Projet d'étude d'Analyse de données et d'Eléments de modélisation statistique

Bekkare Aziza, Besbes Ines, Mac Yanis, Phung Anh Minh

2024-02-02

Contents

1 Introduction

Dans ce projet, on étudie des données issues du site web Atmo-Occitanie. Le but est d'analyser les émissions de 10 différents polluants atmosphériques des EPCI (Etablissements Publics de Coopération Intercommunale) de la région Occitanie de 2014 à 2019.
A l'aide d'une approche d'analyse de données et de statistique, on cherche à expliquer par différents facteurs les tendances d'émissions de ces polluants au fil des années.

2 Statistiques descriptives

Afin de découvrir les données et de mieux les appréhender, on commence par effectuer des statistiques descriptives sur le jeu mis à notre disposition.

Le jeu de données compte 984 lignes et 36 colonnes, soit 36 variables et 984 individus. On y retrouve les émissions de polluants au format *numeric*.
En s'intéressant aux effectifs, on s'aperçoit que l'on a le même nombre d'EPCI par année, et qu'il y a tous les départements également donc il n'y a pas de données manquantes.

On commence dans un premier temps par convertir nos variables qualitatives en facteur, grâce à la fonction $as.factor$.

On observe que certains EPCI sont à cheval entre plusieurs départements. Cependant, ils sont tous bien situés en Occitanie.

On s'intéresse maintenant aux quatre variables qualitatives : *libEPCI*, *TypeEPCI*, *nomdepart* et *anne_inv*. Ces dernières doivent être considérées comme des facteurs à plusieurs modalités.

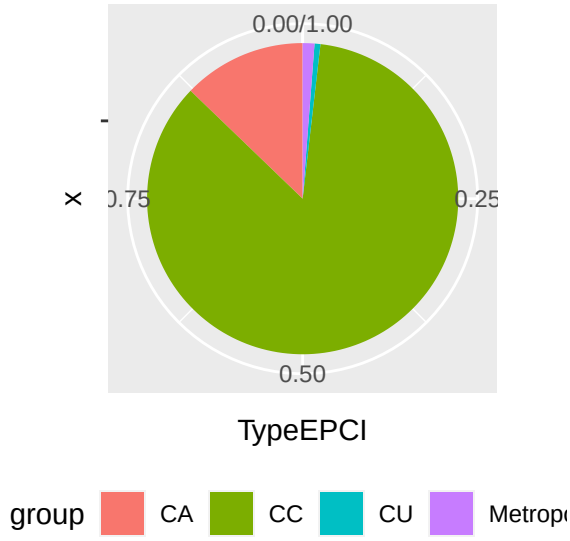


Figure 1: Répartition des types d'EPCI

On observe que les types d'EPCI sont majoritairement CC ou CA, et qu'il y a très peu d'EPCI de type CU ou $Métropole$. On pourra éventuellement chercher à fusionner plus tard ces 2 types d'EPCI avec les types CC ou CA selon les résultats de notre Analyse en Composantes Principales.

2.0.1 Etude des polluants

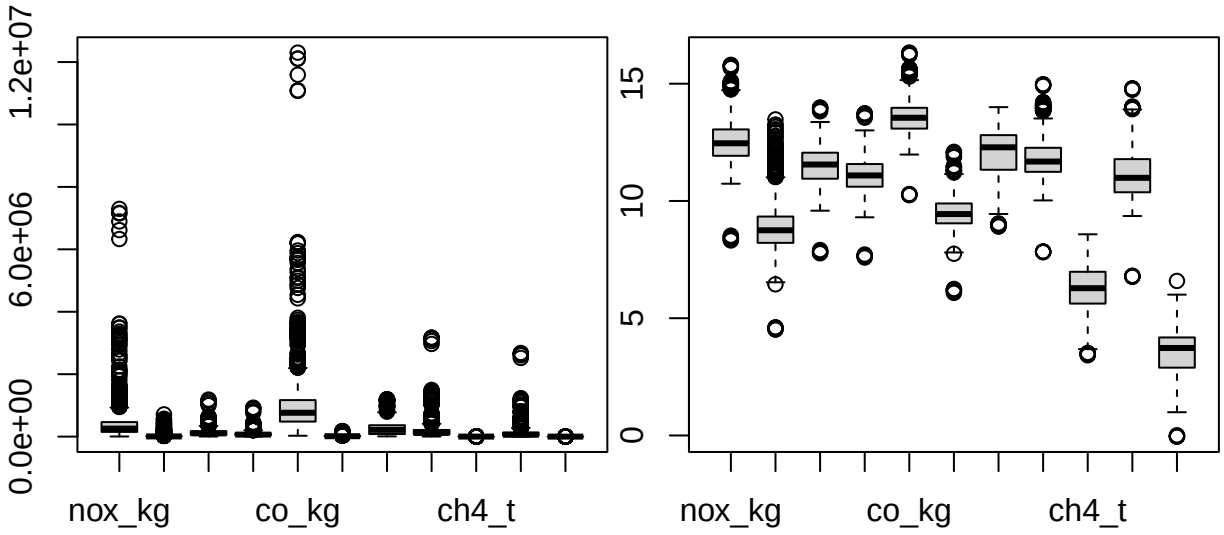


Figure 2: Représentation des données des polluants avant (Gauche) et après (Droite) application de la fonction logarithme

On observe que nos valeurs sont trop dissymétriques, qu'il y a beaucoup d'outliers, et que les valeurs de co_kg ont une plus grande magnitude que les autres et par conséquent écrasent les autres. La variable co_kg risque donc de prendre une performance trop importante par rapport aux autres lors de l'ACP. Il faut donc transformer nos données. Ici, on opte pour une transformation log pour obtenir des distributions symétriques d'ordre de grandeur similaire avec beaucoup moins d'outliers, ce qu'on observe effectivement sur le second boxplot.

Ces résultats sont également visibles lorsqu'on trace les histogrammes des polluants avant et après transformation. On remarque que cette transformation logarithmique des polluants rend la distribution de ces données plus symétrique et plus proche d'une distribution gaussienne.

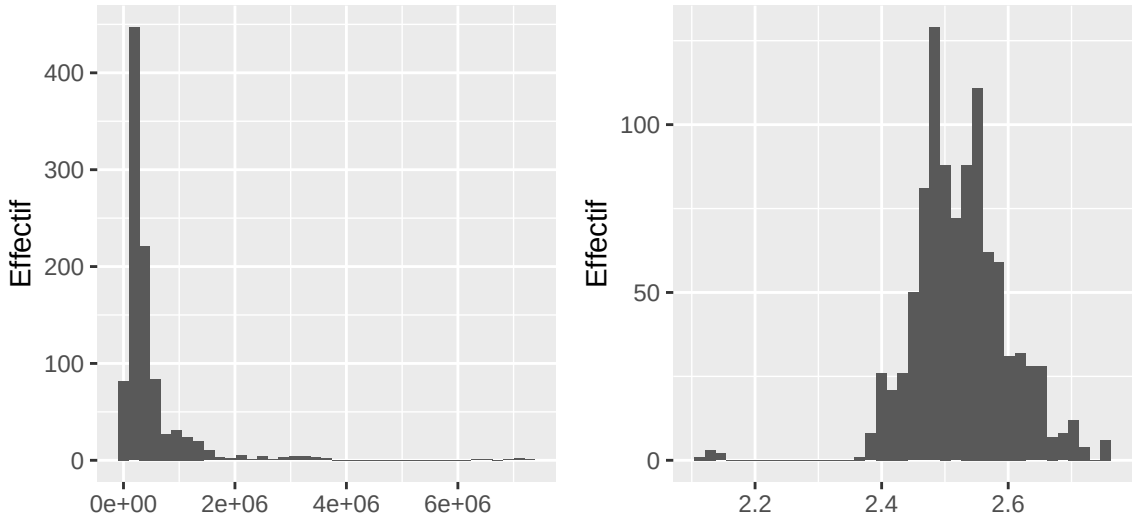


Figure 3: Histogrammes des données des polluants avant (Gauche) et après (Droite) application de la fonction logarithme

3 Analyse en composantes principales

On va à présent réaliser une analyse en composantes principales afin de réduire les dimensions de notre jeu de données. On utilise la fonction **PCA** du package **FactoMineR** pour afficher le graphe des individus et celui des variables.

