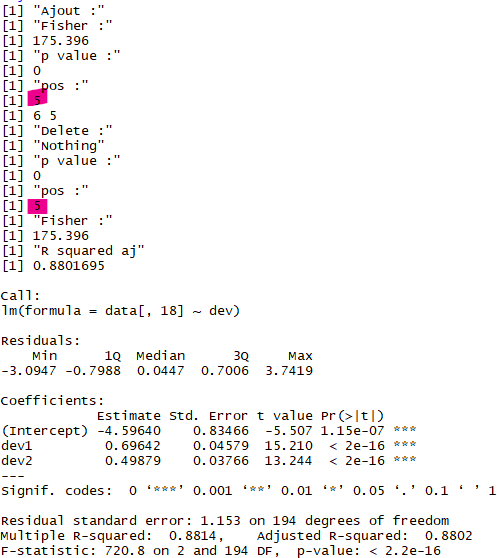


FIGURE 1.17 – l’étape 1 de la méthode pas à pas

##### \*Etape 2 :

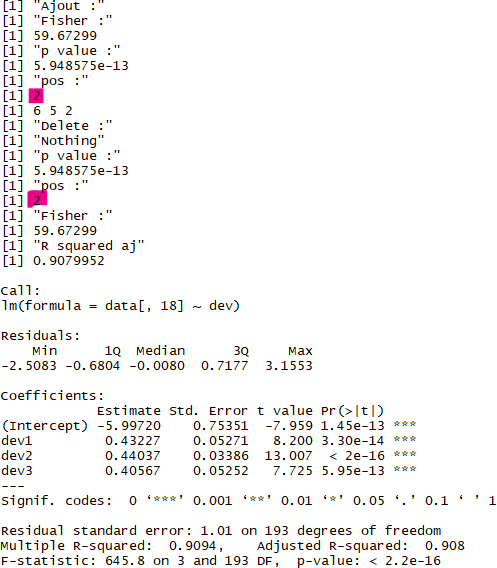


FIGURE 1.18 – l’étape 2 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* La variable entrante est celle ayant comme position 2 ie **Present\_Tmin** car *p* −

*value <* 10% .

* + On a pas de variable sortante car même si **Present\_Tmin** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 3 :

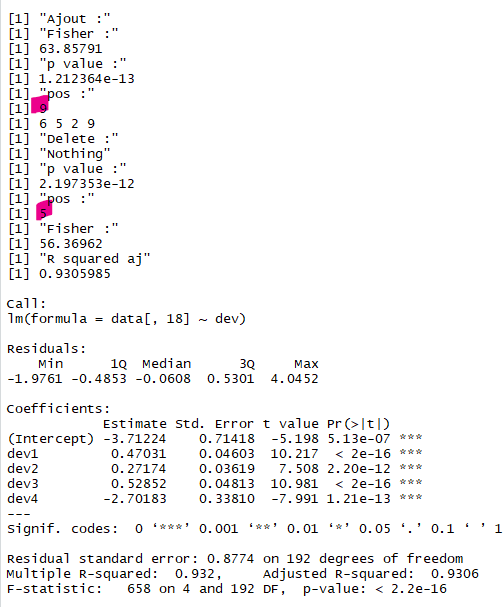


FIGURE 1.19 – l’étape 3 de la méthode pas à pas D’aprés la figure au dessus, On a :

* + La variable entrante est celle ayant comme position 9 ie **CC1** car *p* − *value <* 10%
  + On a pas de variable sortante car même si **Present\_Tmin** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 4 :

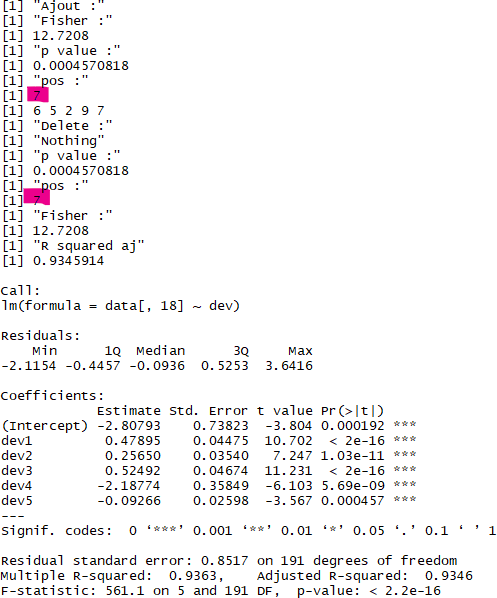


FIGURE 1.20 – l’étape 4 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* La variable entrante est celle ayant comme position 7 ie **WS** car *p* − *value <* 10%
* On a pas de variable sortante car même si **WS** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 5 :

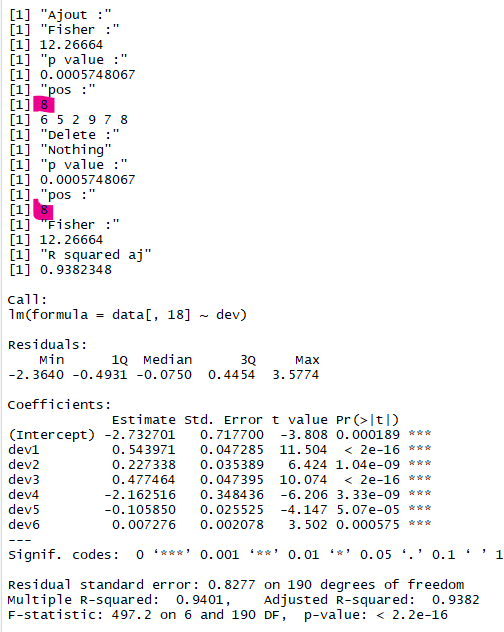


FIGURE 1.21 – l’étape 5 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* + La variable entrante est celle ayant comme position 8 ie **LH** car *p* − *value <* 10%
  + On a pas de variable sortante car même si **LH** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 6 :

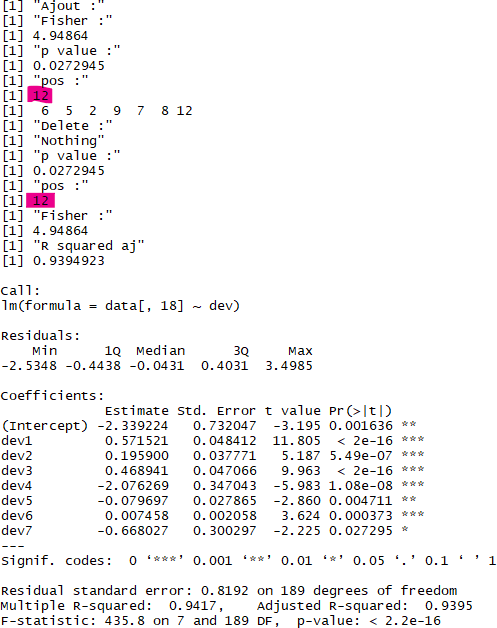


FIGURE 1.22 – l’étape 6 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* La variable entrante est celle ayant comme position 12 ie **CC4** car *p* − *value <*

10%

* On a pas de variable sortante car même si **CC4** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 7 :

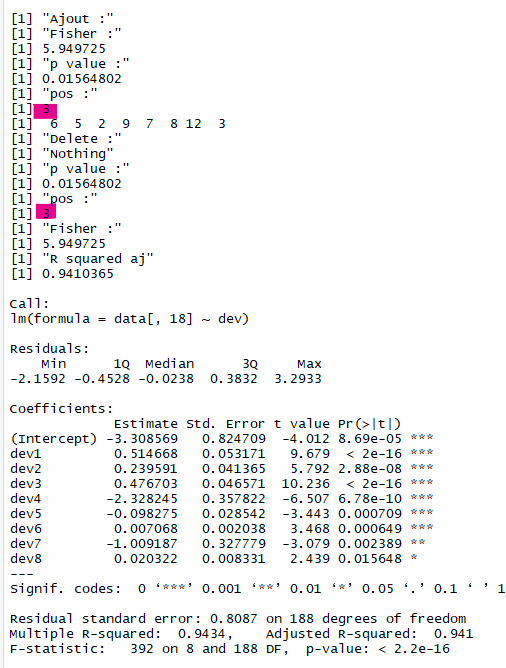


FIGURE 1.23 – l’étape 7 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* + La variable entrante est celle ayant comme position 3 ie **RHmin** car *p* − *value <*

10%

* + On a pas de variable sortante car même si **RHmin** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 8 :

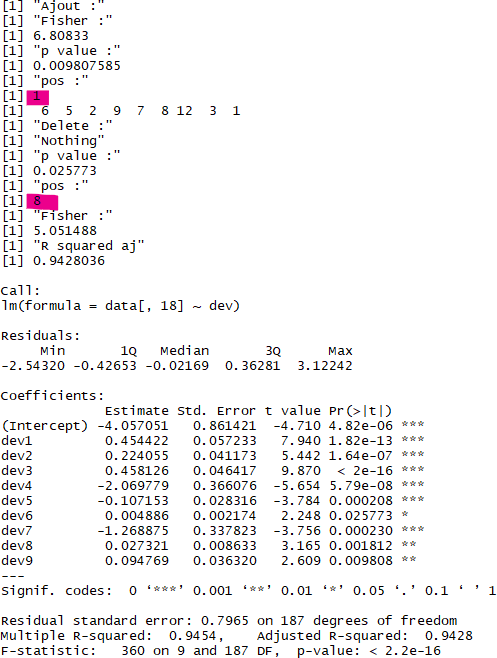


FIGURE 1.24 – l’étape 8 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* La variable entrante est celle ayant comme position 1 ie **Present\_Tmax** car *p* − *value <* 10%
* On a pas de variable sortante car même si **LH** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 9 :

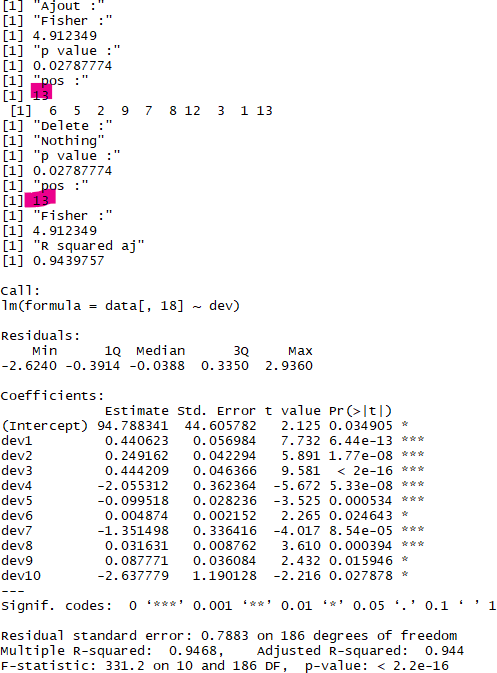


FIGURE 1.25 – l’étape 9 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* + La variable entrante est celle ayant comme position 13 ie **lat** car *p* − *value <* 10%
  + On a pas de variable sortante car même si **lat** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 10 :

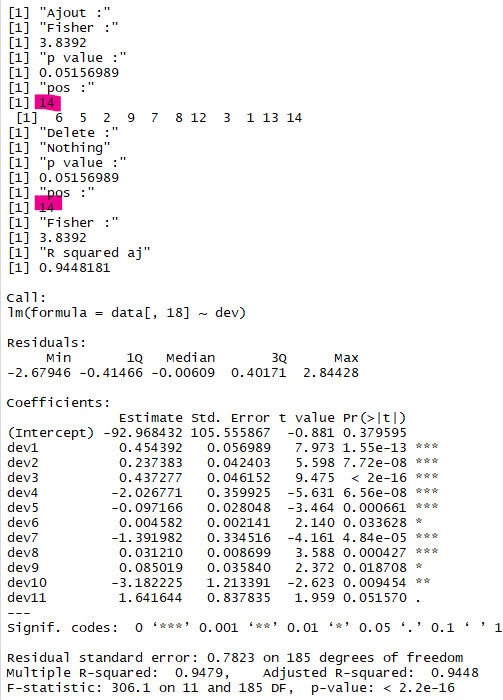


FIGURE 1.26 – l’étape 10 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* La variable entrante est celle ayant comme position 14 ie **lon** car *p* − *value <* 10%
* On a pas de variable sortante car même si **lon** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 11 :

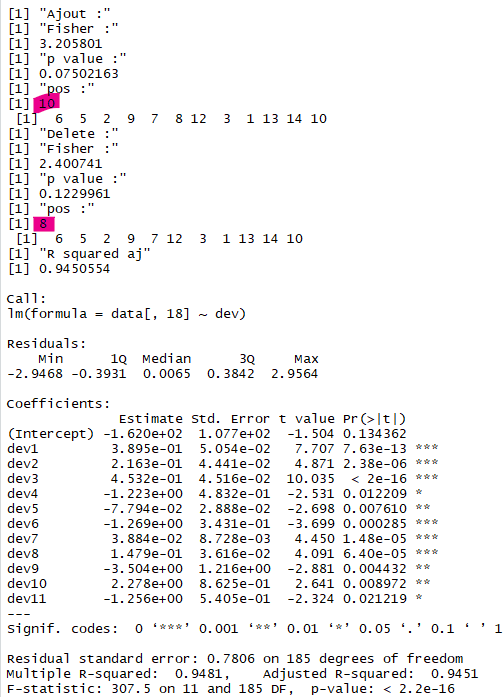


FIGURE 1.27 – l’étape 11 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* + La variable entrante est celle ayant comme position 10 ie **CC2** car *p* − *value <*

10%

* + La variable sortante est celle ayant comme position 8 ie **LH** car sa *p* − *value >*

10%

##### \*Etape 12 :

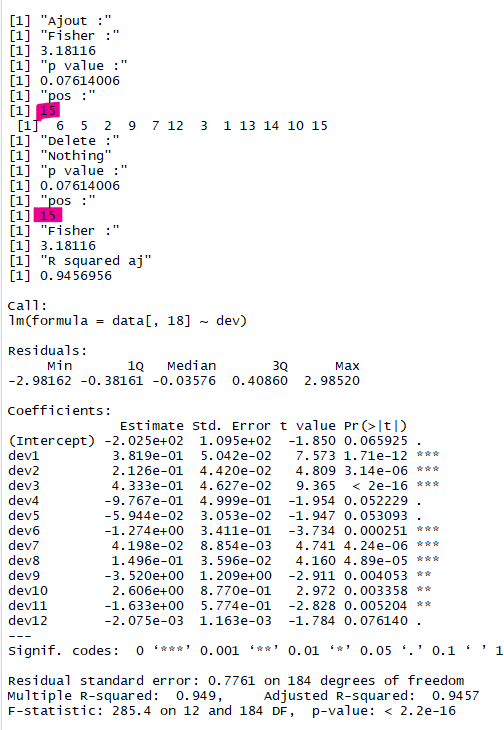


FIGURE 1.28 – l’étape 12 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* La variable entrante est celle ayant comme position 15 ie **DEM** car *p* − *value <*

10%

* On a pas de variable sortante car même si **DEM** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 13 :

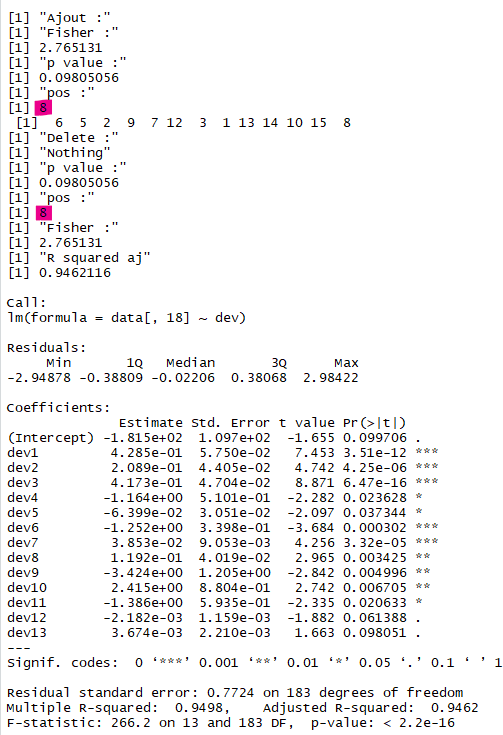


FIGURE 1.29 – l’étape 13 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* + La variable entrante est celle ayant comme position 8 ie **LH** car *p* − *value <* 10%
  + On a pas de variable sortante car même si **LH** la variable ayant le plus petit F, sa *p* − *value <* 10%, donc pour cette étape on n’a pas de retrait ai cours de cette étape.

