Projet d'étude de Statistiques

Maxime Baba, Alexandre Demarquet, Félix de Brandois, Tristan Gay

2024-02-24

Contents

List of Figures

Introduction

Ce rapport analyse les émissions de polluants atmosphériques dans la région Occitanie sur une période de 2014 à 2019, à partir des données fournies par le site web Atmo-Occitanie.

On utilise le jeu de données suivant : `Data-projetmodIA-2324.csv`. Ces données comprennent les émissions de divers polluants tels que les oxydes d'azote, le dioxyde de soufre, les particules en suspension, le monoxyde de carbone, le benzène, l'ammoniac, les gaz à effet de serre, le méthane, le dioxyde de carbone et le protoxyde d'azote, ainsi que des informations sur les EPCI (Etablissements Publics de Coopération Intercommunale) telles que leur nom, leur code d'identification, leur département d'appartenance, leur latitude, leur longitude et leur type.

Nous allons pour cela utiliser plusieurs méthodes d'analyse de données, et également de modèle linéaire, afin de pouvoir anlyser ce jeu de données.

1 Analyse descriptive des données

On commence par interpréter les éléments du jeu de données.
Il est composé de différentes observations de polluants ainsi que de la date et du lieu de l'observation.

1.1 Analyse unidiemensionnelle

On s'intéresse dans un premier temps aux variables quantitatives du jeu de données (et en particulier aux émissions de polluants).
La figure 1 présente une visualisation de quelques variables quantitatives brutes.

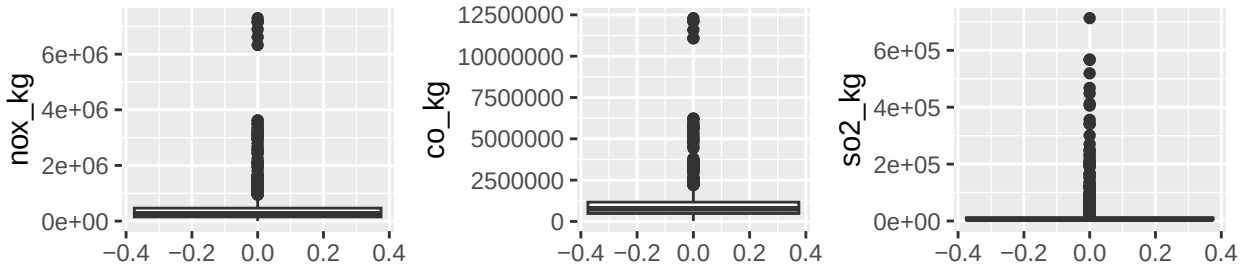


Figure 1: Boxplot des variables nox_kg,co_kg,so2_kg

On observe une très grande variance de certaines données comme co_kg. En observant l'histogramme des données quantitatives, on observe une distribution fortement asymétrique. Ainsi, si l'on souhaite effectuer des analyses sur ces données (comme par exemple une analyse en composantes principales), nos résultats seront biaisés par la variance et l'asymétrie des données. On transforme donc les données, comme présenté à la figure suivante.

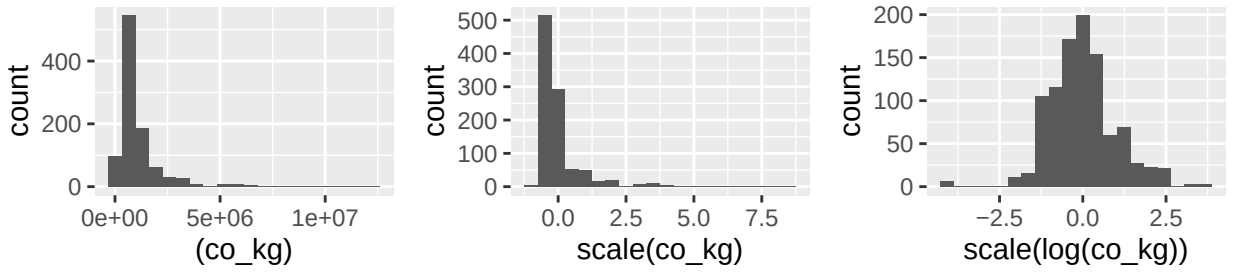


Figure 2: Histogramme de la variable co_kg en brute, scale et scale(log())

La transformation la plus adaptée est la transformation `scale(log())` : Elle permet de mettre les données à la même échelle et de réduire l'asymétrie des données pour avoir une distribution plus proche d'une loi normale.
Par la suite, on manipule les variables quantitatives transformées `scale(log())`.