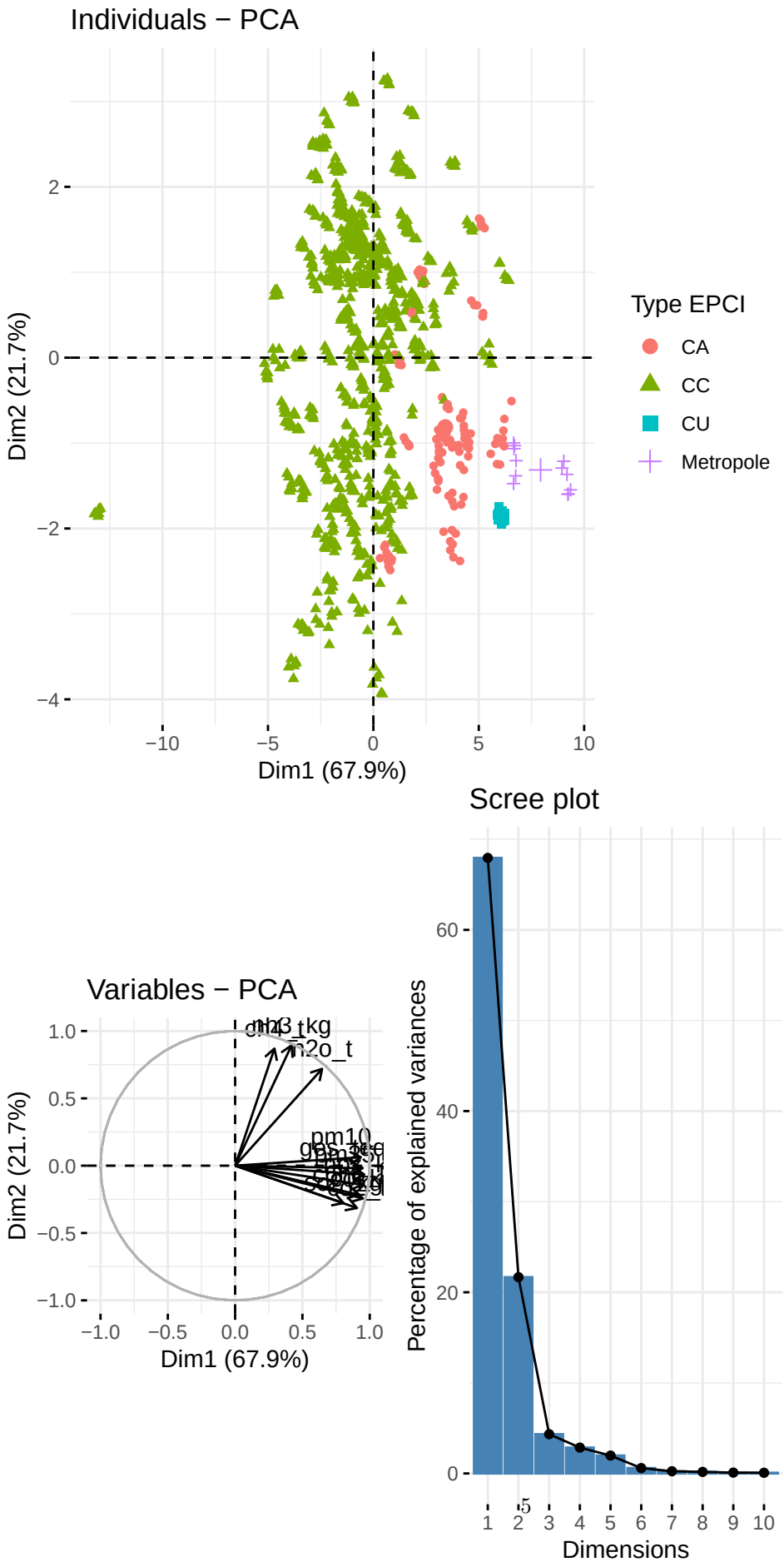


Figure 4: Résultats de l'ACP

Le graphe des variables permet d'interpréter les deux premières composantes principales en les écrivant comme une combinaison linéaire des variables originales. Notons $I = ch4_t, nh3_kg, n2o_t$ et $J = nox_kg, so2_kg, pm10_kg, pm25_kg, co_kg, c6h6_kg, ges_teqco2, co2_t$.
Le cercle des corrélations permet d'écrire : $PC1 = \alpha \cdot J$ et $PC2 = \beta \cdot I$ où α et β sont des poids positifs des variables. $PC1$ est donc proportionnel à la moyenne des variables dans J et $PC2$ à la moyenne des variables dans I. Le graphique des pourcentages d'inertie permet de conserver les deux premières composantes principales car elles représentent 88,95% de l'information contenue dans le jeu de données.
On remarque dans le graphe des individus qu'il y a des valeurs aberrantes. On va donc utiliser la fonction **pca.outlier** de la librairie **mt** qui utilise la distance de Mahalanobis afin de détecter les outliers. Cela suppose d'utiliser des données gaussiennes, ce qui est bien le cas ici.

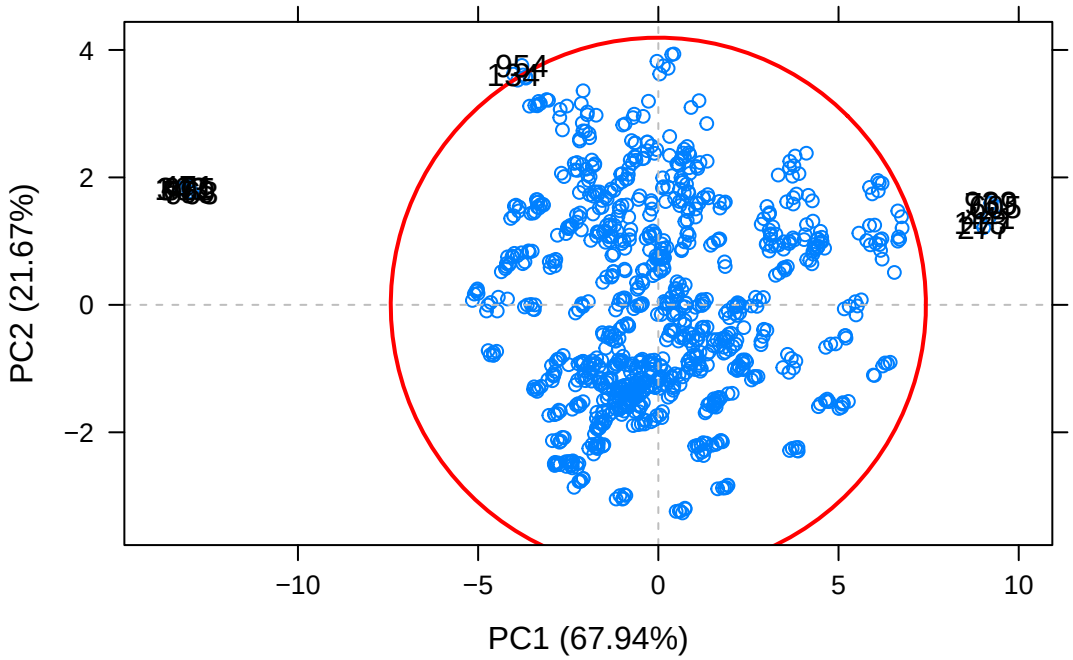


Figure 5: Détection des outliers avec la distance de Mahalanobis

Une fois les outliers détectés, on crée un nouveau jeu de données les excluant.

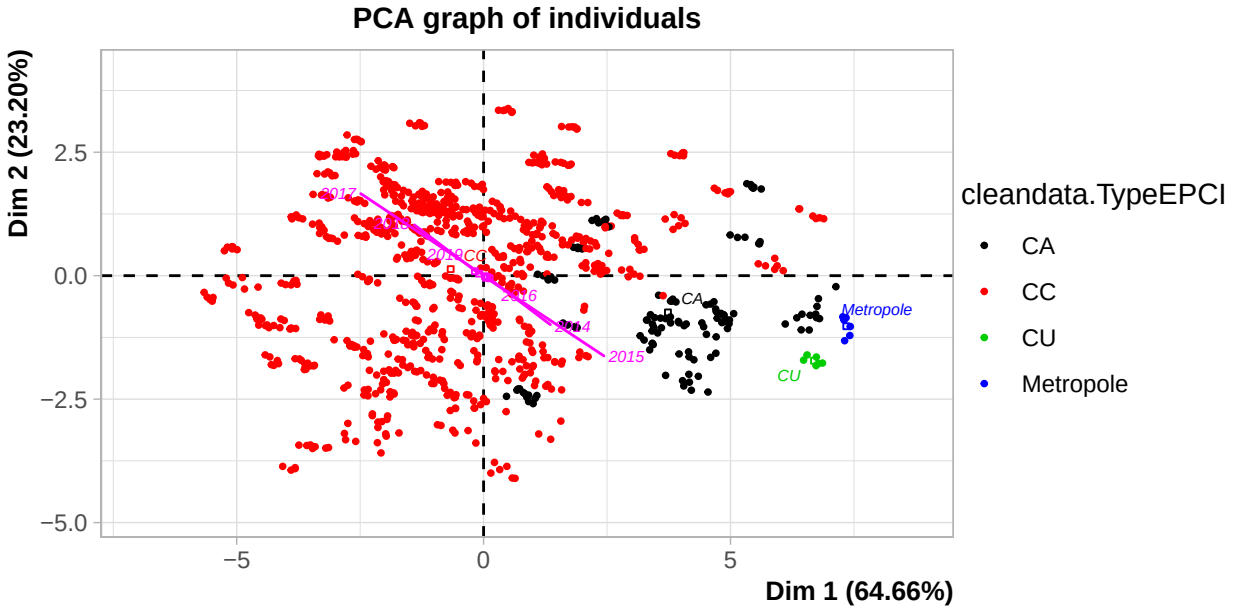


Figure 6: Graphe des individus avec les données nettoyées et transformées

On remarque que l'année n'a pas d'influence sur les deux premières dimensions. Cependant, le type d'EPCI impacte fortement la première dimension. A présent, on décide de regrouper les TypeEPCI CA, CU et $Metropole$ à cause de la différence d'effectifs. On regroupe les types d'EPCI CA, CU et $Metropole$ sous le sigle CA.

4 Analyse Linéaire Discriminante

On effectue une analyse linéaire discriminante afin d'expliquer et prédire l'appartenance d'un individu à une classe à l'aide du reste des données. Il s'agit d'une ACP sur les centroïdes des classes avec une métrique de Mahalanobis. On utilise la LDA pour prédire le dépassement d'émission de méthane de 1000 tonnes par an et prédire le Type d'EPCI.

4.1 Prédiction du dépassement de méthane de 1000 tonnes par an

Dans un premier temps, on explore le dépassement d'émission de méthane de 1000 t par an. On crée la variable *depSeuil* qui vaut 1 si le méthane dépasse 1000 t et 0 sinon. Ensuite, on divise le jeu de donnée en deux afin d'avoir un jeu d'entrainement *x_train* (80 % du jeu initial) et un de test *x_test* (20 % restant). On conserve seulement les variables numériques et la colonne que l'on veut prédire. On utilise la fonction **lda** qui calcule les coefficients de l'ACP. Avec le résultat de la LDA, on utilise la fonction **predict.lda** sur *x_test* et on obtient les prédictions de la variable *depSeuil*.