##### \*Etape 14 :

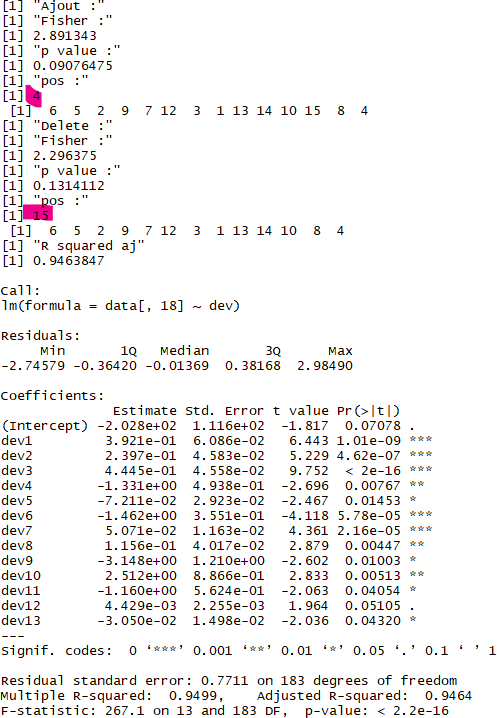


FIGURE 1.30 – l’étape 14 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* La variable entrante est celle ayant comme position 4 ie **RHmax** car *p* − *value <*

10%

* La variable sortante est celle ayant comme position 15 ie **DEM** car *p* − *value >*

10%

##### \*Etape 15 :

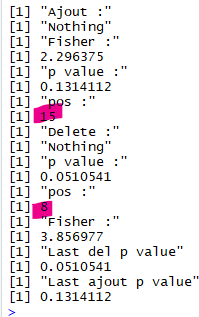


FIGURE 1.31 – l’étape 15 de la méthode pas à pas

D’aprés la figure au dessus, On a :

* + On a pas de variable entrante car même si **DEM** la variable ayant le grand petit F sa *p* − *value >* 10%, donc pas d’ajout.
  + On a pas de variable sortante car même si **LH** la variable ayant le plus petit F, sa

*p* − *value <* 10%, donc pas de retrait.

Donc la procédure s’arrête, et on a alors le modèle ayant comme variables explica- tives **Tmin\_lapse**, **Tmax\_lapse**, **Present\_Tmin**, **CC1**, **WS**, **LH**, **CC4**, **RHmin**, **Present\_Tmax**, **lat**, **lon**, **CC2**, **RHmax**

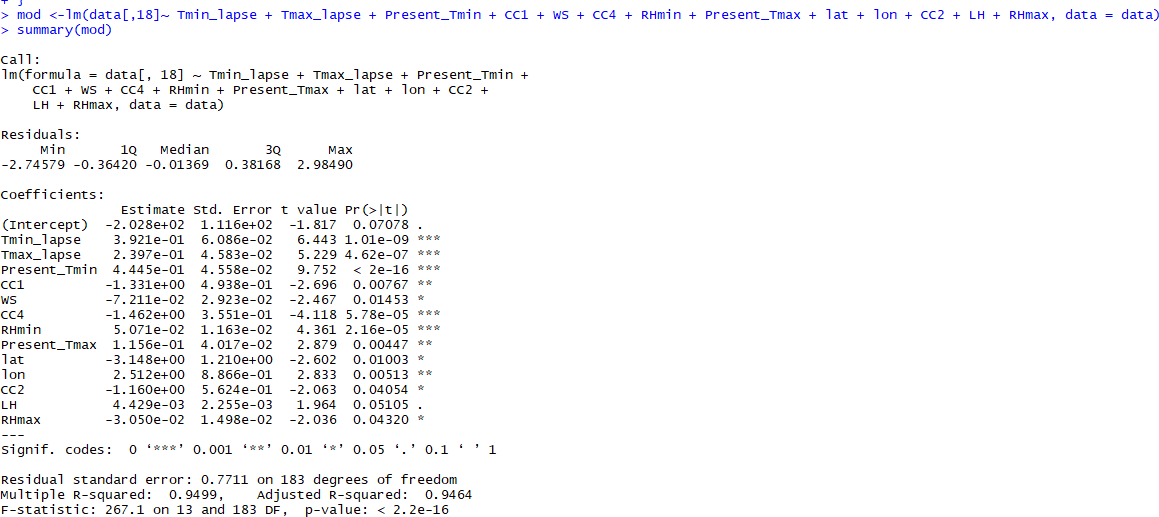


FIGURE 1.32 – Sommaire du modèle obtenu par la méthode pas à pas

Et on résume cette procédure par le tableau suivant :

| **Étape** | **Variable entree** | **Variable sortie** | **Radj** | **F** | **p-value** |
| --- | --- | --- | --- | --- | --- |
| 0 | Tmin\_lapse | - | 0.773 | 668.44 | < 2.2e-16 |
| 1 | Tmax\_lapse | - | 0.8802 | 720.8 | < 2.2e-16 |
| 2 | Present\_Tmin | - | 0.908 | 645.8 | < 2.2e-16 |
| 3 | CC1 | - | 0.9306 | 658 | < 2.2e-16 |
| 4 | WS | - | 0.9346 | 561.1 | < 2.2e-16 |
| 5 | LH | - | 0.9382 | 497.2 | < 2.2e-16 |
| 6 | CC4 | - | 0.9395 | 435.8 | < 2.2e-16 |
| 7 | RHmin | - | 0.941 | 392 | < 2.2e-16 |
| 8 | Present\_Tmax | - | 0.9428 | 360 | < 2.2e-16 |
| 9 | lat | - | 0.944 | 331.2 | < 2.2e-16 |
| 10 | lon | - | 0.9448 | 306.1 | < 2.2e-16 |
| 11 | CC2 | LH | 0.9451 | 307.5 | < 2.2e-16 |
| 12 | DEM | - | 0.9457 | 285.4 | < 2.2e-16 |
| 13 | LH | - | 0.9462 | 266.2 | < 2.2e-16 |
| 14 | RHmax | DEM | 0.9464 | 267.1 | < 2.2e-16 |
| 15 | - | - | 0.9464 | 267.1 | < 2.2e-16 |

### | Tests de validation du modèle obtenu par cette méthode

#### | Test d’homoscédasticité

C’est une représentation graphique où on trace le résidu par rapport à les valeurs prédites.

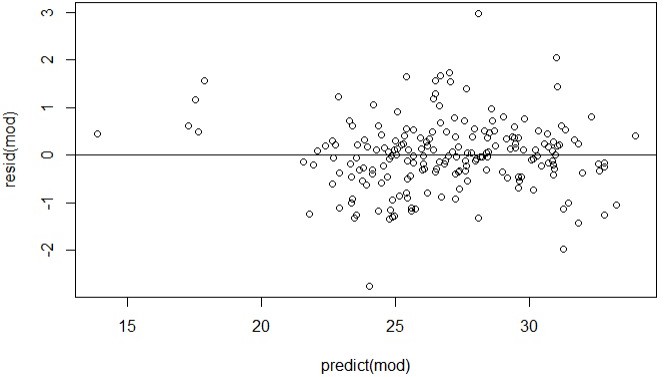


FIGURE 1.33 – Test d’homoscédasticité du modèle obtenu par la méthode pas à pas D’après la figure au-dessus, la distribution des points est assez homogène.

#### | Tests de normalité

Ils permettent de savoir si une série de données suit une loi normale.

##### \* Test de shapiro-Wilk :

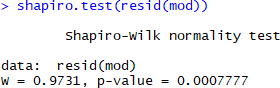


FIGURE 1.34 – Test de Shapiro-Wilk du modèle obtenu par la méthode pas à pas

On a **p-value = 0.000777** ≤ **0.05** ie la p-value est significative. L’échantillon ne suit donc pas une loi normale.

##### \* Test de Kolmogorov-Smirnov :

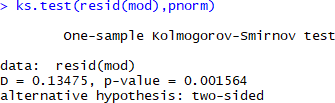


FIGURE 1.35 – Test de Ks du modèle obtenu par la méthode pas à pas

On a **p-value = 0.001564** ≤ **0.05** ie la p-value est significative. L’échantillon ne suit donc pas une loi normale.

#### | Recherche des valeurs aberrantes



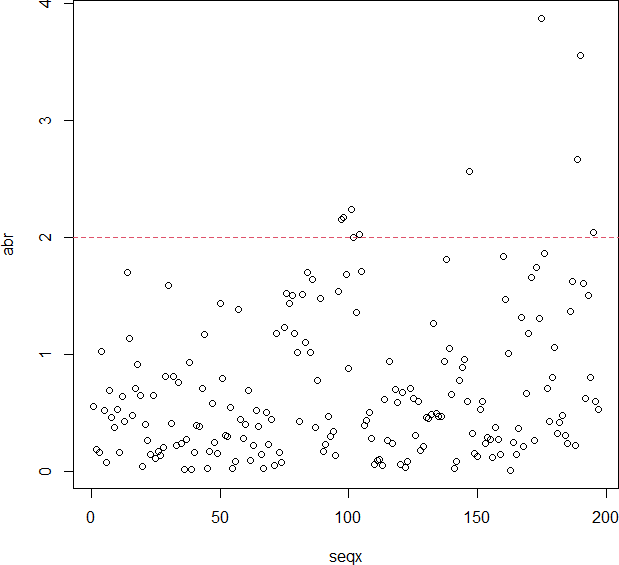


FIGURE 1.36 – Les valeurs aberrantes du modèle obtenu par la méhode pas à pas On remarque qu’il y a 9 valeurs aberrantes.

### | Le critère AIC du modèle obtenu par la méthode de pas à pas



FIGURE 1.37 – AIC du modèle obtenu par la méthode pas à pas

On remarque qu’on a pu diminué l’AIC de notre modèle de 474.7056 jusqu’à 472.145.

## | Comparaison et Conclusion

Dans ce chapitre, on a pu appliquer la régression linéaire multiple à notre jeu de données, et étudier les différentes relations entre les variables explicatives et la variable dépendante. On a vu plusieurs méthodes pour développer le meilleur modèle de pré- diction.

D’abord on s’est partie par un modèle qui utilise toutes les variables ayant un AIC = 474.7056, ensuite on s’est passé au modèle obtenu par la procédure Step qui présente un

AIC = 471.6749 significativement inférieur au modèle du départ et *R*2

*adj*

= 0.9468, d’où

il s’avère être très utile pour l’amélioration du modèle initiale.

Et enfin, un modèle obtenu par la méthode de sélection des variables pas à pas basée

sur Fisher qui présente un AIC = 472.145 et *R*2

*adj*

= 0.9464 pour une précision de 0.10 et

AIC = 471.674 pour une précision de 0.15.

De plus, tout au long ces modèles on a réaliser différents test de validation, à savoir le test d’homoscédasticité où on a remarqué une distribution homogène pour les deux mo- dèles, les tests de normalité où on a constaté que l’absence de la normalité des résidus pour les deux modèles.

→De là, on peut dire que le modèle obtenu par la procédure de Step est plus efficace.

**2**

# Classiﬁcation

## | Préparation du jeu des données

Avant d’appliquer la méthodes k-means et celle de CAH, qui sont l’objectif de ce chapitre, au jeu des données, les variables doivent être centrées et réduites.

D’abord, on commence par le centrage et la réduction des variables pour obtenir le jeu des données suivant :

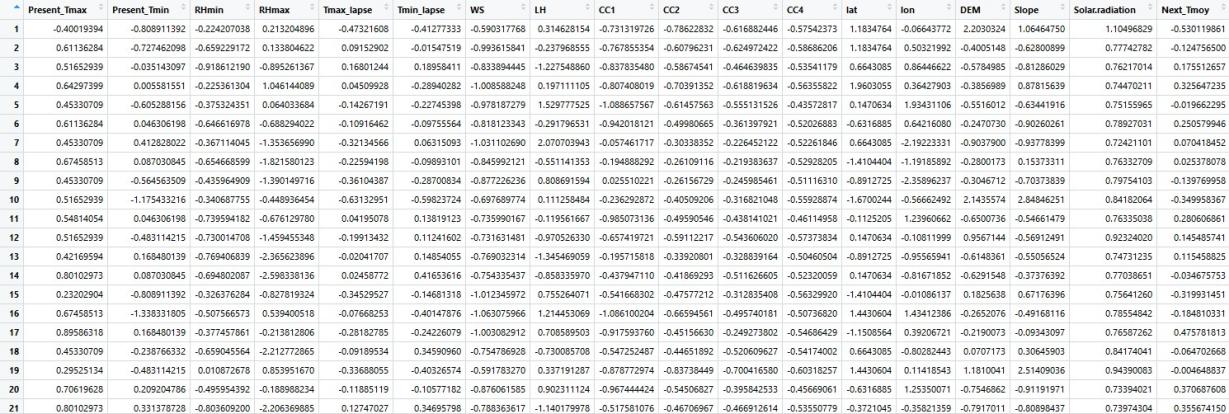


FIGURE 2.1 – Aperçu au jeu des données centrées réduites

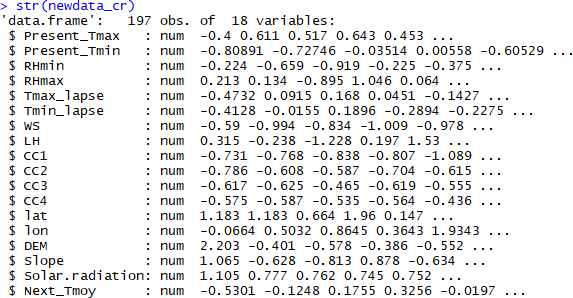


FIGURE 2.2 – Dictionnaire jeu des données centrées réduites

## | La méthode des centres mobiles K-means

Dans un premier temps, on choisit un nombre de classe aléatoire tel que k = 4.

En appliquant le K-means au jeu des données, on trouve le graphe suivant :

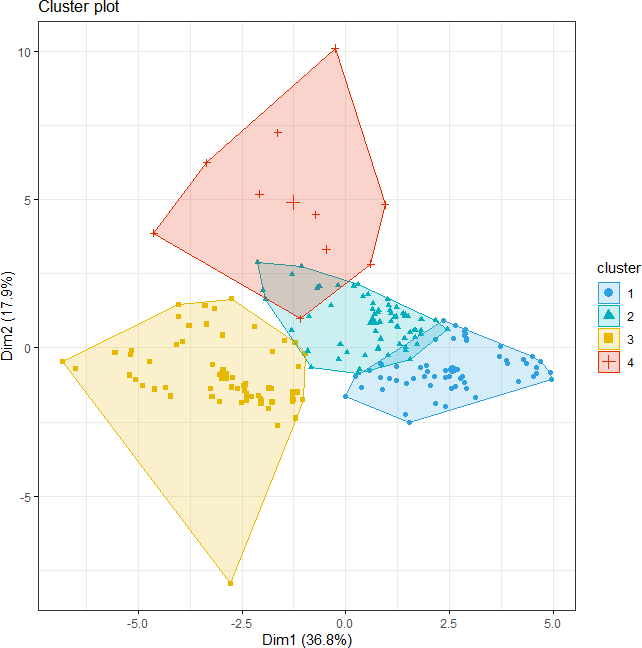


FIGURE 2.3 – Clusters crées par le k-means pour k=4

### | Taux d’inertie

On pose N étant le plus petit entier tel que le taux d’inertie expliquée de la classifi- cation à N classes est supérieur à 0.95.