On étudie ensuite la corrélation entre les variables quantitatives.

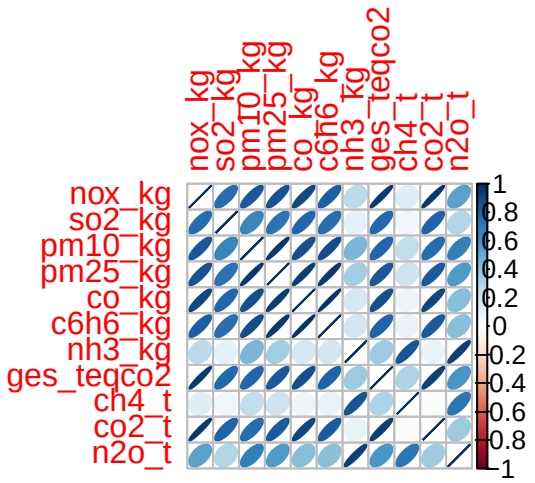


Figure 3: Corrélation entre les variables

L'analyse de la figure 3 nous permet d'identifier rapidement les relations significatives entre nos variables. Les ellipses fortement allongées suggèrent une corrélation plus forte, tandis que les ellipses plus circulaires indiquent une corrélation plus faible. Par exemple, on note une forte corrélation entre nox_kg et ges_teqco2.

1.2 Analyse multidimensionnelle

A partir de notre jeu de données, on va chercher à résumer l'information en un nombre de variables synthétiques plus faible. On effectue pour cela deux types d'analyses : une analyse en composante principale (ACP) et une analyse en composante multiple (MCA).

1.2.1 Analyse en Compomantes Principales (ACP) des variables quantitatives

On s'interesse aux variables quantitatives (émissions de polluants). On cherche à visualiser les individus dans un espace de dimension réduite. Nous effectuons donc une ACP sur les variables quantitatives.

On affiche dans un premier temps le cercle des corrélations.

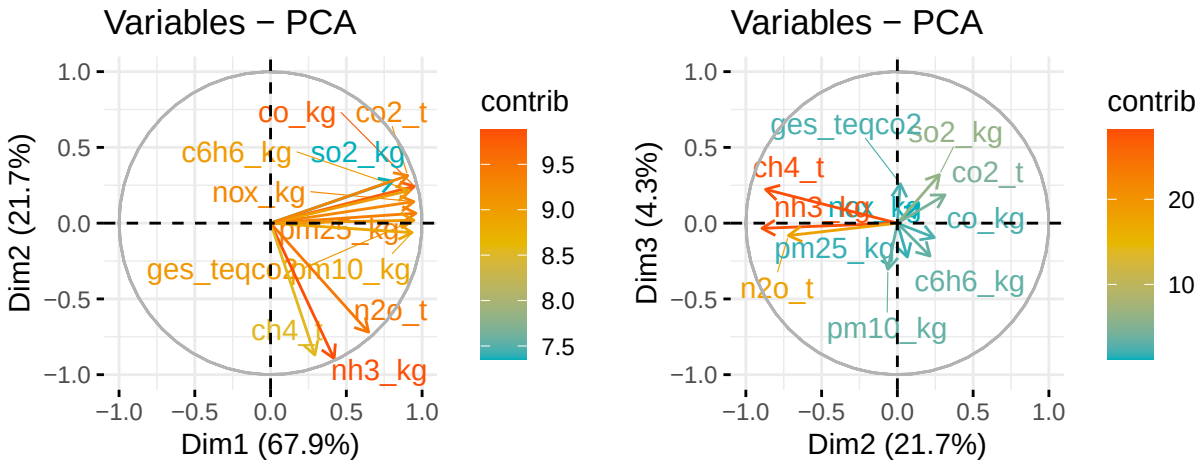


Figure 4: Cercle des corrélations

Le premier axe est une combinaison linéaire de co_kg, co2_t, c6h6_kg, nox_kg, pm25_kg, ges_teqco2 et pm10_kg :

$$C_1 = \alpha_1 \text{co\_kg} + \alpha_2 \text{co2\_t} + \alpha_3 \text{c6h6\_kg} + \alpha_4 \text{nox\_kg} + \alpha_5 \text{pm25\_kg} + \alpha_6 \text{ges\_teqco2} + \alpha_7 \text{pm10\_kg} \quad \text{avec } \alpha_i > 0$$