Pour évaluer les performances de la LDA, on calcule le taux d'erreur de classification. On obtient un taux de 0.0529412. Le taux d'erreur est très faible donc cela implique que presque toutes les données d'entrainement ont étés bien classées.

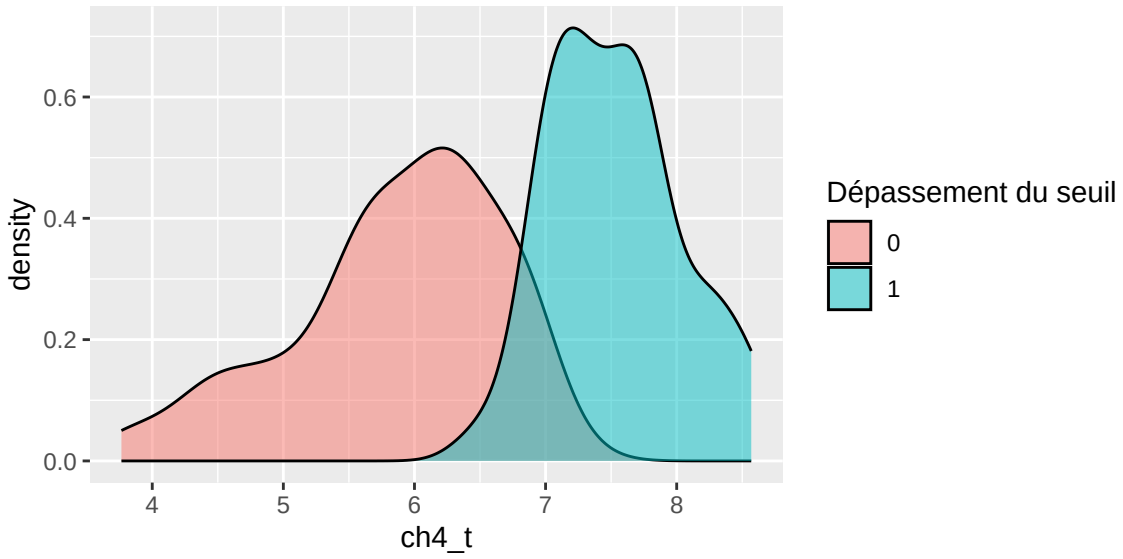


Figure 7: Résultats de LDA pour prédire le dépassement d'émission de méthane de 1000 t par an

Sur la figure 7, on peut observer la densité de probabilité que l'individu dépasse le seuil de 1000 t par an en fonction de la quantité de *ch4_t*. On remarque que plus la quantité de méthane est élevée, plus il y a de chance que le seuil de 1000 tonnes soit dépassé. Cela est cohérent avec ce que l'on aurait prédit.

Dans ce cas, on obtient un taux de 0.0529412. Le taux d'erreur est à nouveau très faible donc la prédiction est satisfaisante.

4.2 Prédiction du type d'EPCI

Afin de représenter les résultats de la LDA pour prédire le type d'EPCI, on crée une nouvelle variable *moypolluant* qui est la moyenne des quantité de polluant pour chaque individu.

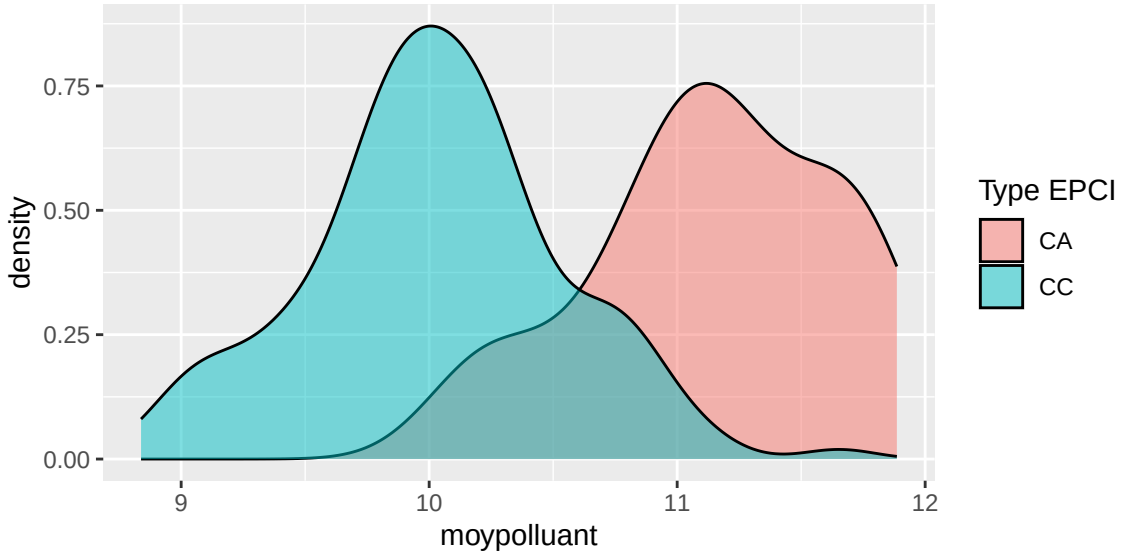


Figure 8: Résultats de LDA pour prédire le type EPCI en fonction de la moyenne des polluants

Sur la figure 8 on observe la densité de probabilité d'appartenir à un type d'EPCI. On remarque que plus la

moyenne des quantités de polluants est élevée plus le type d'EPCI à des chances d'être CA.

Pour la LDA avec 4 facteurs, certaines classes ont moins d'individus donc on entraine moins. Les résultats sont donc meilleurs lorsque l'on regroupe les types d'EPCI.

5 Clustering

Dans cette partie, on cherche une classification de nos données avec plusieurs méthodes différentes, puis on les compare entre elles.

Pour améliorer la classification de nos données, on a décidé de les utiliser après avoir retiré les valeurs aberrantes (outliers), car la présence de ces dernières affecte significativement la performance des algorithmes de classification dans la création de clusters. En éliminant les outliers, on vise à améliorer substantiellement la qualité des clusters générés.

Par ailleurs, on effectue une classification sur les données sur une année uniquement, car après l'étude des statistiques descriptives, on s'est rendus compte qu'il n'y avait pas de grande variation des données d'une année à l'autre. On choisit l'année 2019 arbitrairement, car c'est l'année la plus récente donc à priori la plus significative aujourd'hui, on aurait également pu envisager de moyenner nos données sur chaque année.

5.1 Classification par Kmeans

Dans un premier temps, on propose de rechercher une classification des données par la méthode des K-means. On utilise les critères d'inertie intra-classe et le critère silhouette pour déterminer le nombre de classes à retenir.

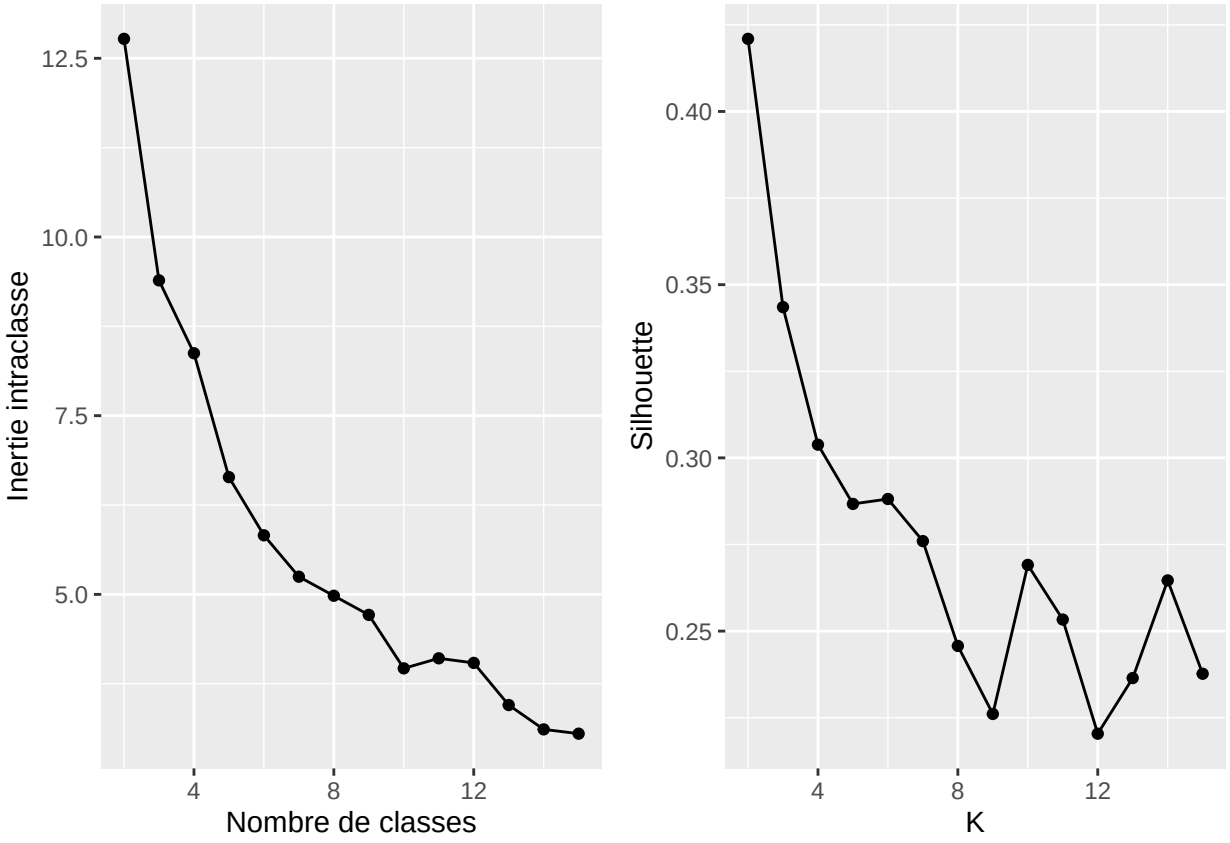


Figure 9: Critères Silhouette et intra-classe pour la méthode K-means

Sur la figure 9, on observe un coude pour 5 classes. On retient une classification à 5 classes avec le critère d'inertie intra-classe.