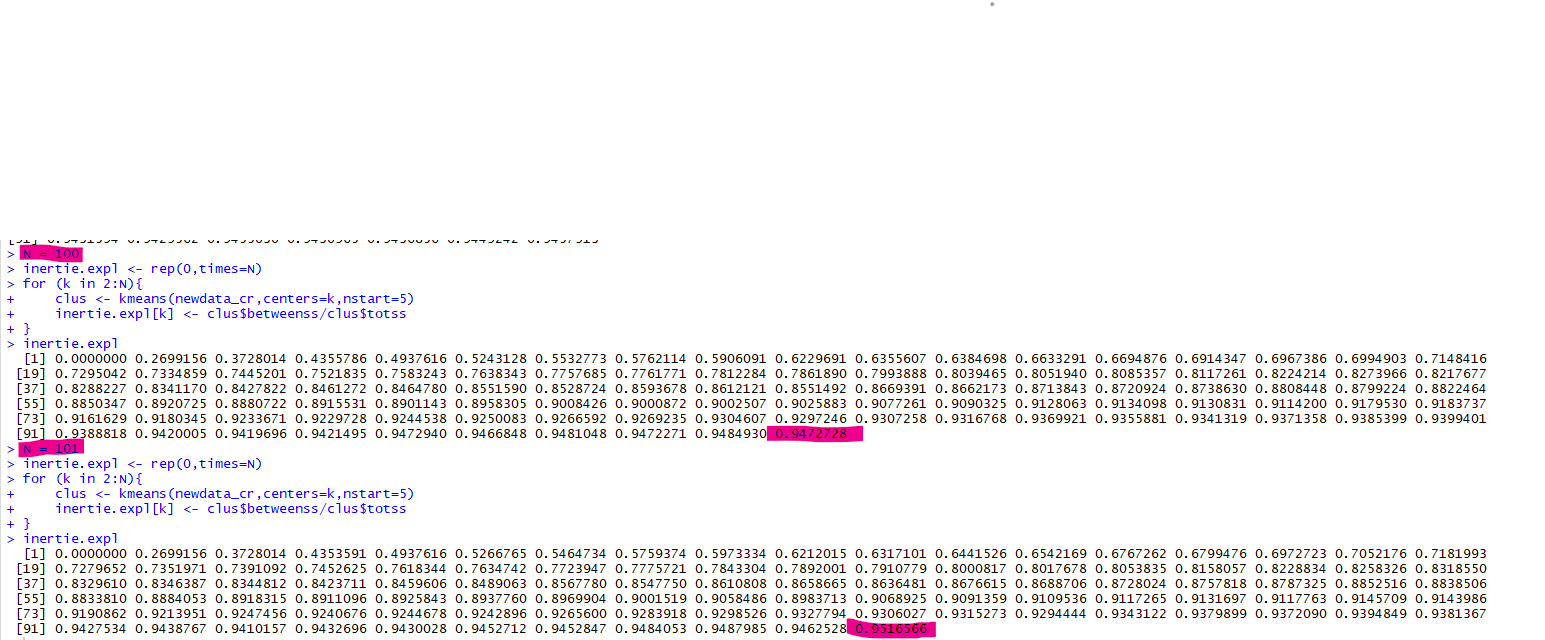


FIGURE 2.4 – Recherche du N le plus adéquat

On a (*max*(*inertie*.*expl*) = 0.9516566) ≥ 0.95 pour N = 101 et (*max*(*inertie*.*expl*) =

0.9472728) *<* 0.95 pour N = 100.

Donc On opte pour N = 101 afin qu’elle soit le plus petit entier tel que le taux d’inertie expliquée de la classification à N classes est supérieur à 0.95.

### | Détermination de nombre de classe Nc

Afin de détecter le nombre adéquat de groupes, on utilise alors la méthode Taux d’inertie expliquée par la partition.

Durant cette procédure, on fait varier le nombre de groupes et on surveille l’évolu- tion de ce taux, jusqu’à la satisfaction de la relation suivante : *var*(*I*2 ) , I étant le vecteur

*var*(*I*)

de taille N des taux d’inertie expliquée et étant le vecteur des (N-Nc) dernières valeurs des taux d’inertie expliquée .

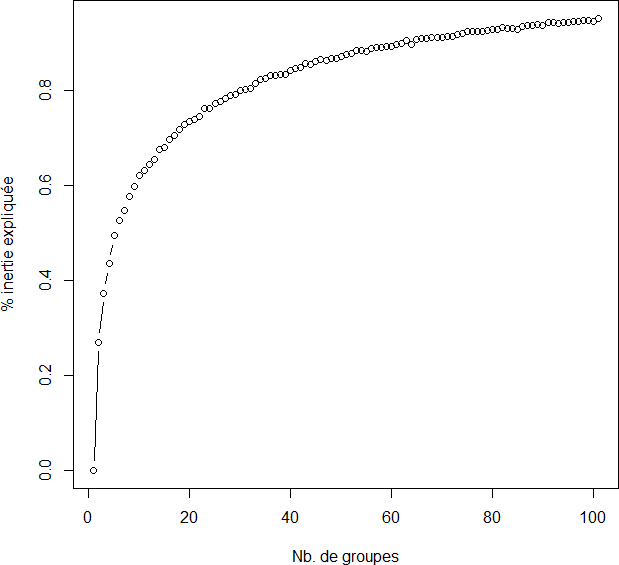


FIGURE 2.5 – Évolution de l’inertie en fonction du Nb. de groupes

D’après la figure au dessus, on remarque qu’à partir de la classe K=20, l’augmenta- tion de l’inertie expliquée diminue. Donc on part de Nc = 20 pour trouver celui le plus adéquat.

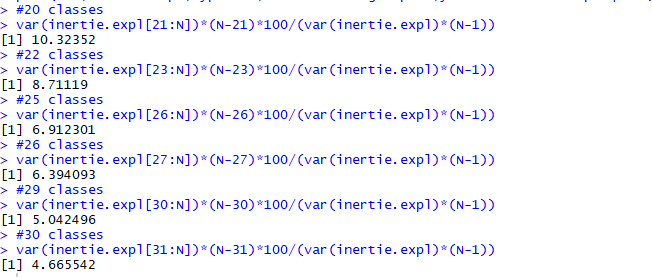


FIGURE 2.6 – Recherche du Nc le plus adéquat

D’après la figure au dessus, on a pour Nc = 30 : *var*(*I*2 )∗(*N*−31)∗100 = 4.665542 *<* 5

*var*(*I*)∗(*N*−1)

Donc le nombre de classes à retenir en utilisant la méthode de la diapositive 22 du cours est **Nc = 30**.

Par suite, en appliquant le K-means au jeu des données avec k = 30, on trouve le graphe suivant :

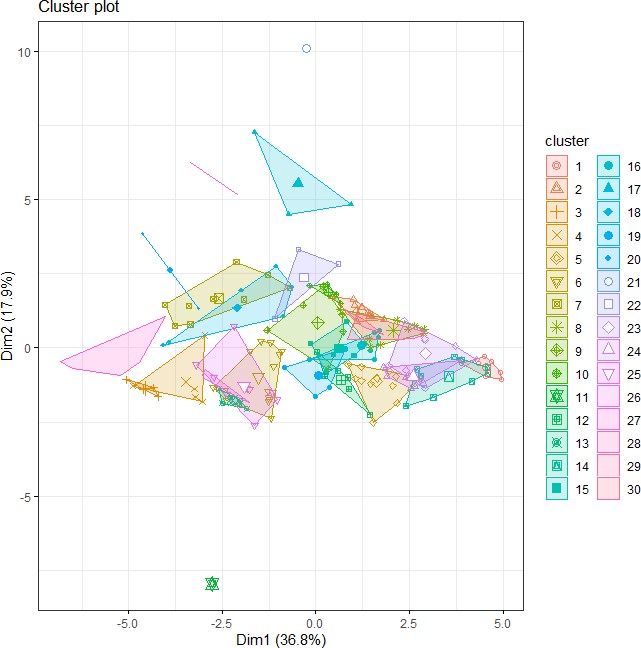


FIGURE 2.7 – Clusters crées par le k-means pour k=30

## | La Classiﬁcation Ascendante Hiérarchique CAH

En appliquant la méthode de CAH; HCPC dans library FactomineR, après le cen- trage et réduction des variables on trouve le dendogramme la distribution suivante des groupes :

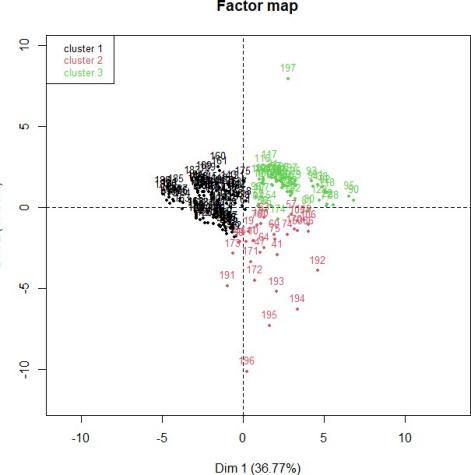
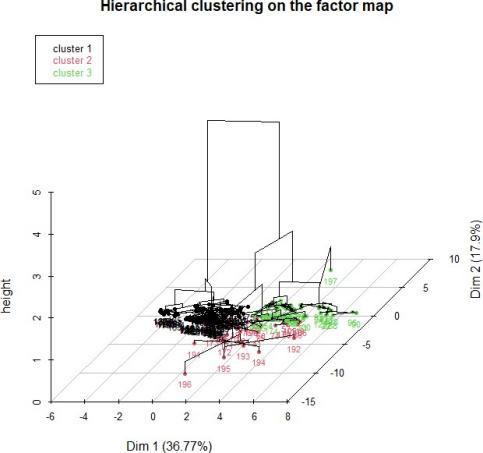
 

FIGURE 2.8 – Carte des facteurs FIGURE 2.9 – Distribution des

groupes

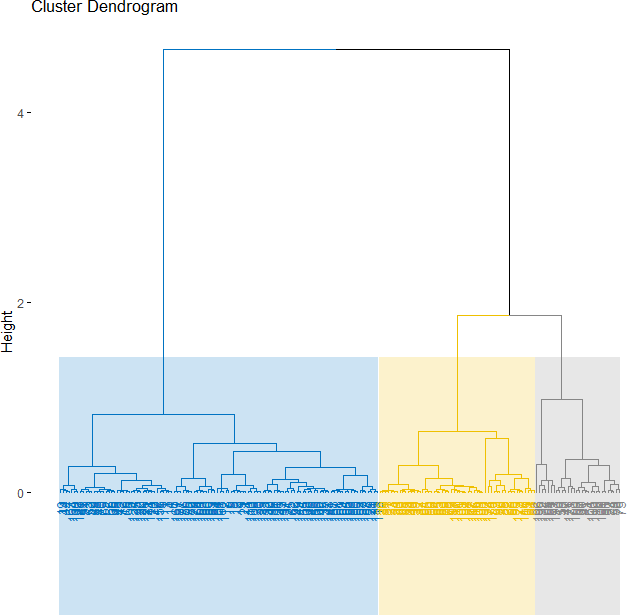


FIGURE 2.10 – Le dendrogramme de la CAH Les graphes au dessus « suggère » un découpage en 3 groupes.

### | Les variables quantitatives les plus corrélées avec la variable classiﬁcation

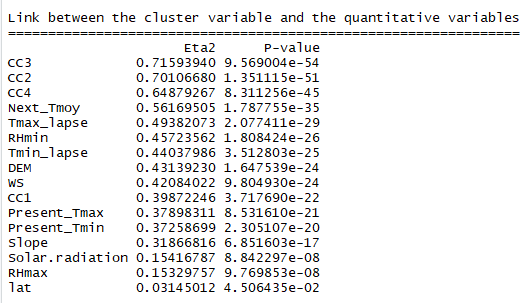


FIGURE 2.11 – Relation entre les variables quantitatives et la variable classification

D’aprés la figure au dessus, on constate que les variables **CC3**, **CC2**, **CC4**, **Next\_Tmoy**, **Tmax\_lapse**, **RHmin**, **Tmin\_lapse** sont les plus corrélées avec la variable classification.

### | Description des classes retenues par variables

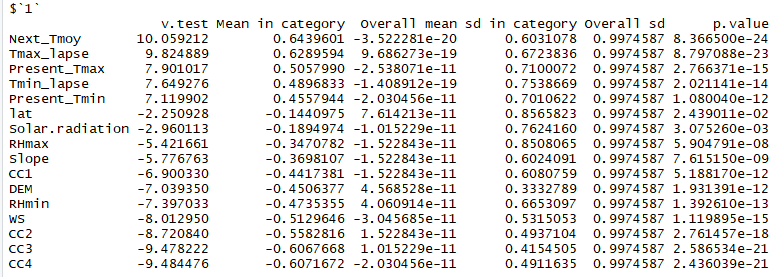


FIGURE 2.12 – Description de la classe 1 retenues par variables

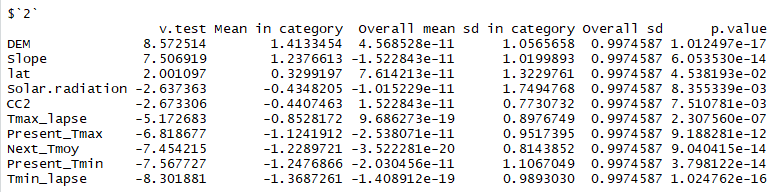


FIGURE 2.13 – Description de la classe 2 retenues par variables

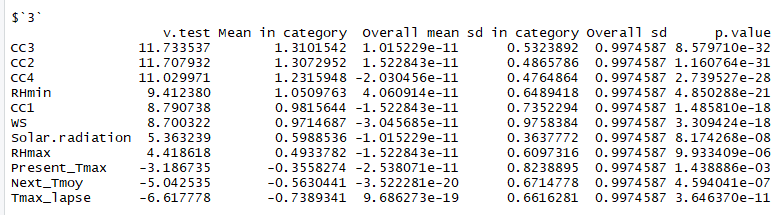


FIGURE 2.14 – Description de la classe 3 retenues par variables

### | Calcul des taux d’inertie

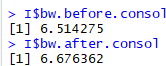


FIGURE 2.15 – Calcul des taux d’inertie

## | Comparaison

D’après les résultats de notre étude dans ce chapitre, on constate que la méthode des centres mobiles (k-means) possède un taux d’inertie plus élevée par rapport à la classi- fication ascendante hiérarchique CAH, cette différence est due aux nombres de classes premièrement qui diffèrent (30 pour k-means et 3 pour CAH), et deuxièmement aux nombres d’individus dans chaque classe. En effet, le nombre de classe dans la méthode de k-means, est supérieur au nombre des individus dans une classe dans la méthode de CAH.

D’autre part, k-means ne trouve pas nécessairement la configuration plus optimale correspondant à la fonction objective minimale, car le nombre de classes est fixé au dé- part ce qui donne une incertitude sur les résultats, par contraire CAH permet de choisir le nombre de classes de façon optimal.

**3**

# Analyse en Composante Principale

Dans cette partie, on va entamer l’analyse factorielle et plus exactement l’analyse en composante principale. On va essayer d’identifier les éventuelles similarités entre les individus, et d’identifier le comportement et la relation qu’existe entre les variables.

D’abord, on commence par un aperçu de notre jeu de données.

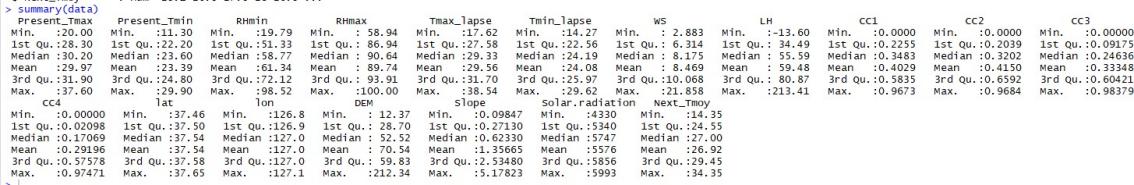


FIGURE 3.1 – Sommaire détaillé du jeu de données Ensuite, on présente la matrice de corrélation entre les variables.