Le deuxième axe est une combinaison linéaire de n2o_t, nh3_kg et ch4_t :

$$C_2 = \beta_1 \text{n2o\_t} + \beta_2 \text{nh3\_kg} + \beta_3 \text{ch4\_t} \quad \text{avec } \beta_i < 0$$

On a également le pourcentage de variance expliquée par chaque axe à la figure 5.

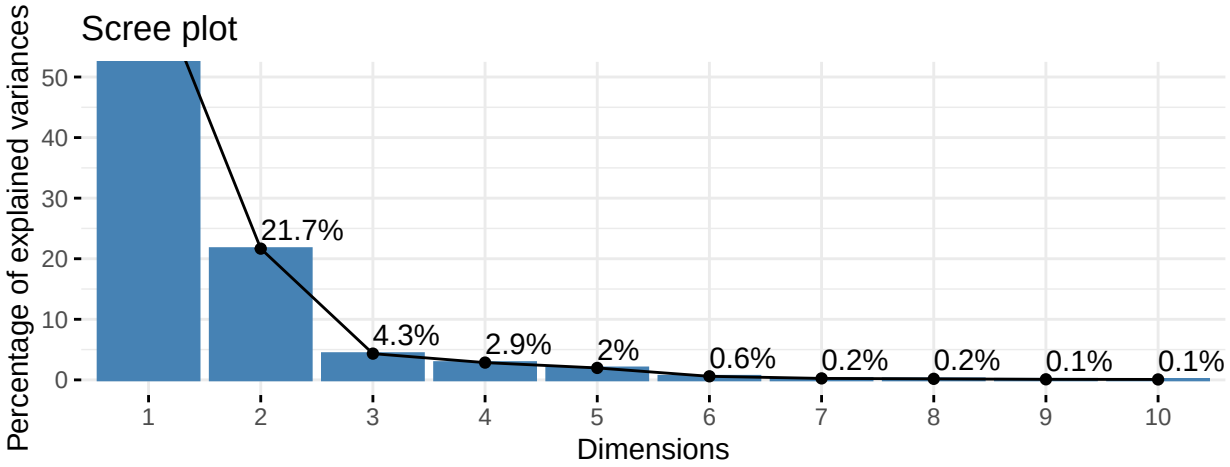


Figure 5: Pourcentage de variance expliquée par chaque axe

On retrouve bien le fait que les deux premiers axes expliquent presque 90% de la variance.

On visualise maintenant les individus dans le plan factoriel des deux premiers axes principaux en fonction de l'année puis du type d'EPCI.