D'autre part, on observe un premier changement de pente à 2 classes. On retient donc une classification à 2 classes avec le critère silhouette.

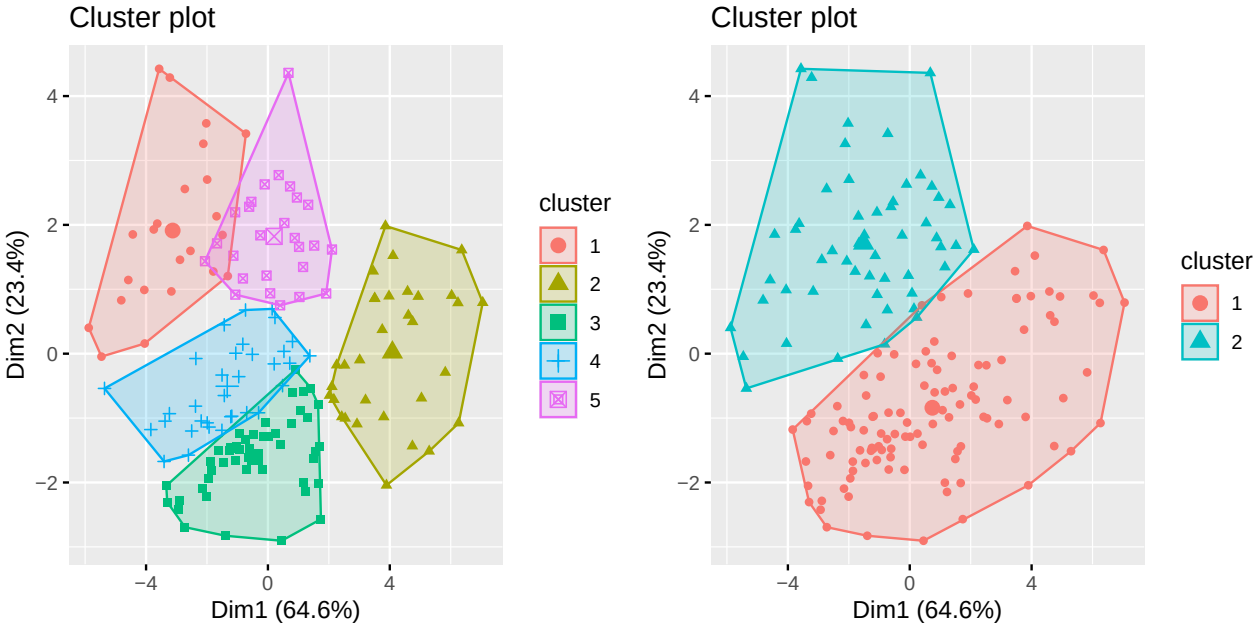


Figure 10: Visualisation des classes obtenues par silhouette sur les plans de l'acp

```
##   cluster size ave.sil.width
## 1       1  108          0.43
## 2       2   53          0.39
```

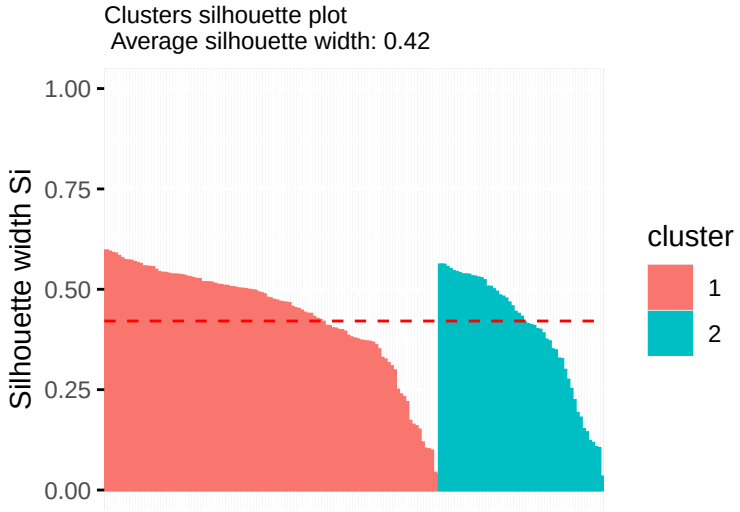


Figure 11: Tracé des Si

On obtient une classification peu fiable avec le critère silhouette, le Si moyen est faible et la forme du Si de la classe 2 est loin d'être rectangulaire, la classification à 2 classes par kmeans ne donne donc pas de résultat très satisfaisant.

```
## [1] 0.2745385
```

```
##
##      1  2  3  4  5
##   1  0 32 47 26  3
##   2 23  0  0  7 23
```

Les individus de la classe 1 de la classification par K-means à 2 classes se retrouvent répartis dans les classes 1, 2 et 4 de la classification par K-means avec 5 classes. Les individus de la classe 2 se retrouvent dans les classes restantes.

```
## [1] 0.07505219
```

```
## [1] -0.0544875
```

```
##
##     CA CC
##   1  0 23
##   2 19 13
##   3  0 47
##   4  1 32
##   5  3 23
```

Il ne semble pas y avoir de lien dans la classification par K-means avec 5 et 2 classes entre le type d'EPCI et les individus dans chaque classe.

On verra lors de nos recherches de classification par modèle de mélange qu'en réalité les modèles diagonaux et sphériques ne sont pas forcément très adaptés à nos données. La méthode K-means reposant sur la notion de distance euclidienne et ne pouvant pas faire de séparation non convexe, n'est peut-être pas adaptée à nos données.

5.2 Classification hiérarchique

Dans cette partie, on propose de rechercher une classification de nos données par une méthode de classification hiérarchique. On utilise les critères de Calinski-Harabasz et le critère silhouette pour déterminer le nombre de classes à retenir. On compare également des méthodes différentes, notamment celles utilisant le lien moyen et la méthode de Ward.

5.2.1 Méthode du lien moyen

On commence par étudier la méthode utilisant le lien moyen.

$$D(C_k, C_{k'}) = \frac{1}{|C_k||C_{k'}|} \sum_{i \in C_k} \sum_{l \in C_{k'}} d(x_i, x_l)$$

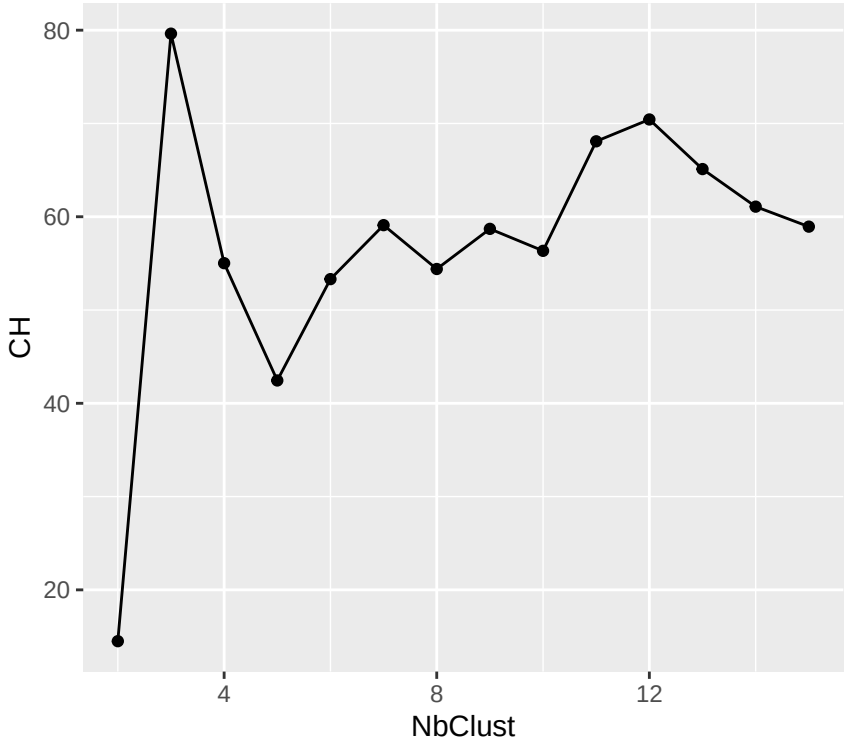


Figure 12: Critère CH pour le lien moyen

En maximisant le critère de Calinski-Habarasz sur une classification hiérarchique utilisant une méthode du lien moyen, on obtient une classification avec 3 classes.