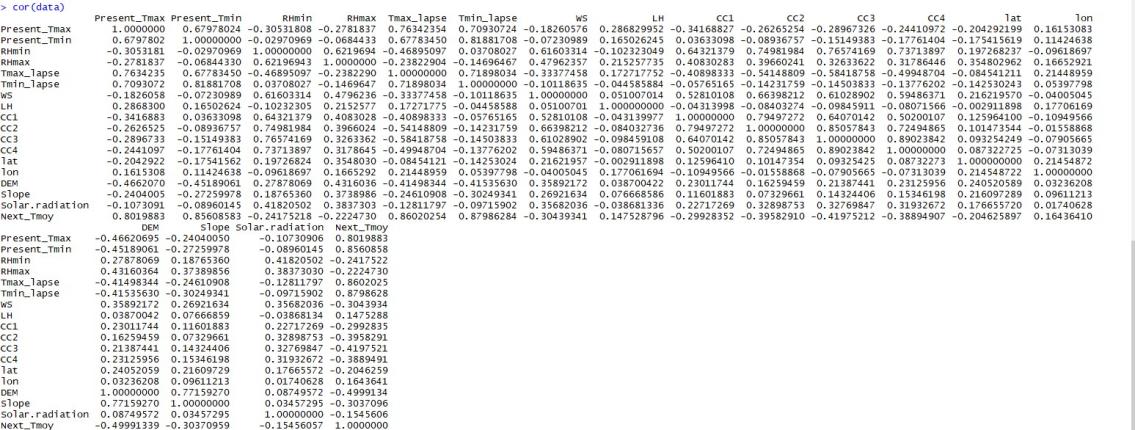


FIGURE 3.2 – corrélation entre les variables

## | ACP

### | Application de ACP normé

Comme la fonction **PCA()** dans **FactoMineR** normalise les données automatique- ment pendant l’ACP, on a pas besoin besoin de faire cette transformation avant.

En appliquant l’ACP au jeu des données, on trouve le graphe suivant :

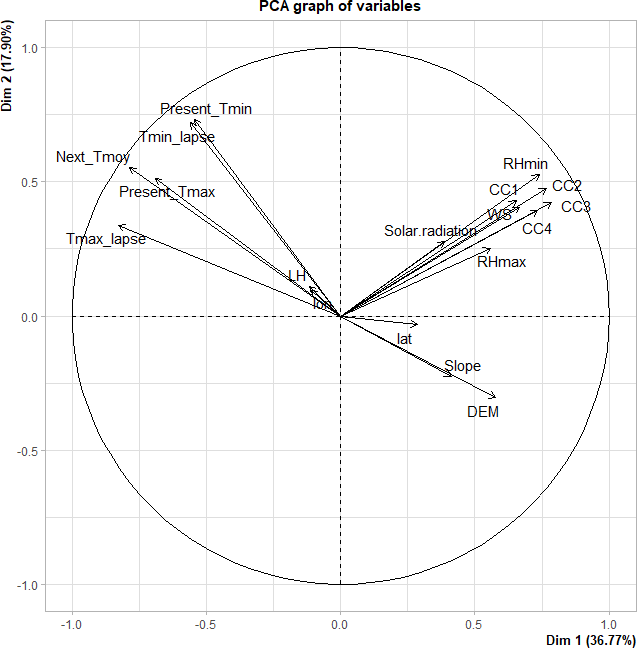


FIGURE 3.3 – Graphes de l’ACP des variables

### | La justiﬁcation du centrage et le réduction

* + - * Si on ne réduit pas le nuage : une variable à forte variance va attirer tout l’effet de l’ACP vers elle.

Dans notre cas Le centrage et la réduction des variables va de plus permettre d’éli- miner le problème de l’unité, et par la suite toutes les variables peuvent traiter de la même manière indépendamment de leurs unités , ce qui justifie largement l’approche.

### | Calcul des indices KMO et MSAI

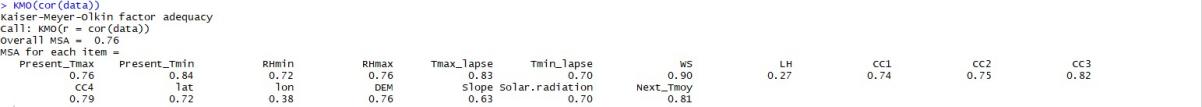


FIGURE 3.4 – Les indices KMO et MSAI

D’après la figure au dessus, on a la valeur du test de KMO est égale à 0.76 peut être qualifiée de moyenne ou même de méritoire. Elle nous indique qu’il y a une forte corrélation entre les variables.

Pour les MSAIs, on a :

* Toutes les variables sauf **LH** ont un MSAI supérieur à 0,6 donc elles sont bien représentées.
* **LH** a un MSAI égale à 0.27 donc elle est mal représentée.

⇒ D’après ce qui précède, on peut conclure que nos variables sont fortement corrélées les une aux autres.

### | Valeurs propres

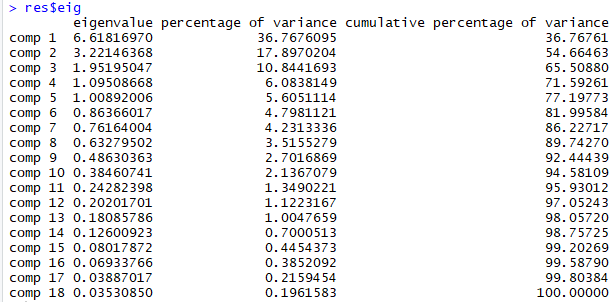


FIGURE 3.5 – Valeurs propres

La première colonne du graphe en dessus désigne la valeur des valeurs propres pour chaque composante, la deuxième colonne montre la variance tandis que la troisième calcule les variances cumulées.

Toutes ces valeurs vont être utile pour nous dans la suite de ce travail pour la déter- mination des nombres de composante.

### | Graphiques des valeurs propres

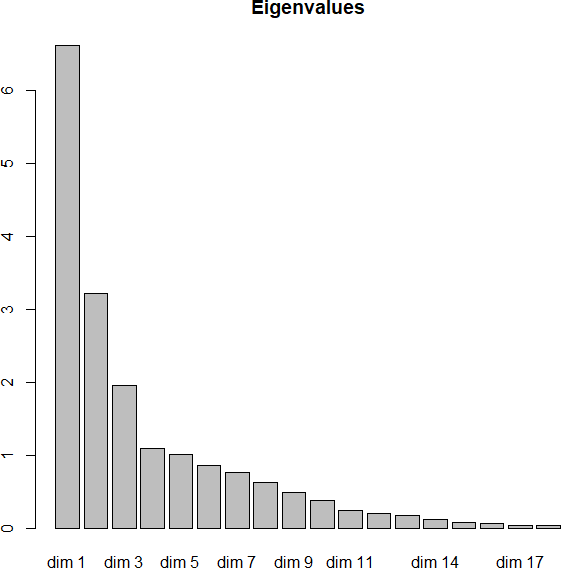
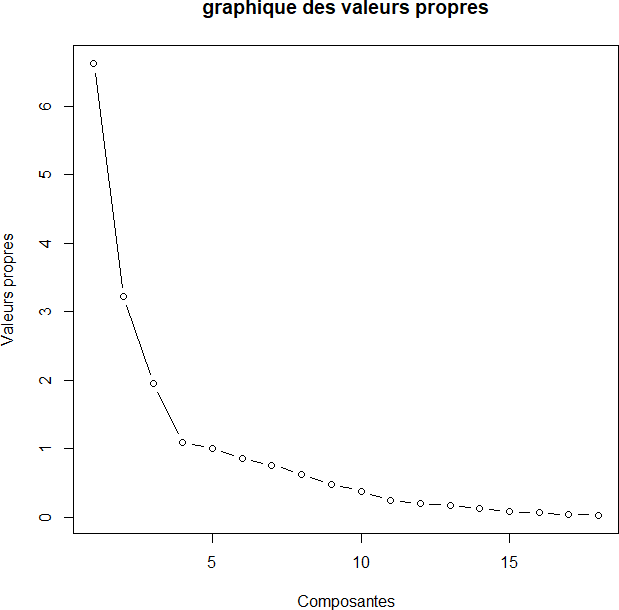


FIGURE 3.6 – Graphiques des valeurs propres

Les graphiques ci-dessus montre l’évolution des valeurs propres.

On remarque que les valeurs propres diminue d’une façon logarithmique et après la 5ème valeur les valeurs propres devient inférieur à 1.

### | La dimension du sous espace

Pour déterminer la dimension du sous espace, on utilise la régle suivante :

*Var*(*val*\_*prop*\_*limines*) *var*(*val*\_*prop*) *<* 0.05



FIGURE 3.7 – Dimension du sous espace

D’après la figure ci dessus, on constate que la dimension du sous espace est 3.

## | Nuage des variables

### | Calcul du cos2 des variables sur le sous espace

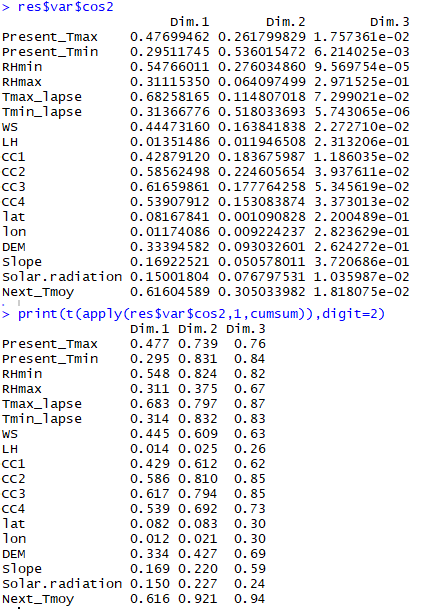


FIGURE 3.8 – cos2 des variables sur le sous espace

### | Distinction des variables

* Les variables bien représentées sont : **Next\_Tmoy**, **CC3**, **CC2**, **Tmin\_lapse**, **Tmax\_lapse**, **RHmin**, **Present\_Tmin**.
* Les variables moyennement représentées sont : **RHmax**, **Present\_Tmax**, **WS**, **CC1**, **CC4**, **Dem**.
  + Les variables faiblement représentées sont : **lon**,**LH**, **Solar.radiation**,**Slope**, **lat**.

### | La contribution des variables dans chaque axe du sous es- pace

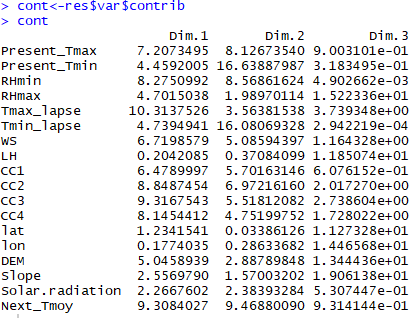


FIGURE 3.9 – Contribution des variables dans chaque axe du sous espace

### | Application de la CAH au tableau des contributions des va- riables aux axes du sous espace

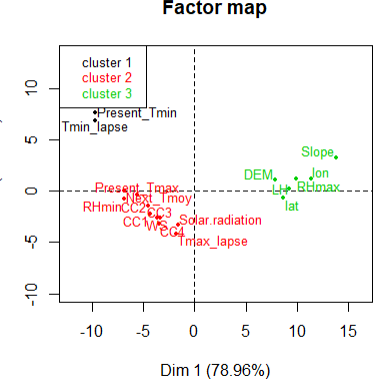
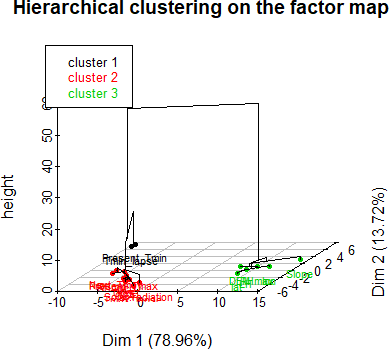
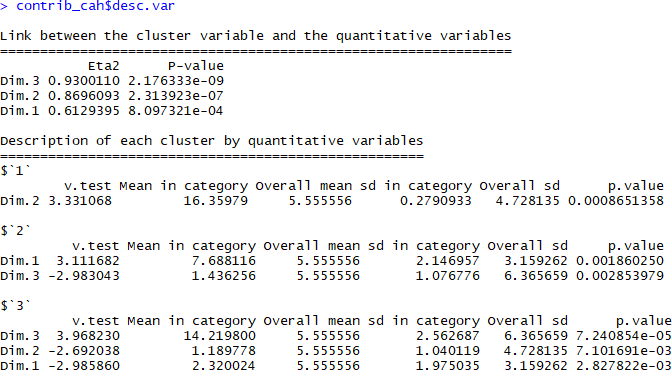
 

FIGURE 3.10 – Carte des facteurs FIGURE 3.11 – Distribution des

groupes



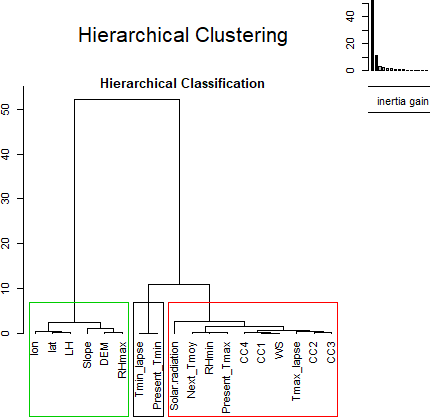
FIGURE 3.12 – Le dendrogramme de la CAH

FIGURE 3.13 – Description de la contribution aux axes par groupes

Les figures ci dessus suggère un découpage en 3 groupes. Autrement dit, les va- riables sont regroupées en 3 classes selon leur contribution aux axes principaux.

Et on remarque de plus que les variables dans le groupes 3 contribuent à l’axe 3 et ne contribuent pas aux axes 1 et 2, les variables dans le groupes 2 contribuent à l’axe 1 et ne contribuent pas à l’axe 3, alors que les variables dans le groupes 1 contribuent à l’axe 2 .