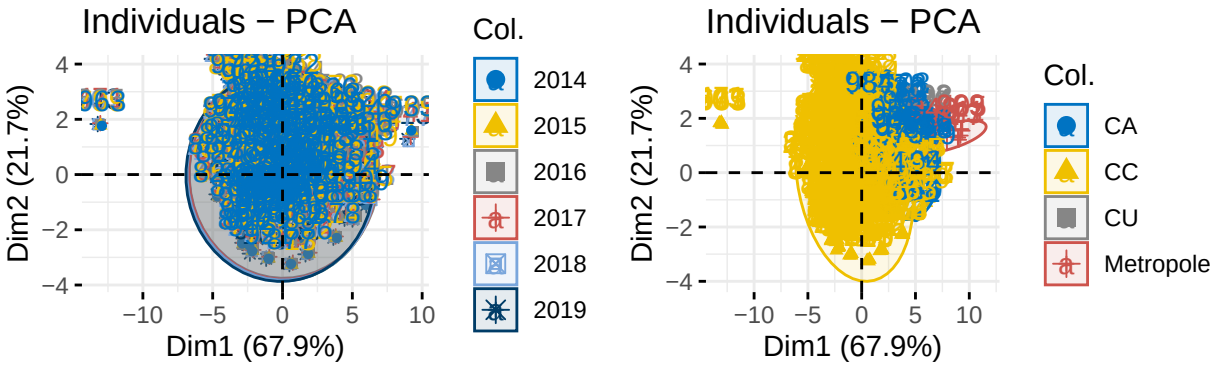


Figure 6: ACP des variables quantitatives

On observe sur la figure 6 que l'année ne semble pas beaucoup influer sur les variables. En effet, les points de différentes couleurs sur la figure de gauche se superposent, montrant que l'année ne permet pas de distinguer clairement les individus. En revanche, le type d'EPCI influe lui beaucoup plus, car on remarque que les points de différentes couleurs sont bien séparés.

1.2.2 Réduction de dimension (MCA)

Dans cette partie, on cherche à effectuer une réduction de dimension pour les polluants et du type EPCI. Nous allons donc utiliser une MCA (Multiple Correspondance Analysis). Les polluants sont des variables quantitatvives nous avons donc besoin de discrétiser ces variables. Nous allons former un nombre fini d'intervalles qui formeront les modalités des nouvelles variables qualitatives.

Nous allons aussi retirer les valeurs aberrantes c'est-à-dire en-dehors des quantiles (voir figure 1). En effet, la MCA est sensible aux valeurs extrêmes car elle vise à maximiser la variance des données. Les outliers, en raison de leur nature inhabituelle, peuvent influencer significativement la variance et ainsi biaiser les résultats de l'analyse.

Les données quantitatives sont enrichies en incluant la colonne avec la variable qualitative, puis les données quantitatives sont transformées en données qualitatives afin de réaliser une Analyse en Composantes Principales (MCA) à l'aide de FactoMineR.

Ensuite, nous appliquons l'Analyse en Composantes Principales à l'aide de la bibliothèque factoMineR, en variant les intervalles de découpage des données quantitatives en données qualitatives.

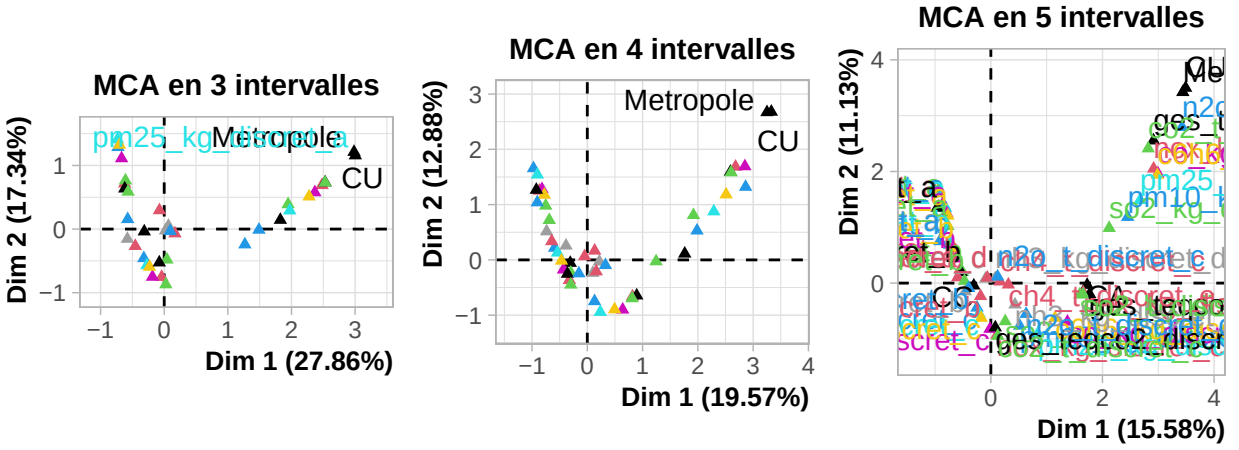


Figure 7: MCA avec découpage des données en 3, 4 et 5 intervalles

L'analyse des résultats de la MCA révèle une structure significative lorsque les variables sont regroupées selon un découpage en trois intervalles. Dans ce scénario, les variables partageant le même découpage d'intervalles présentent un regroupement cohérent, suggérant une association claire entre ces catégories.