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

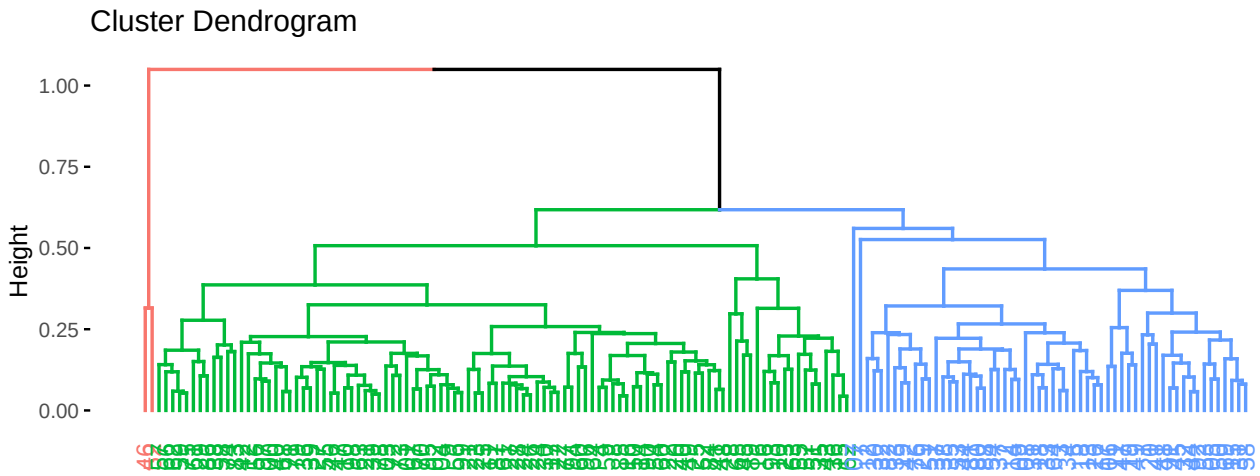


Figure 13: Dendogramme avec le lien moyen

```
## resCAH3
##   1   2   3
##  58 101   2
```

5.2.2 Méthode de Ward

On étudie maintenant les classifications obtenues par la méthode de Ward.

$$D(C_k, C_{k'}) = \frac{|C_k||C_{k'}|}{|C_k| + |C_{k'}|} d(m_k, m_{k'})^2$$

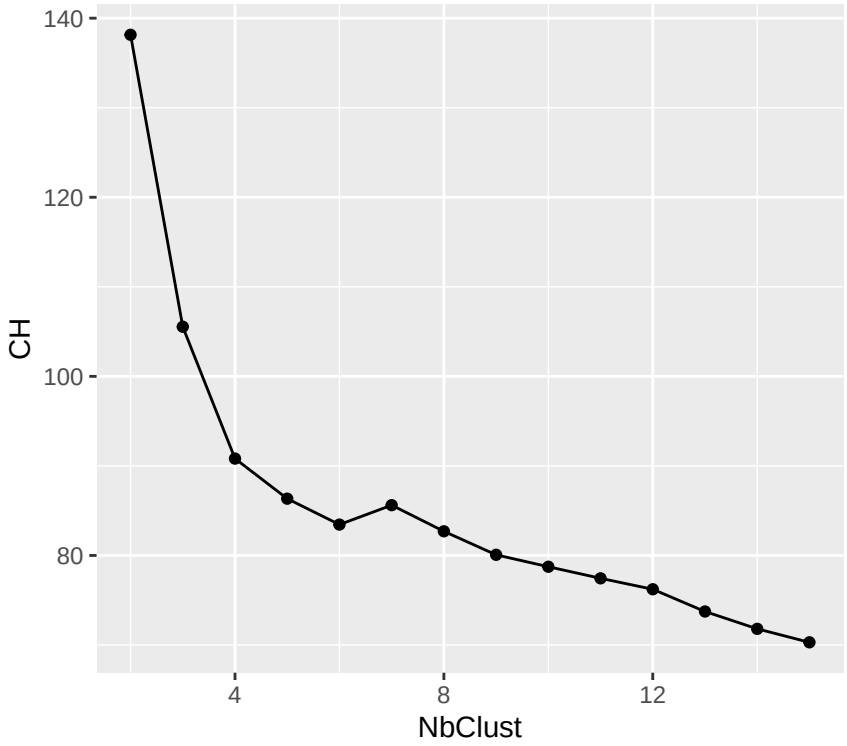


Figure 14: Critère CH pour méthode de Ward

En maximisant le critère de Calinski-Habarasz sur une classification hiérarchique utilisant une méthode de Ward, on obtient une classification avec 2 classes.

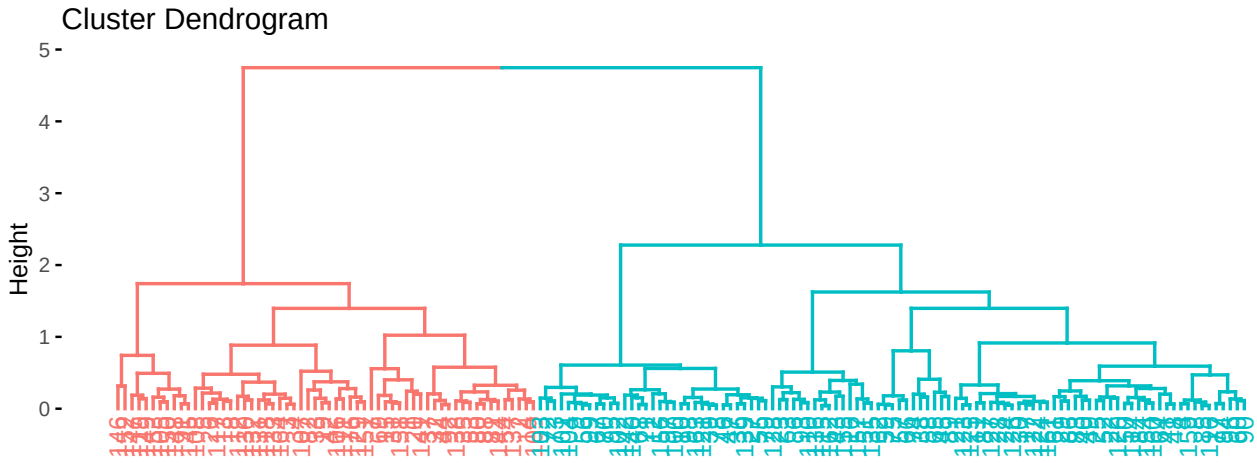


Figure 15: Dendogramme avec le lien Ward

```
##        resCAH3
## resCAH2   1   2   3
##       1  58   0   2
##       2   0 101   0
```

```
## [1] 0.9819439
```

La classification resCAH3 est quasiment la même que la classification resCAH2 car on a seulement ajouté 2 éléments dans la classe 3 de resCAH3.

```
##
## resCAH2 CA CC
##       1  3 57
##       2 20 81
```

```
## [1] -0.03532963
```

```
##
## resCAH3 CA CC
##       1  3 55
##       2 20 81
##       3  0  2
```

```
## [1] -0.04216316
```

On aurait pu imaginer que nos classes feraient la différence entre nos types d'EPCI en séparant les EPCI du type CA et CC mais ce n'est en fait pas le cas.

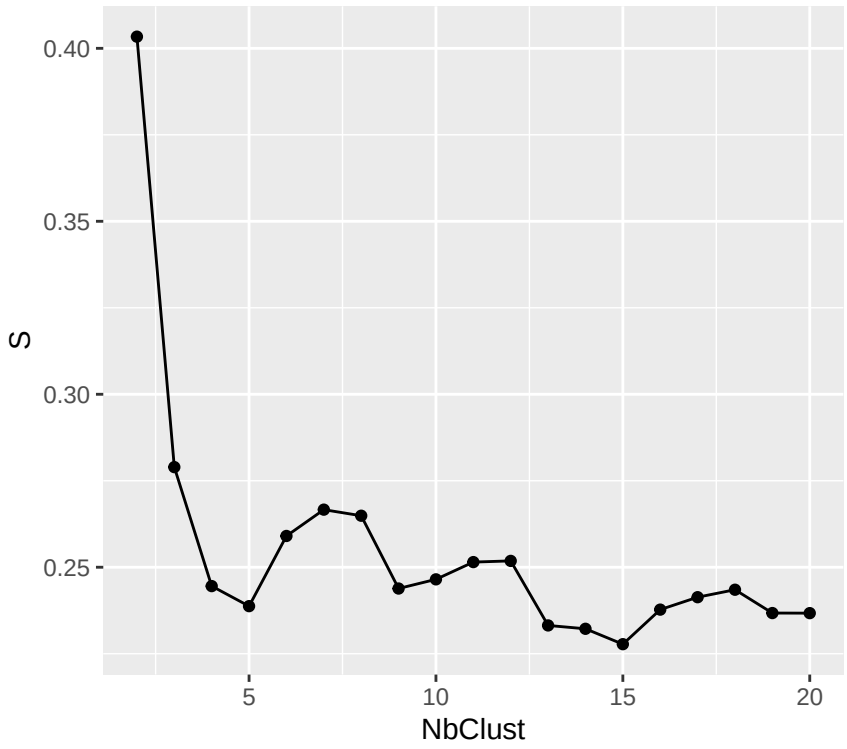


Figure 16: Critère Silhouette pour la méthode de Ward

En cherchant une nouvelle classification avec le critère silhouette, on obtient la même que celle obtenue avec le critère de Calinski-Habarasz.

```
## [1] 0.3063318
```

```
## [1] 0.3040657
```

```
## [1] 0.8315033
```

```
## [1] 0.8181151
```

```
##    resCAH3
##       1   2   3
##   1   7 101   0
##   2  51   0   2
```

La classification obtenue par K-means avec 2 classes semble se rapprocher de la classification hiérarchique avec 2 et 3 classes car on a un adjusted Rand Index d'environ 0.81 qui est très proche de 1. Il semblerait que l'une soit une simplification de l'autre. Les autres classifications ne se rapprochent pas l'une de l'autre autrement.

5.3 Classification par modèle de mélange

Pour finir, on étudie les classifications obtenues par modèle de mélange. Pour déterminer le nombre de classes à retenir, on utilise les critères BIC et ICL.