### | le nuage des variables projeté sur les 2 premiers axes

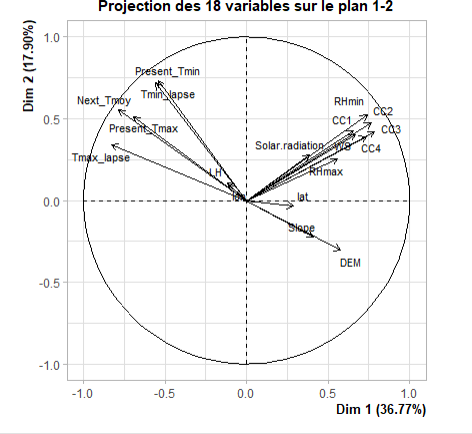


FIGURE 3.14 – le nuage des variables projeté sur les 2 premiers axes

### | Indication les variables qui sont relativement bien corrélées (positivement et négativement) avec les axes du 1er plan factoriel

* Les variables qui sont relativement bien corrélées positivement avec l’axe 1 sont :

**Solar.radiation**, **CC1**, **RHmax**, **RHmin**, **CC2**, **CC3**, **CC4**, **WS**, **lat**, **Slope**, **DEM**.

* Les variables qui sont relativement bien corrélées négativement avec l’axe 1 sont :

##### Tmax\_lapse, Present\_Tmax, Next\_Tmoy, LH, lon, Tmin\_lapse, Present\_Tmin.

* Les variables qui sont relativement bien corrélées positivement avec l’axe 2 sont : **Solar.radiation** , **CC1** , **RHmax**, **RHmin**, **CC2**, **CC3**, **CC4**, **WS**, **Tmax\_lapse**, **Present\_Tmax**, **Next\_Tmoy**, **LH**, **lon**, **Tmin\_lapse**, **Present\_Tmin**.
* Les variables qui sont relativement bien corrélées négativement avec l’axe 2 sont :

**lat**, **Slope**, **DEM**.

## | Nuage des individus

### | Le cos2 des individus sur le sous espace

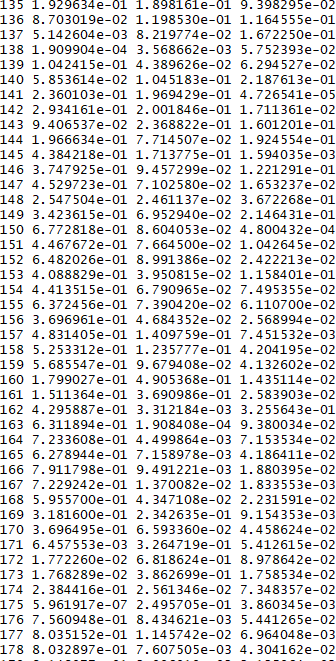
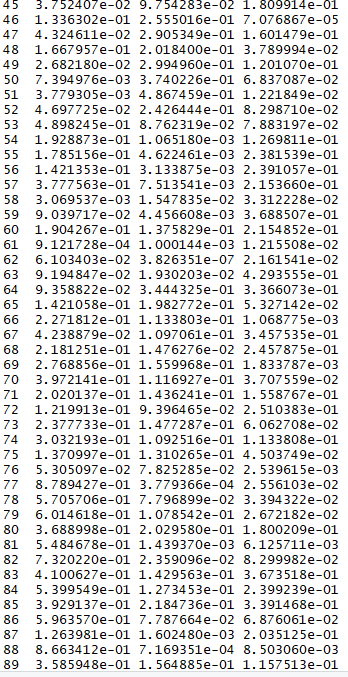
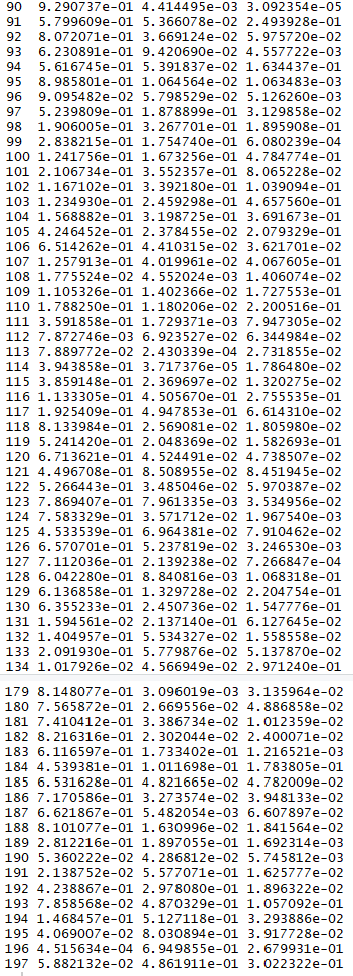
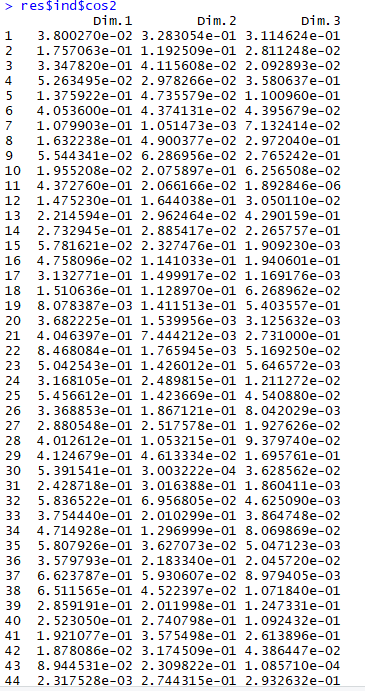


FIGURE 3.15 – le cos2 des individus sur le sous espace

### | Distinction des individus bien représentées, moyennement représentées et faiblement représentées sur le sous espace

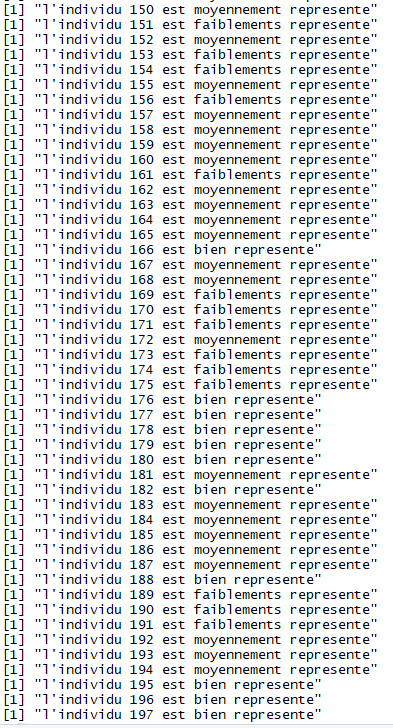
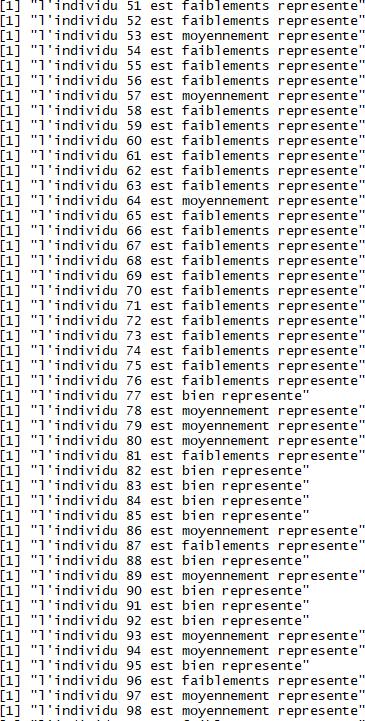
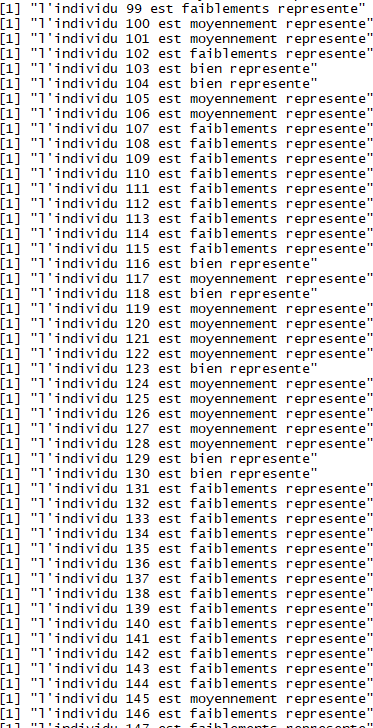
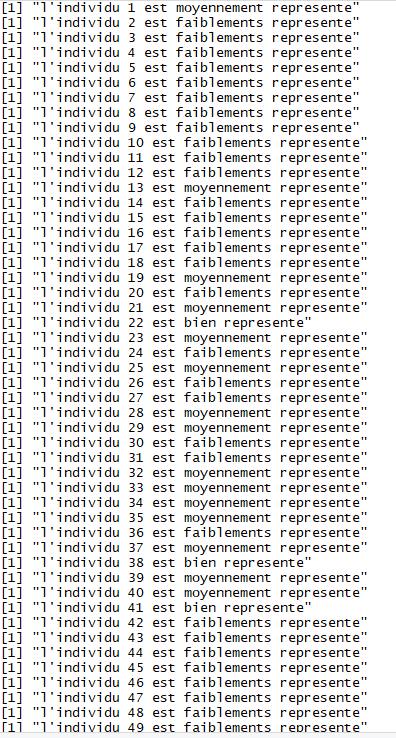


FIGURE 3.16 – Distinction des individus

### | La contribution des individus dans chaque axe du sous es- pace

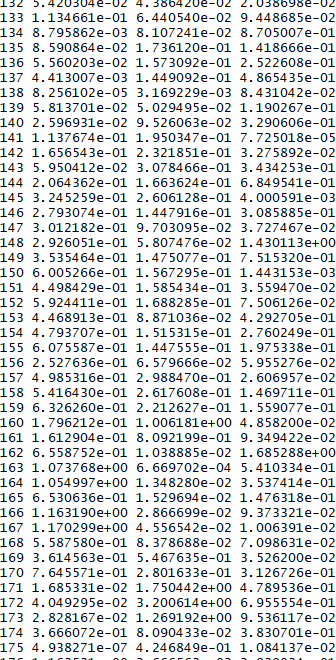
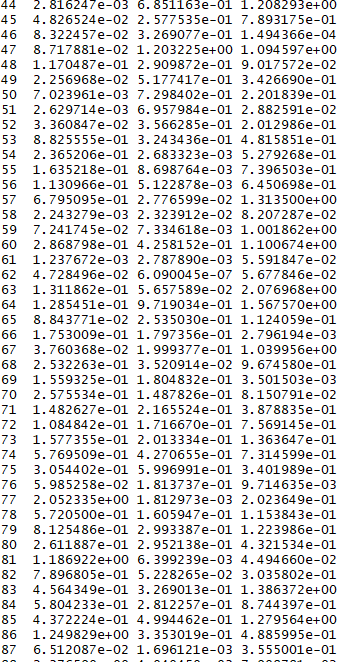
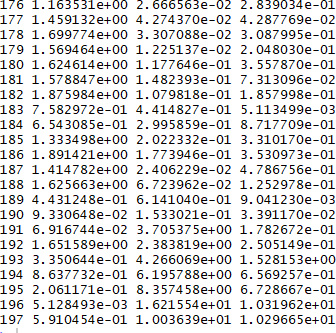
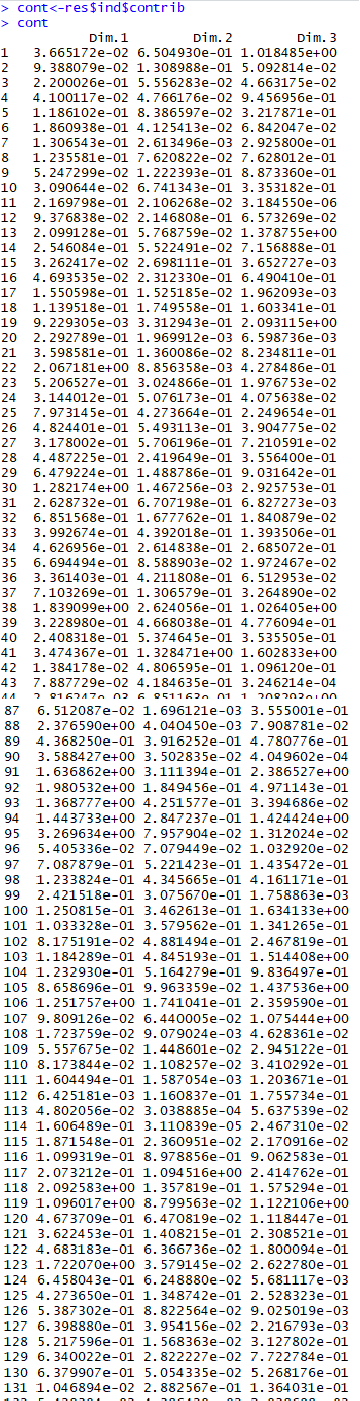


FIGURE 3.17 – Contribution des individus

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

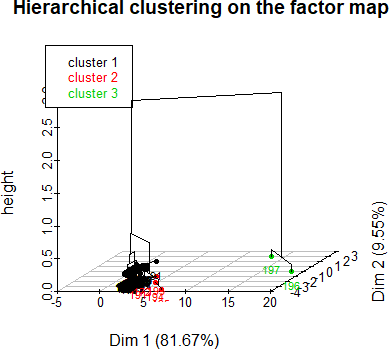
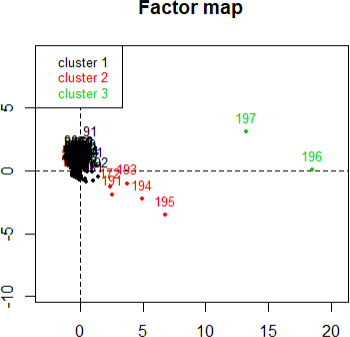
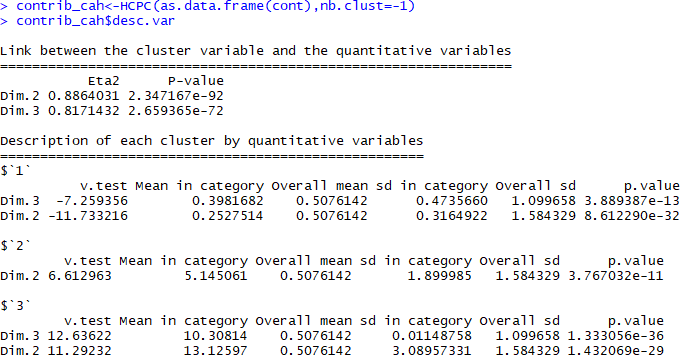


FIGURE 3.18 – Carte des facteurs FIGURE 3.19 – Distribution des



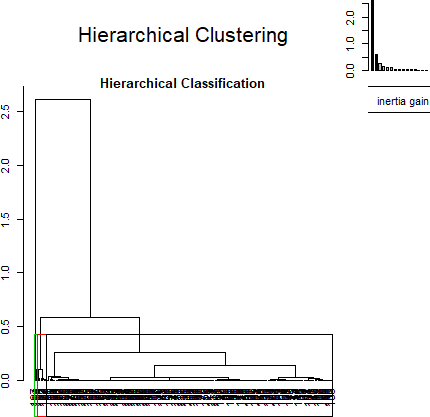
FIGURE 3.20 – Le dendrogramme de la CAH

FIGURE 3.21 – Description de la contribution aux axes par groupes

groupes

Les figures ci dessus suggère un découpage en 3 groupes. Autrement dit, les indivi- dus sont regroupés en 3 classes selon leur contribution aux axes principaux.

Et on remarque de plus que les individus dans le groupes 1 ne contribuent pas aux axes 3 et 2, que les individus dans le groupe 2 contribuent a l’axe 2 , de plus les individus dans le groupe 3 contribuent aux axes 3 et 2 .

Ainsi aucun individu ne contribue a l’axe 1.

### | Le nuage des individus projeté sur les 2 premiers axes

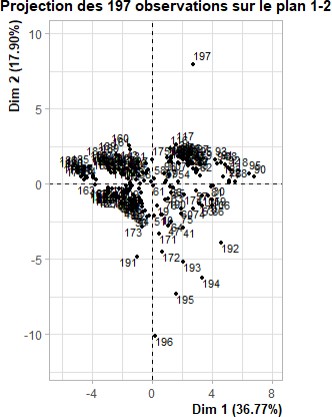


FIGURE 3.22 – le nuage des individus projeté sur les 2 premiers axes

## | Conclusion

Dans ce chapitre, on a pu appliquer l’analyse en composante principale qui est une méthode puissante pour synthétiser les jeu de données décrites par plusieurs variables quantitatives.

En effet, grâce à cette méthode on a réduit les 18 variables en 3 composantes prin- cipales. Chose qui nous a permis de mettre en évidence les relations globales existant entre les variables et les individus, ainsi leurs visualiser dans un plan factoriel.

**4**

# Analyse des Correspondances

**Multiples**

Après avoir étudier dans le chapitre précèdent l’analyse en composante principale dont son principal but est la réduction de la dimension de l’espace, on va voir main- tenant un autre type de l’analyse factorielle qui est l’analyse en composante multiple qui permet d’étudier les liens entre les variables et les individus qui sont décrit par des variables qualitatives.

## | Préparation du jeu de données

### | Discrétisation des variables

D’abord, on part des variables quantitatives comme suit :

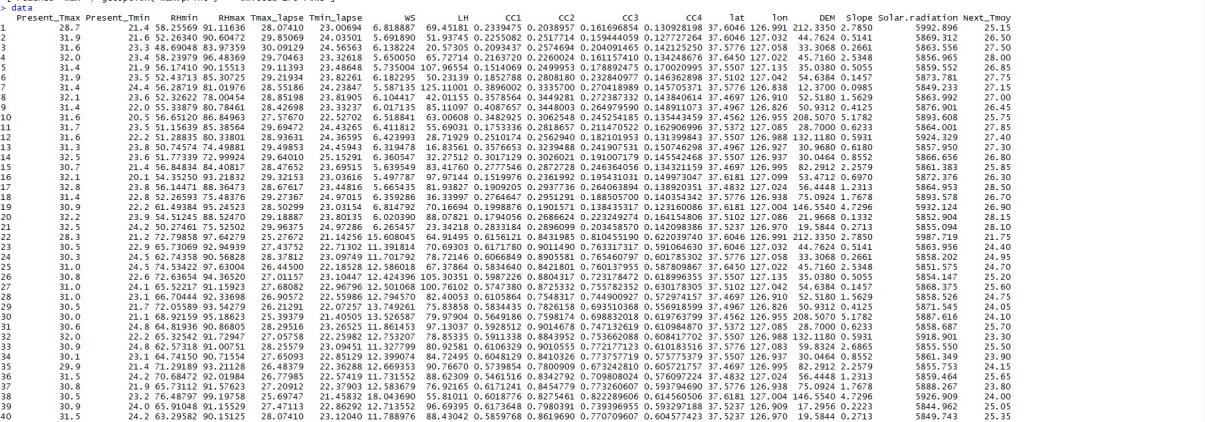


FIGURE 4.1 – Aperçu aux variables quantitatives

Ensuite, la transformation des variables quantitatives en des variables qualitatives est faite par la méthode de la classification kmeans.