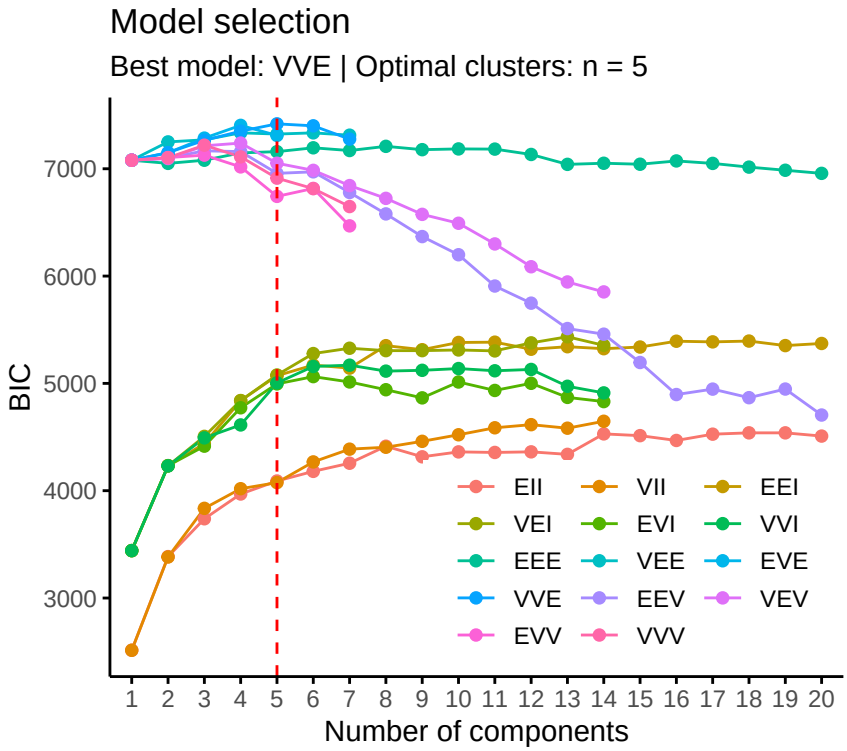


Figure 17: Critère BIC pour tout modèle de mélange

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVE (ellipsoidal, equal orientation) model with 5 components:
##
##  log-likelihood   n  df      BIC      ICL
##        4138.573 161 169 7418.389 7411.288
##
## Clustering table:
##  1  2  3  4  5
## 35 35 49 18 24
```

Tous les modèles de mélange diagonaux et sphériques sont complètement incompatibles avec nos données au vu du tracé de notre critère BIC. On conserve le modèle de mélange VVE à 5 classes.

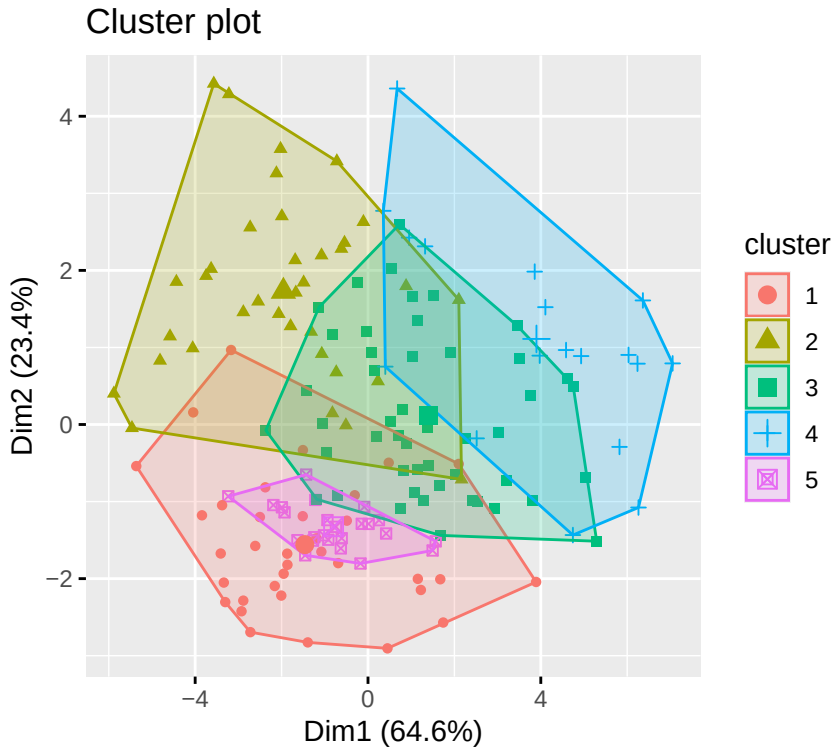


Figure 18: Visualisation des classes obtenues dans le premier plan de l'ACP

On obtient en effet des classes qui ne sont plus diagonales ou sphériques.

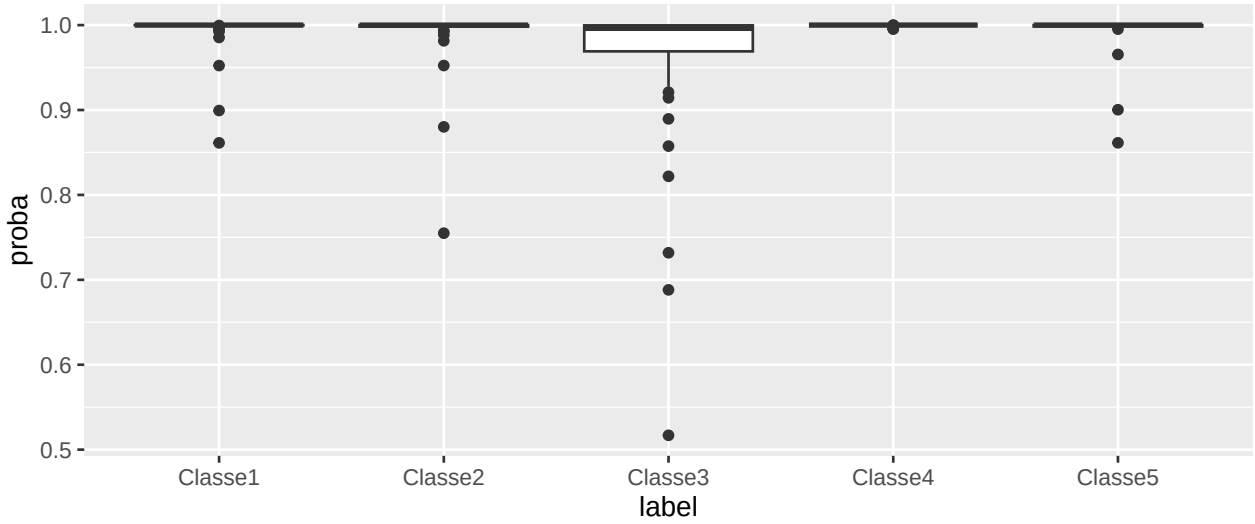


Figure 19: Boxplots des probabilités d'appartenance maximale

On observe que les boxplots sont très serrés et très proches de 1. Il semblerait donc que la plupart des individus aient été attribués à la classe qui leur correspond. Il y a quelques outliers qui sont du au chevauchement entre les classes, mais globalement chaque individu semble avoir été bien classé.

```
## Best ICL values:
##             VVE,5      EVE,4      VVE,6
```

```
## ICL      7411.288 7395.93510 7392.99601
## ICL diff    0.000  -15.35318  -18.29227
```

Le critère ICL et le critère BIC proposent 5 classes et le même type de modèle VVE que pour le critère BIC, les classifications retenues sont donc les mêmes.

```
## [1] 0.2136428
```

```
## [1] 0.1545541
```

```
## [1] 0.1646979
```

```
## [1] 0.1624851
```

Comme prévu, les classifications que l'on a obtenu par les méthodes de K-means et classification hiérarchique ne se rapprochent pas de la classification que l'on a obtenu par modèle de mélange. Le modèle de mélange final obtenu n'étant pas sphérique, on avait peu de chances d'obtenir une classification similaire.

```
## [1] 0.03155391
```

```
##
##     CA CC
##   1  1 34
##   2  0 35
##   3 12 37
##   4 10  8
##   5  0 24
```

Pour toutes les classifications que l'on a obtenu, on obtient à priori pas de lien entre nos classes et le type d'EPCI. On peut donc imaginer que le type d'EPCI n'a pas d'incidence significative sur la concentration en polluant, ce qui tout de même pourrait être contre-intuitif.

5.4 Visualiation sur une carte

On décide à présent de visualiser les émissions de polluants dans les différents types d'EPCI. Pour ce faire, on utilise la librairie leaflet qui permet de créer une carte interactive. Dans un premier temps, on décide de visualiser la quantité des polluants dans les différents EPCI. On observe (géographie demander).

6 Modèle Linéaire

6.1 Le gaz à effet de serre en fonction des variables Type et années

On souhaite expliquer l'émission du gaz à effet de serre *ges_teqco* en fonction des variables *typeEPCI* et *annee_inv*. On commence par le modéliser par un modèle d'ANOVA à 2 facteurs avec interactions.

$$\begin{cases} ges\_teqco2_{ij\ell} = \mu + \alpha_i + \beta_0 + \beta_1 \cdot TypeEPCI_i + \beta_2 \cdot annee\_inv_j + \beta_{12} \cdot TypeEPCI_i \times annee\_inv_j + \epsilon_{ij\ell} \\ \forall i = 1, \dots, I, j = 1, \dots, J, \ell = 1, \dots, n_{ij} \\ (\varepsilon_{ij\ell}) \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

Afin de tester la présence d'interactions, on écrit le modèle additif:

$$\begin{cases} ges\_teqco2_{ij\ell} = \mu + \alpha_i + \beta_0 + \beta_1 \cdot TypeEPCI_i + \beta_2 \cdot annee\_inv_j + \epsilon_{ij\ell} \\ \forall i = 1, \dots, I, j = 1, \dots, J, \ell = 1, \dots, n_{ij} \\ (\varepsilon_{ij\ell}) \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

On utilise la fonction **anova** sur les deux modèles évoqués précédemment.

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ TypeEPCI + annee_inv
## Model 2: ges_teqco2 ~ TypeEPCI * annee_inv
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    963 373.76
## 2    958 373.76  5 0.0039962 0.002      1
```

La pvaleur est de » 0.05 donc on ne rejette pas le sous modèle au risque 5%. Il n'y a donc pas d'interactions entre les variables *typeEPCI* et *anne_inv*.

$$\begin{cases} ges\_teqco2_{ij} = \beta_0 + \beta_1 \cdot TypeEPCI_i + \epsilon_{ij\ell} \\ \forall i = 1, \ldots, I, j = 1, \ldots, n_i \\ (\varepsilon_{ij}) \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ TypeEPCI
## Model 2: ges_teqco2 ~ TypeEPCI + annee_inv
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    968 373.95
## 2    963 373.76  5   0.18696 0.0963 0.9928
```

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ annee_inv
## Model 2: ges_teqco2 ~ TypeEPCI + annee_inv
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    964 615.84
## 2    963 373.76  1    242.08 623.72 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On effectue en parallèle deux tests de Fisher de sous modèle pour tester la nullité des variables année et TypeEPCI. On observe que la pvaleur pour le premier test est `anova2$pvalue` » 0.5 donc on ne rejette pas H0 au risque 5%. L'année n'a pas d'effet significarif sur le modèle. Le second test a une pvaleur < 2.2e-16 donc « 0.05. On rejette H0 au risque 5% donc le type d'EPCI a un effet significatif sur le modèle.

Regardons à présent si on peut également enlever le Type EPCI.

$$\begin{cases} ges\_teqco2_{ij} = \mu + \epsilon_{ij} \\ \forall i = 1, \ldots, I, j = 1, \ldots, n_i \\ (\varepsilon_{ij}) \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ 1
## Model 2: ges_teqco2 ~ TypeEPCI + annee_inv
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    969 616.02
## 2    963 373.76  6    242.26 104.03 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cette fois ci la pvaleur est encore inférieure à 0.05 donc on rejette cette hypothèse de nullité de la variable TypeEPCI au risque 5%.

Vérifions que le modèle avec seulement le Type d'EPCI est bien validé par rapport au modèle avec interactions.

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ TypeEPCI
## Model 2: ges_teqco2 ~ TypeEPCI * annee_inv
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    968 373.95
## 2    958 373.76 10   0.19096 0.0489      1
```

Le modèle est bien validé par rapport au modèle complet car on a une pvaleur égale à .

Pour confirmer nos résultats, on utilise un algorithme descendant de sélection de variable basé sur le critère BIC.

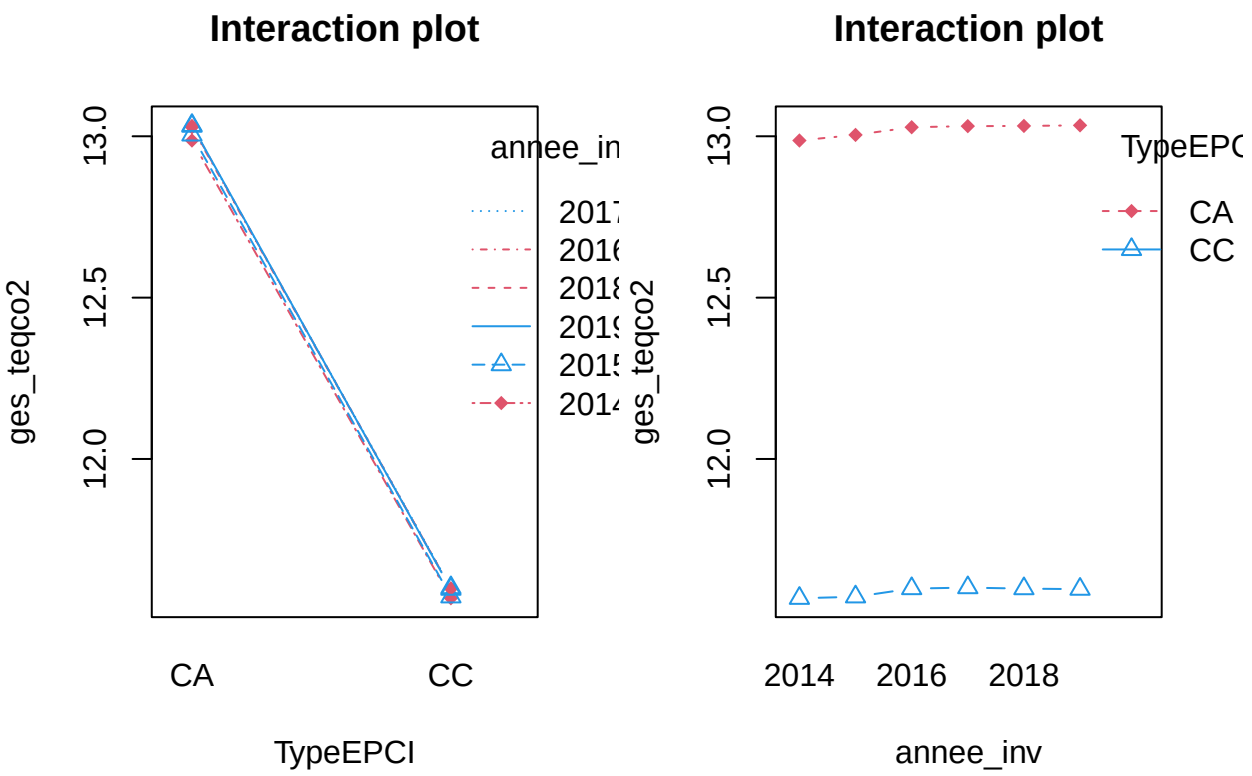
```
## Start:  AIC=-842.55
## ges_teqco2 ~ TypeEPCI * annee_inv
##
##                       Df Sum of Sq    RSS     AIC
## - TypeEPCI:annee_inv  5 0.0039962 373.76 -876.93
## <none>                              373.76 -842.55
##
## Step:  AIC=-876.93
## ges_teqco2 ~ TypeEPCI + annee_inv
##
##             Df Sum of Sq    RSS     AIC
## - annee_inv  5     0.187 373.95 -910.83
## <none>                   373.76 -876.93
## - TypeEPCI   1   242.079 615.84 -399.42
##
## Step:  AIC=-910.83
## ges_teqco2 ~ TypeEPCI
##
##            Df Sum of Sq    RSS     AIC
## <none>                  373.95 -910.83
## - TypeEPCI  1    242.07 616.02 -433.52
##
## Call:
## lm(formula = ges_teqco2 ~ TypeEPCI, data = newcleandata)
##
## Coefficients:
## (Intercept)   TypeEPCICC
##       13.02        -1.43
```

La sélection de variables confirme bien le modèle retenu:

$$\begin{cases} \text{ges\_teqco2}_{ij} = \beta_0 + \beta_1 \cdot TypeEPCI_i + \epsilon_{ij} \\ \forall i = 1, \ldots, I, j = 1, \ldots, n_i \\ (\varepsilon_{ij}) \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

Où ges_teqco_{ij} est la quantité de gaz à effet de serre du TypeEPCI i du jème individu.

On visualise le graphe d'interactions :



On observe que les courbes sont parallèles, il n'y a donc pas d'interaction entre les différents types d'EPCI.

6.2 Gaz à effet de serre en fonction de tous les autres polluants

Dans cette partie, on s'intéresse à la relation entre la quantité de gaz à effet de serre en fonction de tous les polluants. On commence par la modéliser par un modèle de régression linéaire avec interactions entre les facteurs car nous étudions l'influence de variables quantitatives uniquement.

```
## Start:  AIC=-5419.47
## ges_teqco2 ~ (nox_kg + so2_kg + pm10_kg + pm25_kg + co_kg + c6h6_kg +
##     nh3_kg + ch4_t + co2_t + n2o_t)^2
```

Avec une méthode forward on obtient le modèle complet avec interactions. Avec des méthodes backward, on obtient des sous-modèles du modèle avec interaction, le critère BIC choisissant un modèle étant sous-modèle de celui retenu par le critère AIC.

En effectuant un test de sous-modèle entre le modèle complet et le sous-modèle retenu par le critère BIC, on trouve une pvaleur de 0.1488. On ne rejette donc pas H0 au niveau 5%, et on peut travailler avec le sous-modèle retenu par le critère BIC.

6.3 Emission de méthane en fonction de l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année

Dans cette partie, on s'intéresse à la relation entre l'émission de méthane en fonction de l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année. Pour cela, on va mettre en place un modèle d'analyse de la covariance car on veut expliquer une variable quantitave en fonction de variables qualitatives et quantitatives.

Nous allons considérer les interactions et ainsi travailler avec le modèle singulier complet.