* Si le taux de discrétisation est supérieur à 0.5, on transforme la variable en 3 mo-

dalité.

* Sinon, il est possible de la discrétiser en 4 classes.



FIGURE 4.2 – Taux de discrétisation

D’après la figure au dessus, on a pour toutes les variables le taux de discrétisation est supérieur à 0.5, donc pour chacune on aura une transforamation en 3 modalités.

On aura alors un nouveau jeu de données ayant des variables qualitatives comme suit :

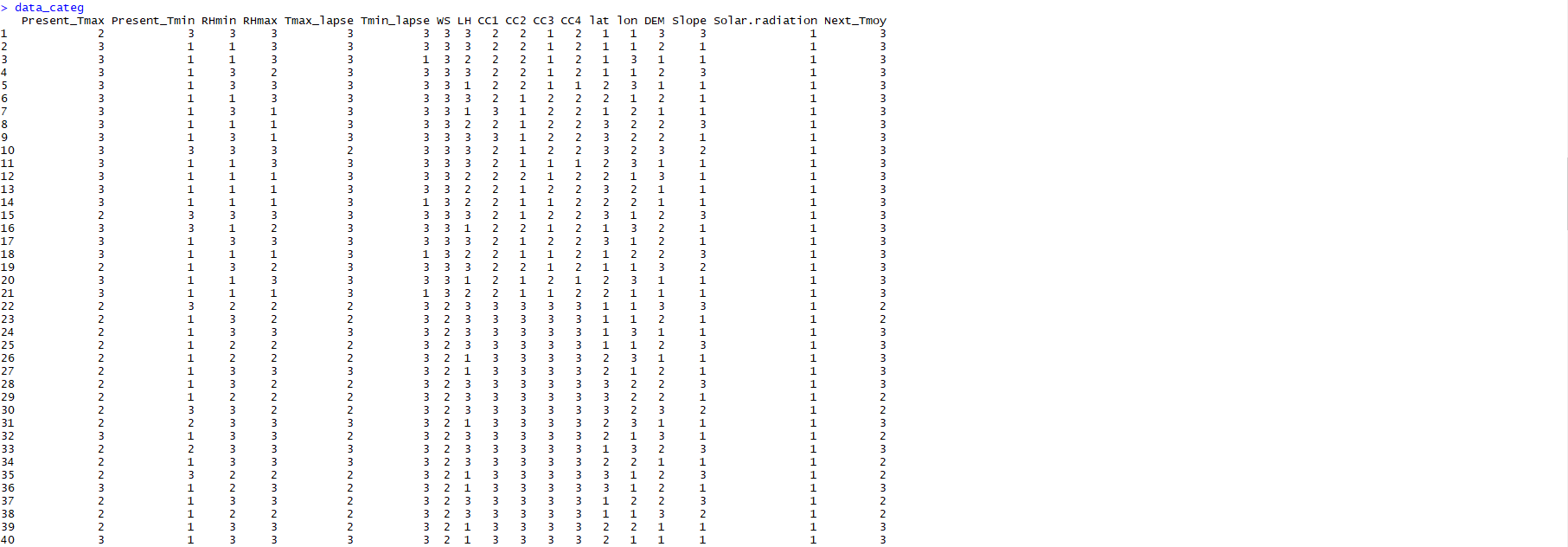


FIGURE 4.3 – Aperçu aux variables qualitatives

Ces tableaux, des variables quantitatives et des variables qualitatives ainsi que de taux de discrétisation de chaque transformation, ont été regroupé dans un fichier Excel. (Voir fichier "Pred\_temp\_moy\_ACM.xlsx")



FIGURE 4.4 – Regroupement dans un fichier Excel

### | Tableau Disjonctif Complet

Grâce au tableau qui contient les modalités de chaque variables, on peut construire un tableau disjonctif. Dans ce dernier, une variable qualitative à 3 modalités est rempla- cée par 3 variables binaires, chacune correspondant à une des modalités.

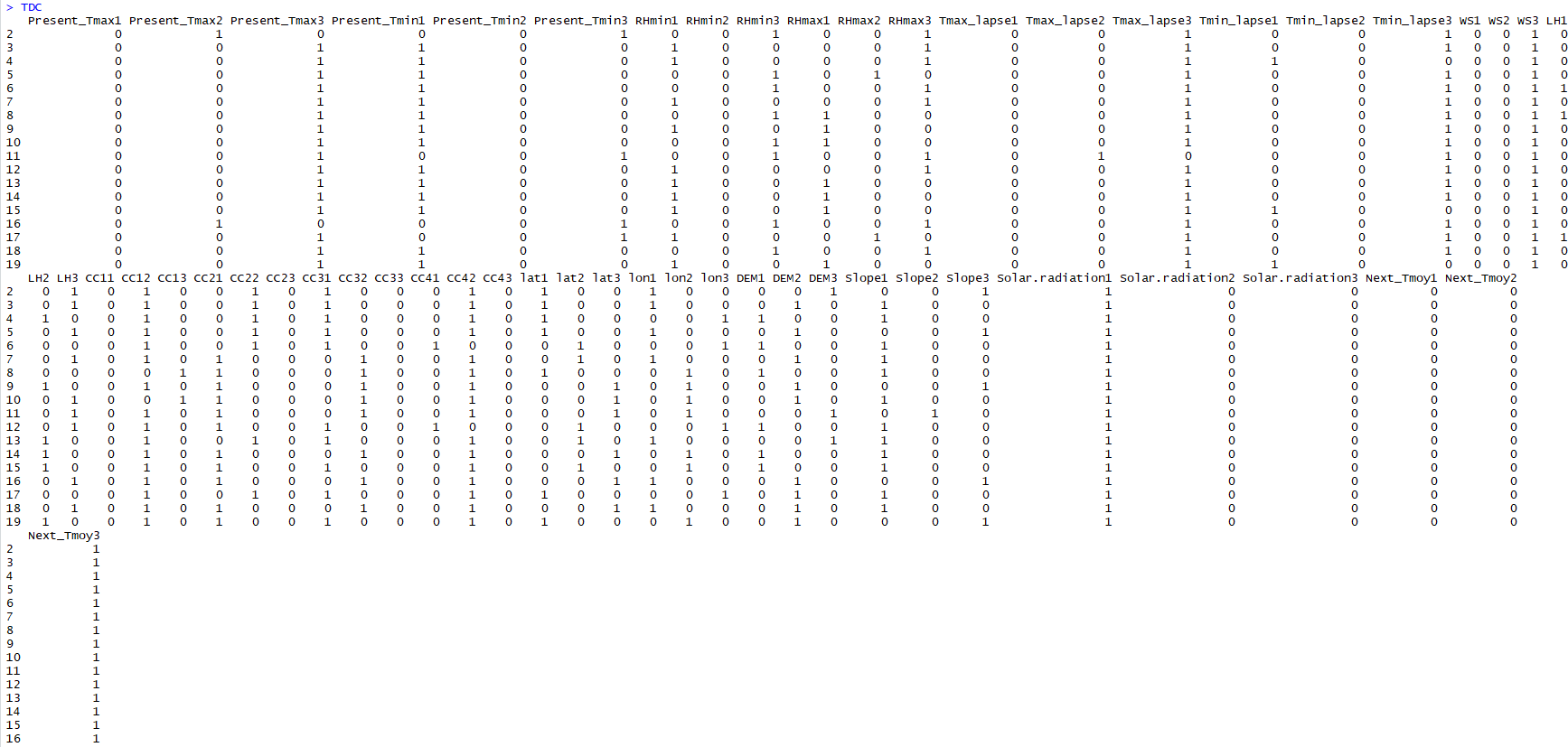


FIGURE 4.5 – Aperçu au tableau disjonctif complet

### | La fréquence de chaque modalité

Le calcul du fréquence de chaque modalité permet de détecter les modalités les plus fréquentes et celles les moins fréquentes. binaires, chacune correspondant à une des modalités.

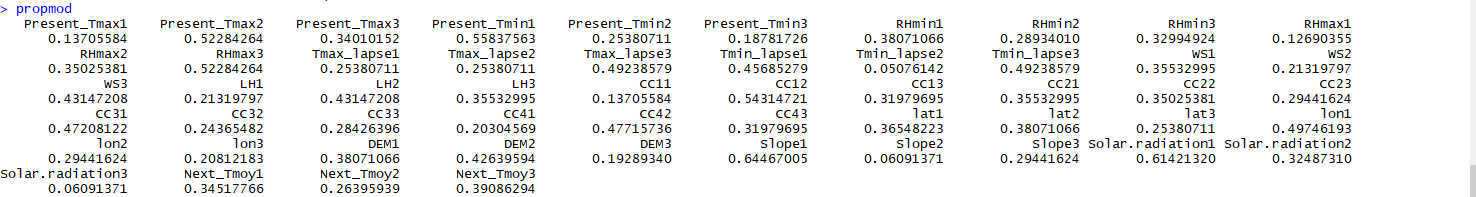


FIGURE 4.6 – La fréquence de chaque modalité

D’après la figure au dessus, on constate que toutes les fréquences sont supérieures à 0.01, donc absence de modalité rare, et par la suite on n’élimine aucune variable.

## | ACM

On applique l’ACM au tableau qualitative.

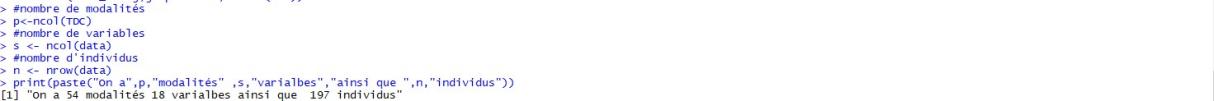


FIGURE 4.7 – Données générales

D’après la figure au dessus, on constate qu’on 54 modalités, 18 variables et 197 ob- servations.

### | Valeurs propres

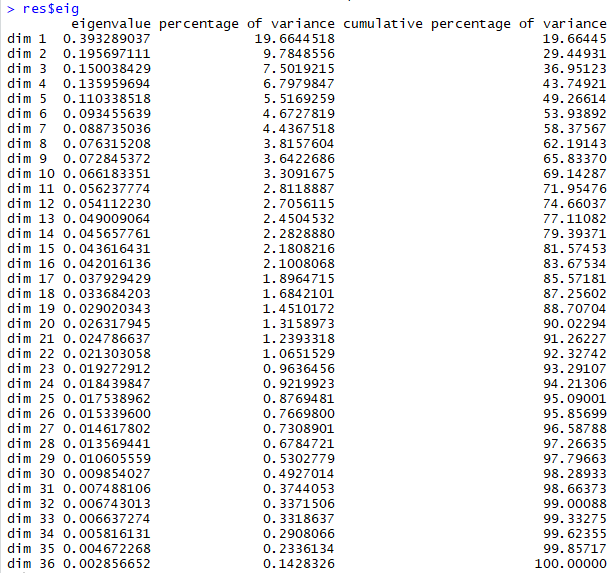


FIGURE 4.8 – Valeurs propres en ACM

La première colonne du graphe en dessus désigne la valeur des valeurs propres pour chaque composante, la deuxième colonne montre la variance tandis que la troisième calcule les variances cumulées.

Toutes ces valeurs vont être utile pour nous dans la suite de ce travail pour la déter- mination des nombres de composantes.

### | Le graphique des valeurs propres

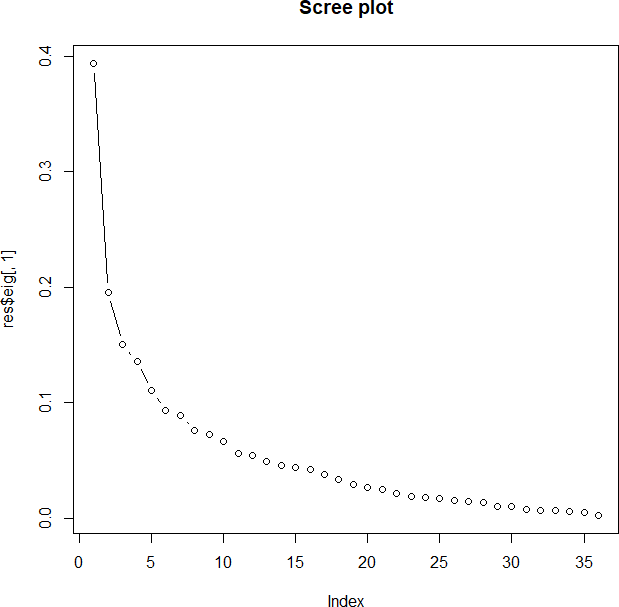


FIGURE 4.9 – Le graphique des valeurs propres en ACM

Les graphiques ci-dessus montre l’évolution des valeurs propres.

On remarque que les valeurs propres diminue d’une façon logarithmique.

### | La dimension du sous espace

Pour déterminer la dimension du sous espace, on utilise la régle suivante :

*Var*(*val*\_*prop*\_*limines*) *var*(*val*\_*prop*) *<* 0.05



FIGURE 4.10 – Dimension du sous espace en ACM

Donc La dimension du sous espace est donc égal a 9 .