Afin de simplifier le modèle, on applique un algorithme de sélection descendante par le test de Fisher en utilisant les critères AIC et BIC.

```
stepAIC(acpolluant,trace=F,direction="backward")
```

```
##
## Call:
## lm(formula = ch4_t ~ annee_inv + nh3_kg + n2o_t + TypeEPCI +
##     nh3_kg:n2o_t + nh3_kg:TypeEPCI + n2o_t:TypeEPCI, data = newcleandata[c(3,
##     10, 12, 14, 15)])
##
## Coefficients:
##       (Intercept)      annee_inv2015      annee_inv2016      annee_inv2017
##           2.20608           -0.04537           -0.09945           -0.19658
##     annee_inv2018      annee_inv2019             nh3_kg              n2o_t
##          -0.31246           -0.29096            0.44271           -1.85056
##        TypeEPCICC       nh3_kg:n2o_t  nh3_kg:TypeEPCICC   n2o_t:TypeEPCICC
##         -13.32544            0.12827            1.47963           -1.35428
```

```
stepAIC(acpolluant,trace=F,direction="backward",k=log(nrow(newcleandata)))
```

```
##
## Call:
## lm(formula = ch4_t ~ annee_inv + nh3_kg + n2o_t + TypeEPCI +
##     nh3_kg:n2o_t + nh3_kg:TypeEPCI + n2o_t:TypeEPCI, data = newcleandata[c(3,
##     10, 12, 14, 15)])
##
## Coefficients:
##       (Intercept)      annee_inv2015      annee_inv2016      annee_inv2017
##           2.20608           -0.04537           -0.09945           -0.19658
##     annee_inv2018      annee_inv2019             nh3_kg              n2o_t
##          -0.31246           -0.29096            0.44271           -1.85056
##        TypeEPCICC       nh3_kg:n2o_t  nh3_kg:TypeEPCICC   n2o_t:TypeEPCICC
##         -13.32544            0.12827            1.47963           -1.35428
```

Au final, on obtient le même modèle simplifié suivant :

$$\begin{cases} ch4\_t_{ijl} = \mu + \sum_{i=2015}^{2019} \beta_{inv_i} \times 1_{(annee\_inv=i)} + \beta_{nh3} \times nh3\_kg + \beta_{n2o} \times n2o\_t + \sum_{j=1}^{4} \beta_{EPCI_j} \times 1_{(TypeEPCI=j)} \\ + \sum_{j=1}^{4} \beta_{nh3\_EPCI_j} \times nh3\_kg \times 1_{(TypeEPCI=j)} + \beta_{nh3\_n2o} \times nh3\_kg \times n2o\_t \\ + \sum_{j=1}^{4} \beta_{n2o\_EPCI_j} \times n2o\_t \times 1_{(TypeEPCI=j)} + \varepsilon_{ijl}, \\ \forall i = 2015, \ldots, 2019, j = 1, \ldots, 4, l = 1, \ldots, n_{ij}. \\ (\varepsilon_{ijl}) \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

6.4 Dépassement d'émission de méthane de 1000 t par an en fonction de l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année.

On veut expliquer le dépassement d'émissions de méthane de 1000 t par an en fonction de l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année. Pour ce faire, on modifie la variable ch4_t, cette dernière prend la valeur 1 si elle est supérieure à 1000 tonnes, 0 sinon. On construit donc un modèle de régression logistique car la variable réponse est binaire.

```
##   annee_inv   nh3_kg ch4_t    n2o_t TypeEPCI
## 1      2019 11.80325 FALSE 2.839897       CC
## 2      2019 11.64862 FALSE 2.983407       CC
## 3      2019 11.17293 FALSE 2.868240       CC
## 4      2019 12.40602 FALSE 3.931355       CC
## 5      2019 11.77857 FALSE 3.204371       CA
```

```
## 6      2019 11.61738 FALSE 4.090337       CA
```

On s'intéresse tout d'abord au modèle complet avec interactions.

Afin de simplifier le modèle, on utilise les critères BIC et AIC.

On commence par faire un test de sous modèle entre le modèle avec interactions et le modèle obtenu avec le critère AIC. On obtient une pvaleur de 0.5528 donc on ne rejette pas H0 au niveau 5% et on peut travailler avec le sous modèle **bestmodel**.

On fait ensuite la même chose, avec les modèles obtenus avec les critères BIC et AIC. On obtient une pvaleur de ... donc on rejette H0 au risque 5%.

Le sous-modèle retenu à la fin est donc **bestmodel**.

7 Régression régularisée

On va à présent utiliser une méthodes de régression régularisée sur les polluants. On les a déjà centrés et réduits. On veut expliquer le gaz à effet de serre en fonction de tous les autres polluants. On utilise donc la régression de Lasso car c'est celle qui fait la sélection de variables et qui rend le modèle plus interprétable.

7.1 Régression Lasso

La régression Lasso est une méthode de régression linéaire qui inclut une pénalité de régularisation L1 sur les coefficients de régression. Cette pénalité a pour effet de réduire certains coefficients à exactement zéro, ce qui peut être utile pour la sélection de variables et pour produire des modèles plus simples et plus interprétables.

Dans un premier temps, on ajuste une régression Lasso en faisant varier λ sur une grille. On stockera le résultat dans la variable.

On trace ensuite les chemins de régularisation de la régression Lasso qui montrent comment les coefficients de régression évoluent en fonction du paramètre de régularisation λ. Pour ce faire, on utilise le paramètre λ qui contrôle l'intensité de la pénalité L1. En augmentant λ, on augmente la quantité de régularisation appliquée aux coefficients, ce qui peut conduire à plus de coefficients étant réduits à zéro.

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```

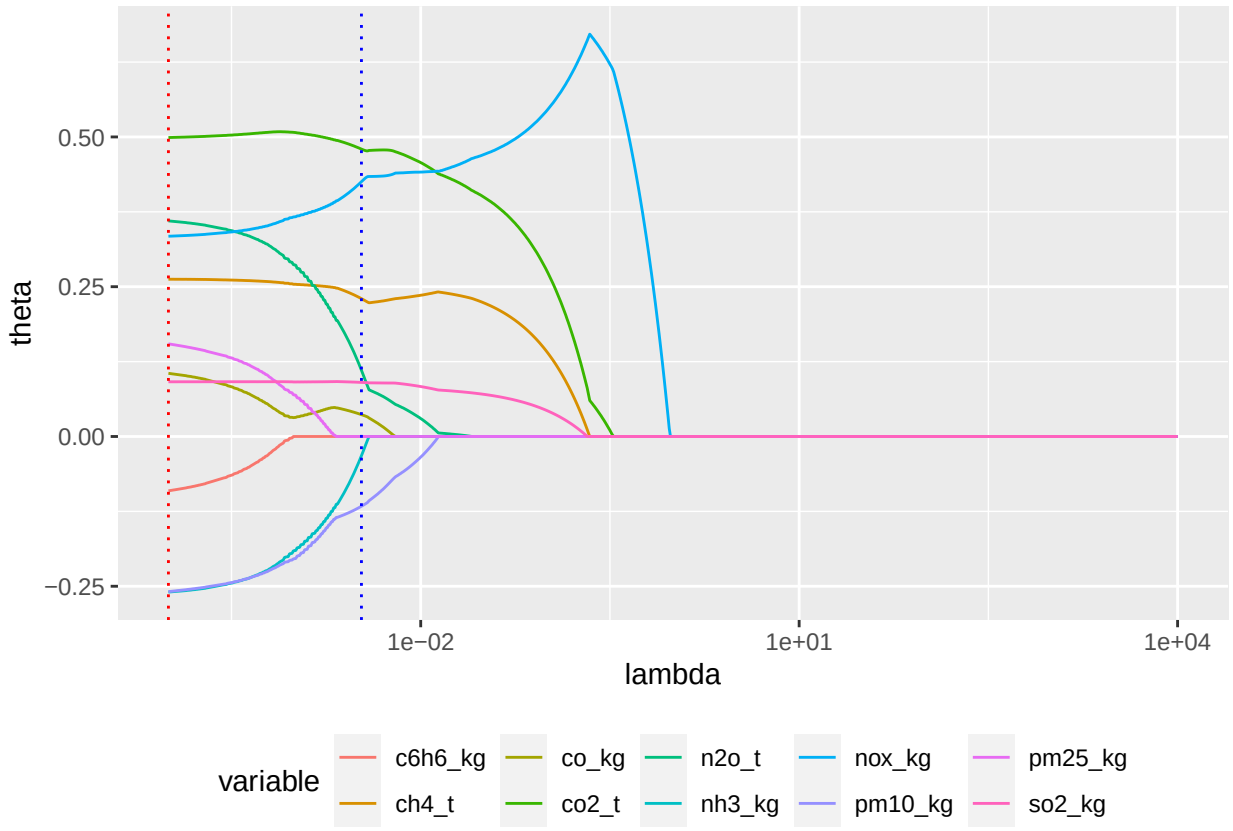


Figure 20: Chemins de régularisation avec la régression Lasso

```
## [1] 1e-04
```

```
## [1] 0.003388442
```

La valeur de λ sélectionnée est : `lambda1se`.

7.2 Sous modèle

```
##              Value SE Coefficient
## nox_kg    0.33427063 NA      nox_kg
## so2_kg    0.09128815 NA      so2_kg
## pm10_kg -0.25891293 NA     pm10_kg
## pm25_kg  0.15461114 NA     pm25_kg
## co_kg     0.10542456 NA       co_kg
## c6h6_kg -0.09071312 NA     c6h6_kg
## nh3_kg  -0.25969692 NA      nh3_kg
## ch4_t     0.26249571 NA       ch4_t
## co2_t     0.49899930 NA       co2_t
## n2o_t     0.35988296 NA       n2o_t
```

```
##              Value SE Coefficient
## nox_kg    0.42531192 NA      nox_kg
## so2_kg    0.09027087 NA      so2_kg
## pm10_kg -0.11646082 NA     pm10_kg
## co_kg     0.03646834 NA       co_kg
## nh3_kg  -0.03207278 NA      nh3_kg
## ch4_t     0.23008360 NA       ch4_t
```

```
## co2_t    0.47970216 NA       co2_t
## n2o_t    0.11055273 NA       n2o_t
```

Etant donné qu'on veut faire de la sélection de variables, on utilise lambda.1se car il annule plus de variables que le lambda.min.

8 Conclusion

Cette analyse des émissions de polluants en Occitanie, fondée sur les types d'EPCI sur cinq ans, pourrait ouvrir la voie à une étude plus vaste et détaillée à l'échelle nationale. En élargissant le champ d'investigation à toute la France et en prolongeant la période d'étude, tout en intégrant des variables supplémentaires telles que la densité de population, on pourrait obtenir une compréhension plus nuancée et complète des tendances de pollution. Cette approche élargie permettrait d'affiner les stratégies de réduction des émissions polluantes. Enfin, ce projet nous a surtout permis de mettre en application nos compétences techniques et de travailler sur un vrai jeu de données.