## | Nuage de Modalités

### | Calcul du cos2 des modalités sur le sous espace

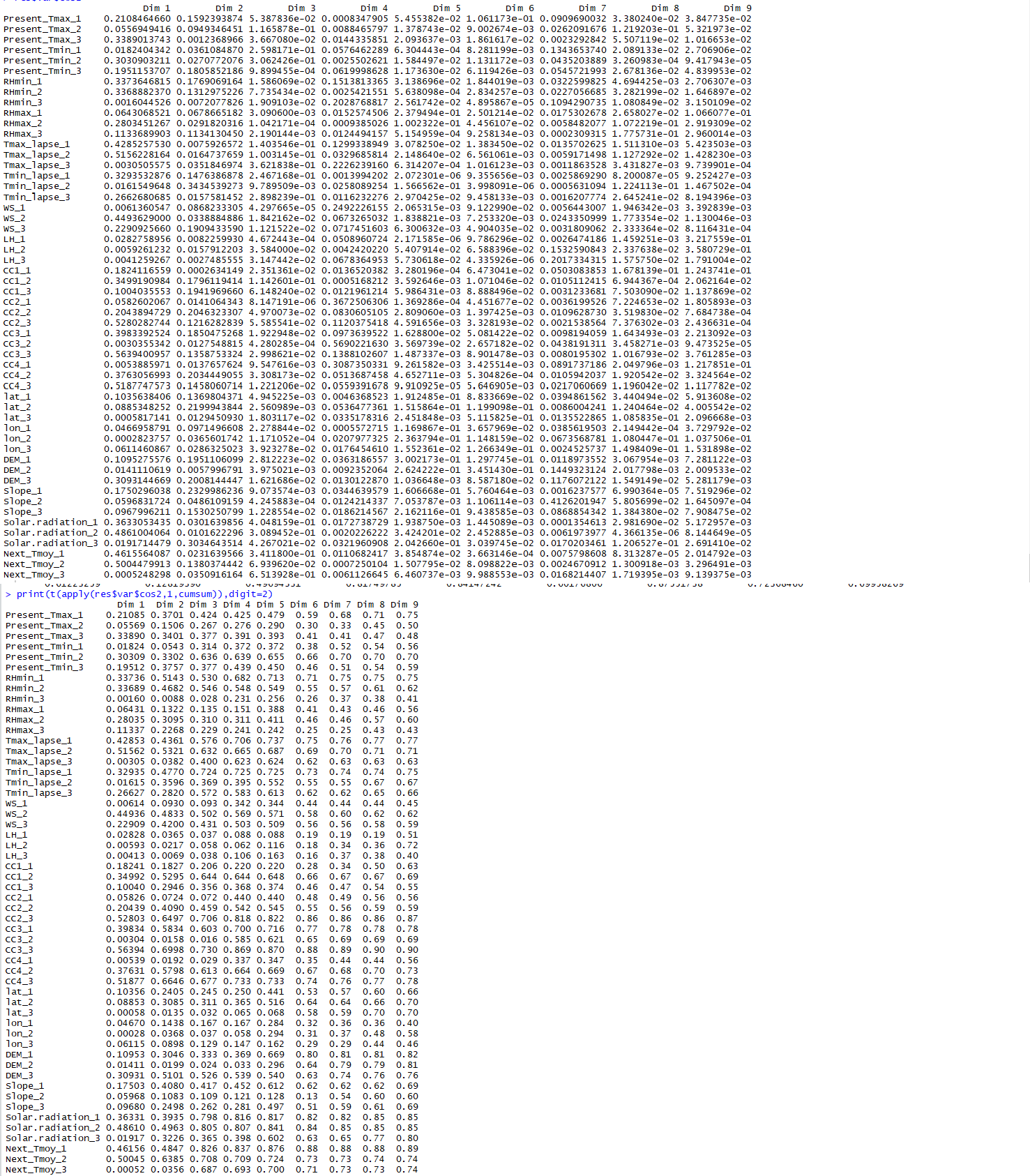


FIGURE 4.11 – cos2 des modalités sur le sous espace

### | Distinction des modalités

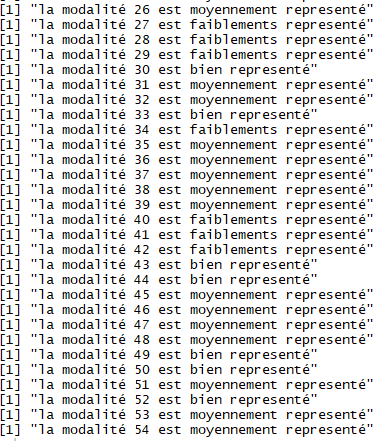
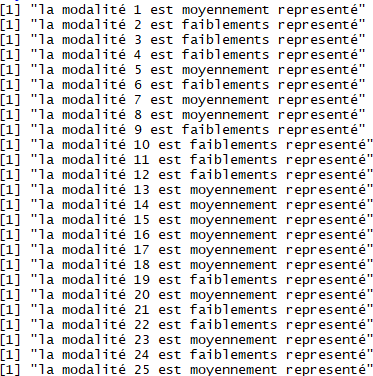


FIGURE 4.12 – Distinction des modalités

### | La contribution des modalités dans chaque axe du sous es- pace

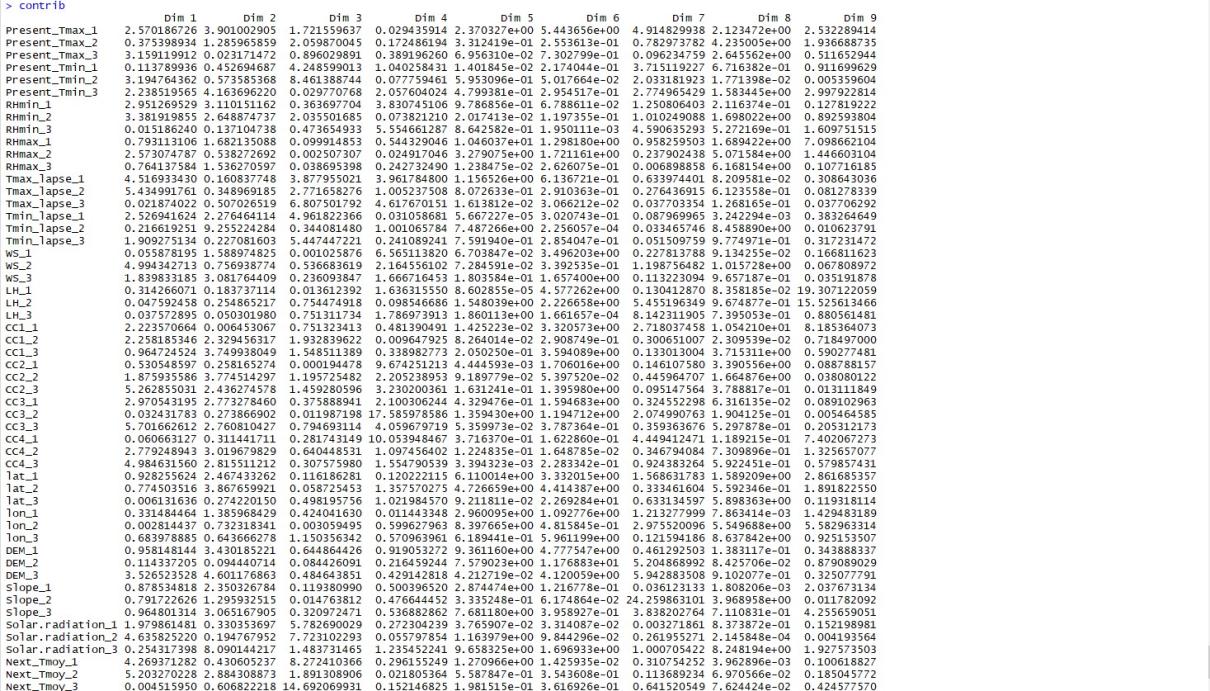


FIGURE 4.13 – La contribution des modalités dans chaque axe du sous espace en ACM

### | Application de la CAH au tableau des contributions des mo- dalités aux axes du sous espace

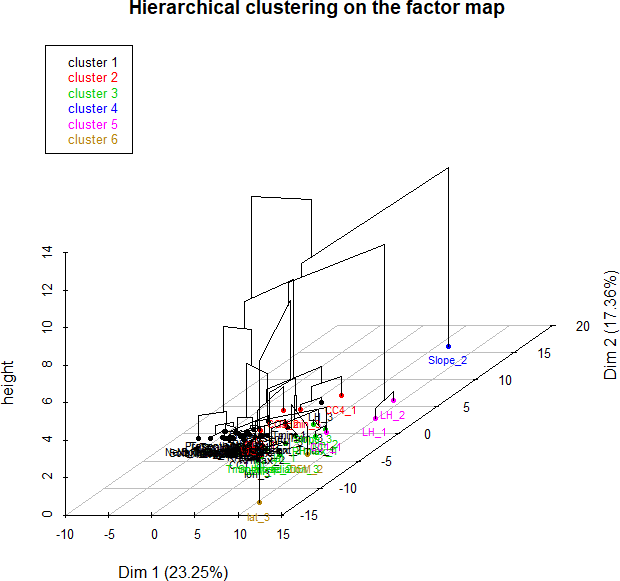
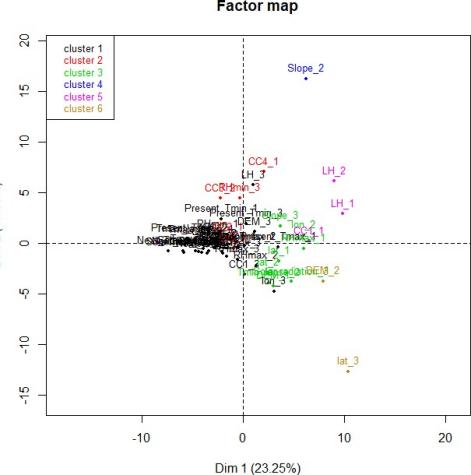


FIGURE 4.14 – Carte des facteurs FIGURE 4.15 – Distribution des

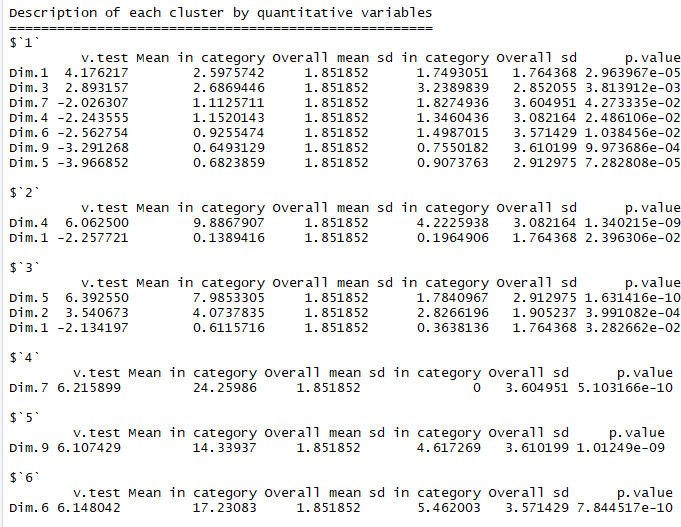
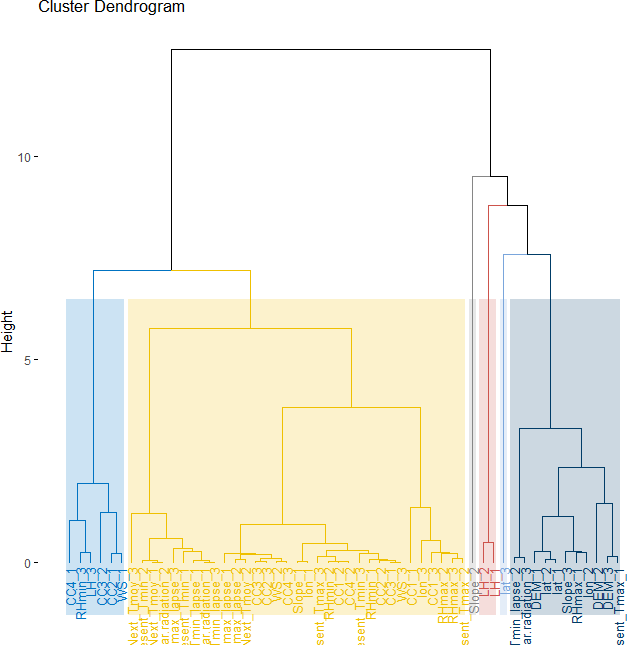


FIGURE 4.16 – Le dendrogramme de la CAH

FIGURE 4.17 – Description des Mo- dalités retenue par classes

groupes

Les figures ci dessus suggère un découpage en 6 groupes. Autrement dit, les moda- lités sont regroupées en 6 classes selon leur contribution aux axes principaux.

Et on constate de plus que :

* Les modalités du premier cluster contribuent aux axes 1 et 3 et ne contribuent pas aux axes 7,8,9,6 et 5.
* Les modalités du deuxieme cluster contribuent a l’axe 4 et ne contribuent pas à l’axe

1.

* + Les modalités du troisieme cluster contribuent aux axes 5,8 et 2.
  + Les modalités du quatrieme cluster contribuent à l’axe 9.
  + Les modalités du cinquieme cluster contribuent à l’axe 7.
  + Les modalités du sixieme cluster cluster contribuent à l’axe 6.

### | Le nuage des modalités projeté sur les 2 premiers axes

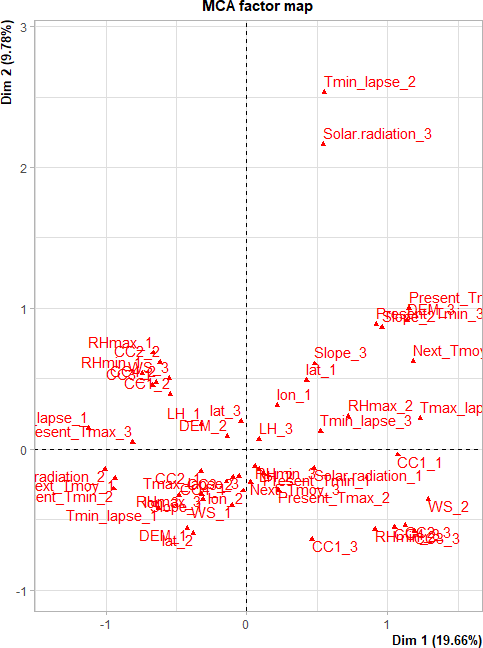


FIGURE 4.18 – Le nuage des modalités projeté sur les 2 premiers axes

## | Nuage des individus

### | Calcul du cos2 des individus sur le sous espace

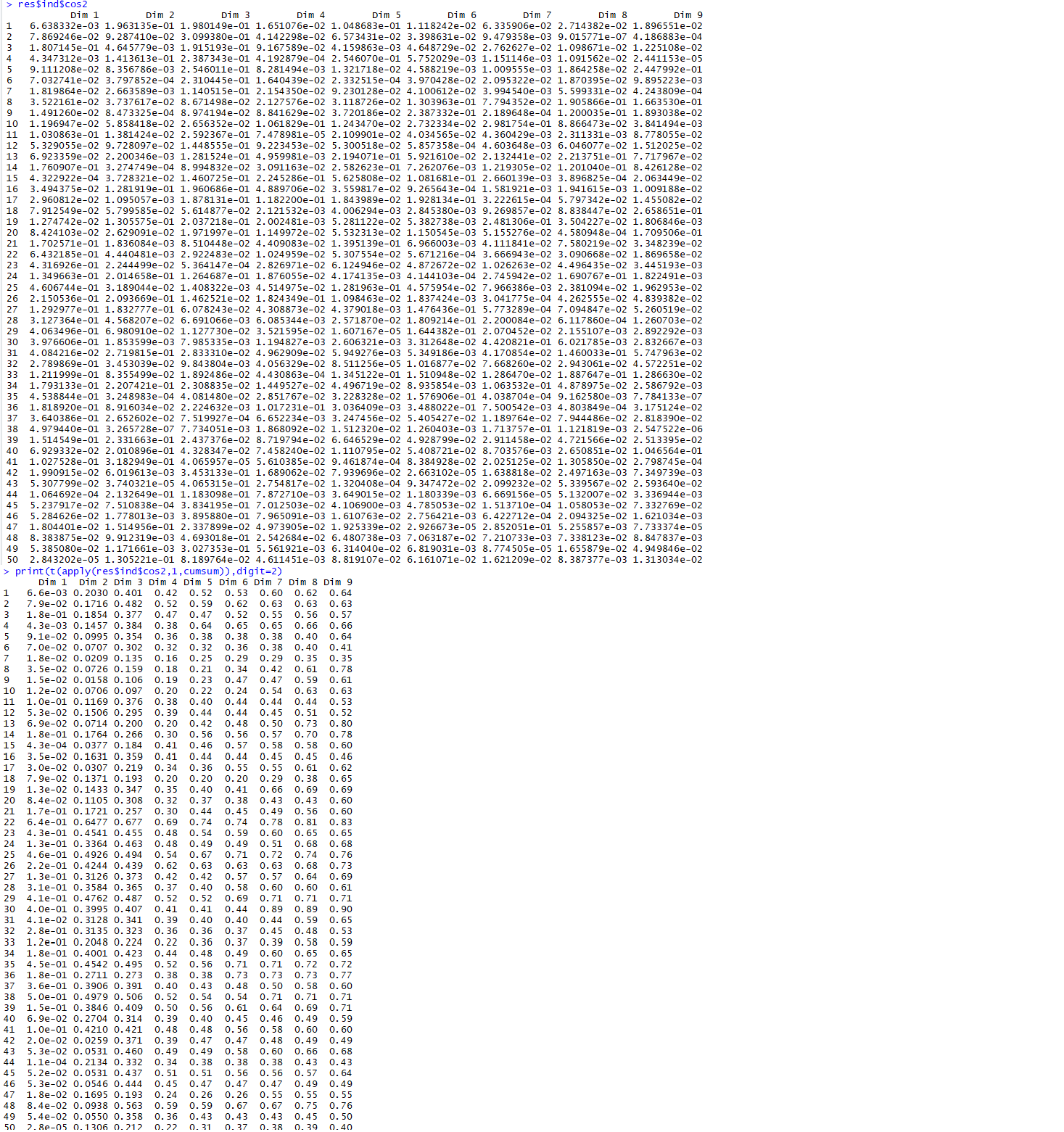
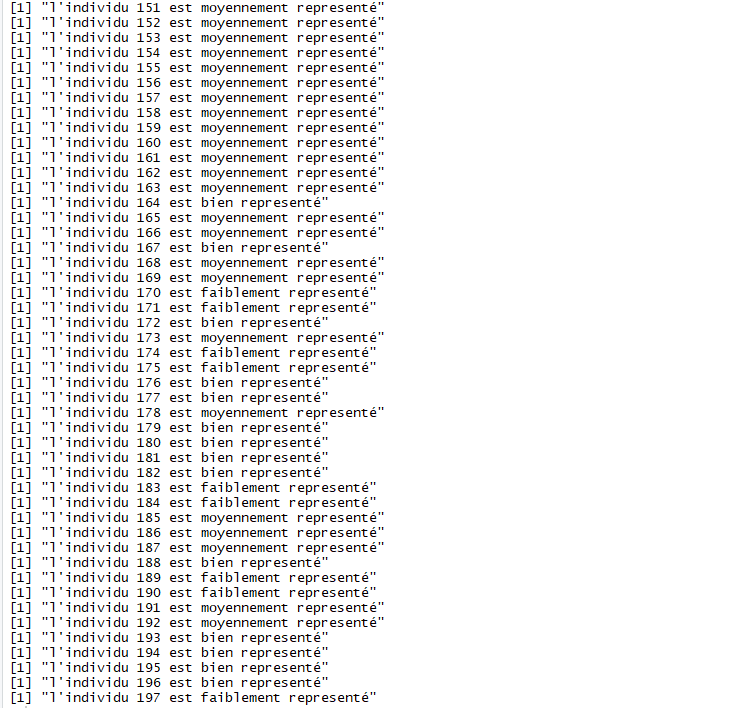
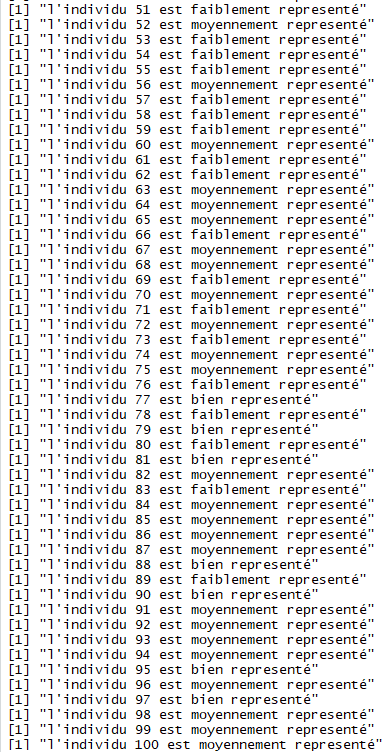
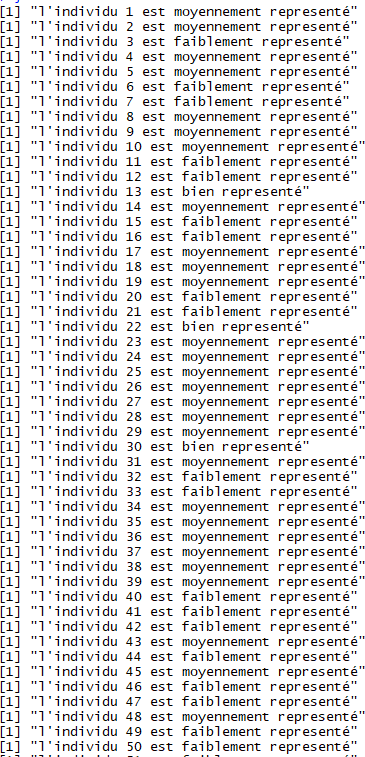


FIGURE 4.19 – Aperçu au cos2 des individus sur le sous espace en ACM

### | Distinction des individus

FIGURE 4.20 – Distinction des individus en ACM



### | La contribution des individus dans chaque axe du sous es- pace

groupes

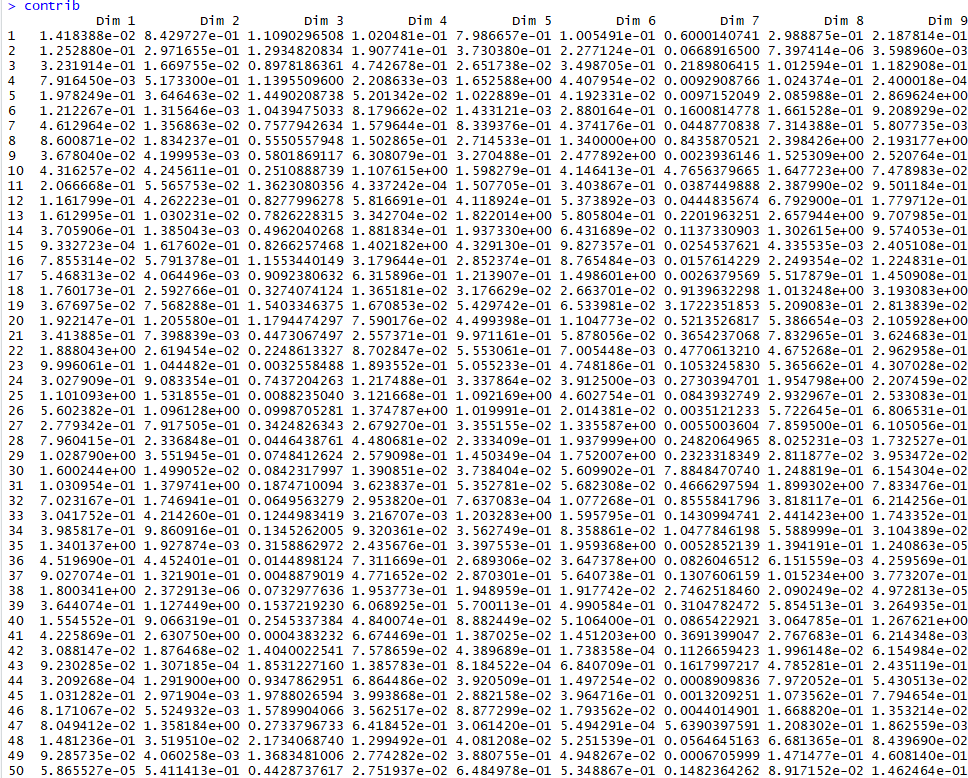


FIGURE 4.21 – Aperçu à la contribution des individus dans chaque axe du sous espace

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

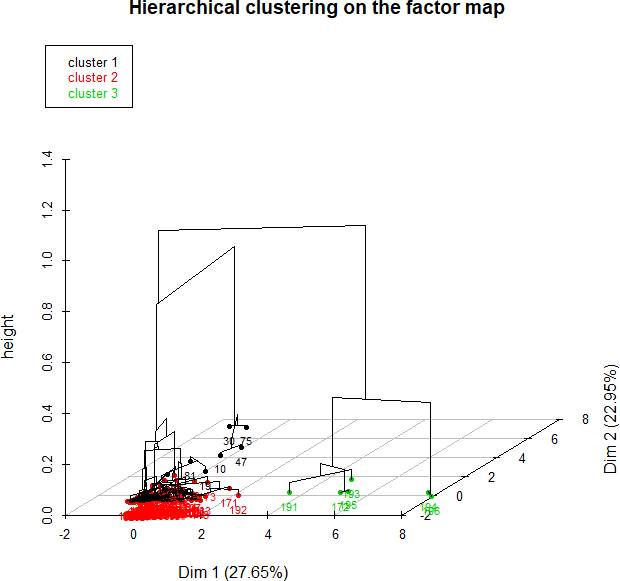
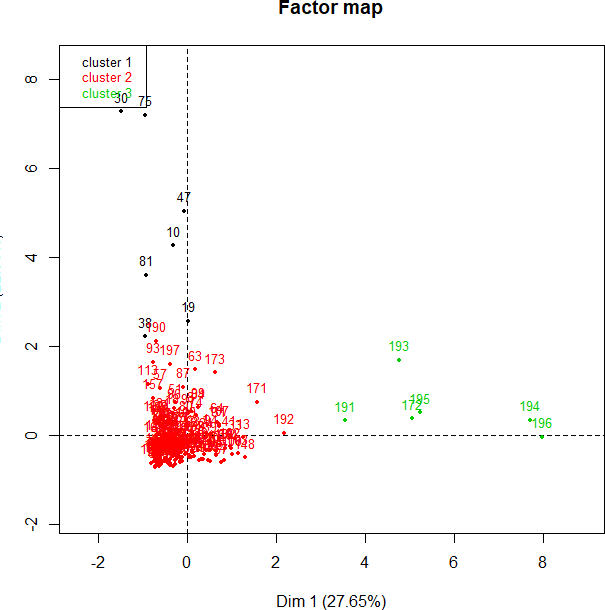
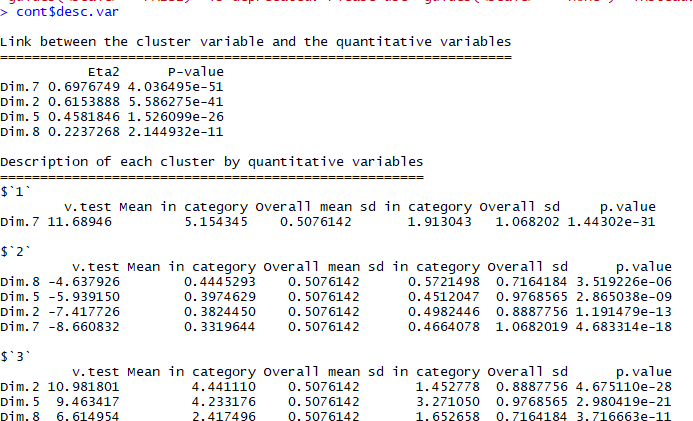


FIGURE 4.22 – Carte des facteurs FIGURE 4.23 – Distribution des



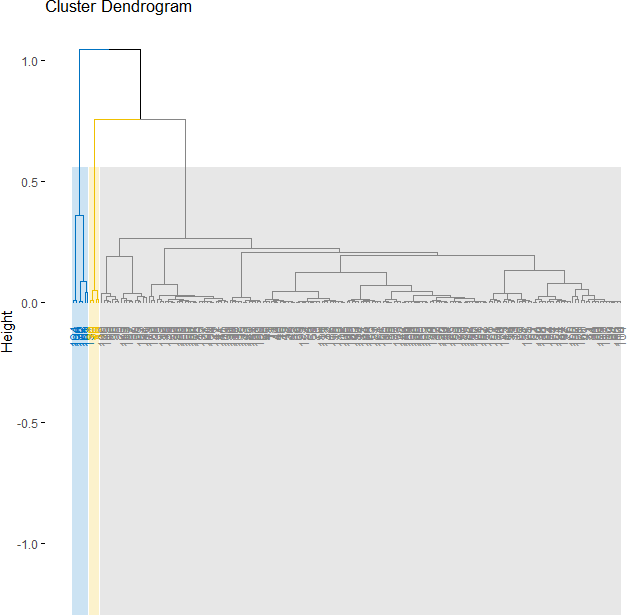
FIGURE 4.24 – Le dendrogramme de la CAH

FIGURE 4.25 – Description des axes retenue par classes

Les figures ci dessus suggère un découpage en 3 groupes. Autrement dit, les indivi- dus sont regroupés en 3 classes selon leur contribution aux axes principaux.

Et on constate de plus que :

* + - * Les individus du premier cluster contribuent tres significativement a l’axe 7.
      * Les individus du deuxieme cluster ne contribuent pas aux axes 8,5,2 et 7.
      * Les individus du troisieme cluster contribuent aux axes 2,5 et 8.

## | Nuage des variables

### | Les coeﬃcients de corrélation des variables avec les projec- tions sur les axes du sous espace

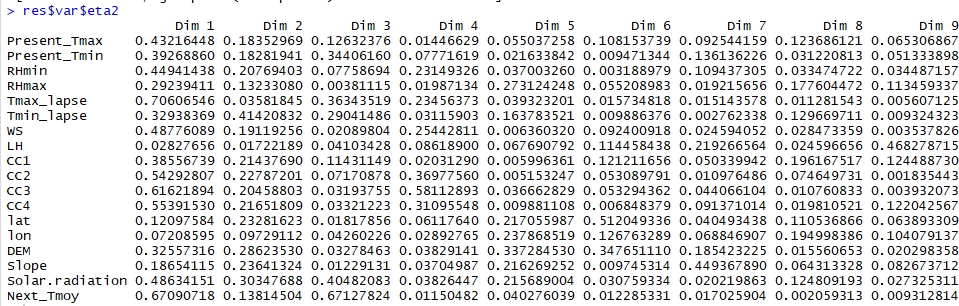


FIGURE 4.26 – Les coefficients de corrélation des variables avec les projections sur les axes du sous espace

### | Le graphique des coeﬃcients de corrélation des variables avec les projections sur le 1er plan factoriel

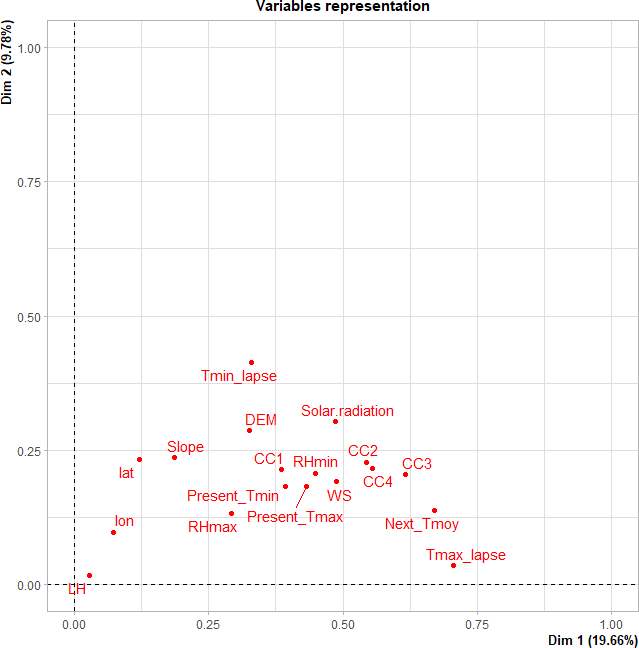


FIGURE 4.27 – Le graphique des coefficients de corrélation des variables avec les projec- tions sur le 1er plan factoriel

## | Conclusion

Dans ce chapitre, on a pu appliquer l’analyse des correspondances multiples qui est une méthode puissante pour synthétiser les jeu de données décrites par plusieurs variables qualitatives.

En effet, grâce à cette méthode on a réduit les 54 modalités en 9 composantes prin- cipales. Chose qui nous a permis de mettre en évidence les relations globales existant entre les modalités et les individus, ainsi leurs visualiser dans un plan factoriel.

# Conclusion générale

La réalisation de ce projet a été une excellente mise en pratique des connaissances en analyse des données acquises tout au long ce module. Et nous a permis de les appro- fondir grâce à sa richesse technique et théorique et aux nombreux notions utilisées.

Durant ce projet, on a pu appliquer la régression linéaire multiple à notre jeu de données, et étudier les différentes relations entre les variables explicatives et la variable dépendante. On a vu plusieurs méthodes pour développer le meilleur modèle de pré- diction.

Ensuite, on a appliqué les différentes méthodes de la classification, à savoir La méthode des centres mobiles K-means et La Classification Ascendante Hiérarchique CAH , qui nous a permis classifier nos variables en des groupes selon des critères bien précis soit le calcul des distances ou le critère de ressemblance.

Et enfin, on a pu entamer les différentes méthodes de l ’analyse factorielle, soit l’analyse en composante principale sur les variables quantitatives, soit l’analyse en correspon- dances multiples sur les variables qualitatives. Ceci nous a permis de réduire la dimen- sionnalité de notre jeu de données et bien visualiser les corrélations entre les variables.