Les deux premiers axes principaux de l'Analyse en Composantes Principales (MCA) capturent un pourcentage significatif de la variance totale, avec des valeurs respectives de 27% et 17%. Ces résultats indiquent que ces axes fournissent une représentation robuste des relations entre les variables, soulignant des patterns structurés dans les données.

Cependant, lorsqu'on effectue un découpage en un plus grand nombre d'intervalles, les pourcentages associés aux axes principaux diminuent, suggérant une dispersion accrue des données. Cela peut être interprété comme une indication que le découpage en trois intervalles offre une simplification pertinente, condensant l'information tout en préservant la structure sous-jacente, tandis qu'un découpage plus fin pourrait introduire du bruit ou de la complexité excessive.

En résumé, l'analyse suggère que le découpage en trois intervalles optimise la représentation des variables, offrant une compréhension significative des relations dans les données, tandis qu'un découpage plus fin pourrait conduire à une perte de clarté et à une dilution de l'information utile.

2 Classification des EPCI

On cherche à classer les EPCI en fonction de leurs émissions de polluants. On utilise pour cela différentes méthodes de classification.

2.1 Clustering

On met en place différents algorithmes de clustering :

2.1.1 Méthodes des k-means

On détermine combien de classes choisir en observant le comportement de l'inertie intraclasse en fonction du nombre de classes (k) :

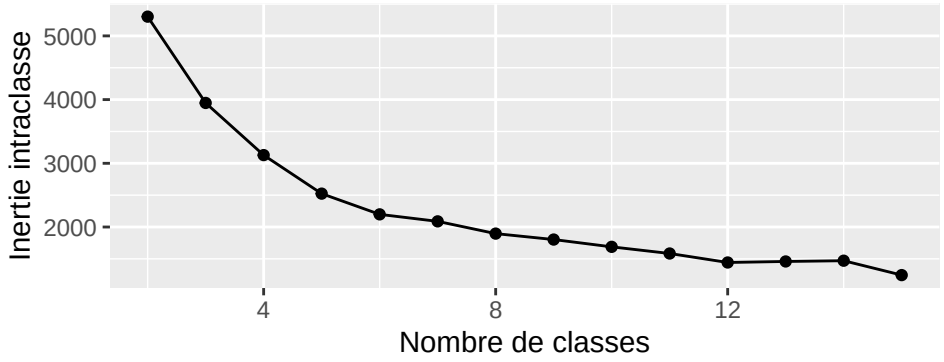


Figure 8: Determination du nombre de clusters optimal

On observe un coude sur le graphe de l'inertie itntraclasse à partir de K = 5 classes. On choisit donc 5 classes d'après le critère des K-means, et on obtient ainsi le résultat suivant :

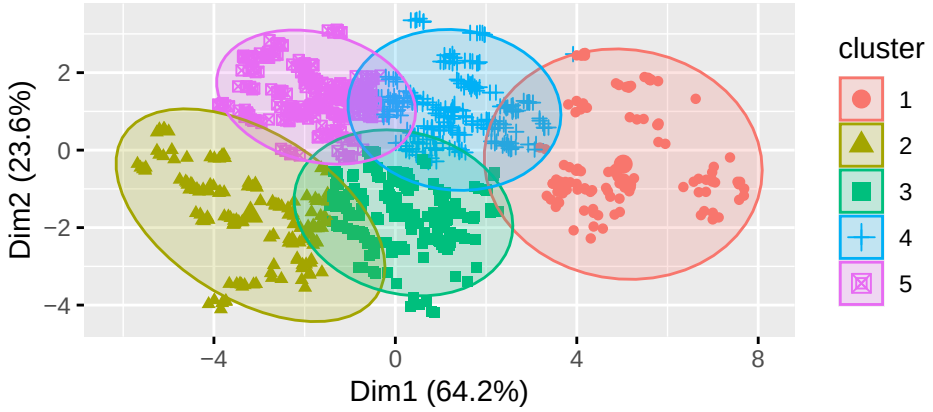


Figure 9: K-means avec K=5

2.1.2 Critère de sélection Silhouette

Toujours en faisant varier k, on extrait la moyenne des indices de silhouette de chaque cluster, afin d'obtenir le graphe suivant :

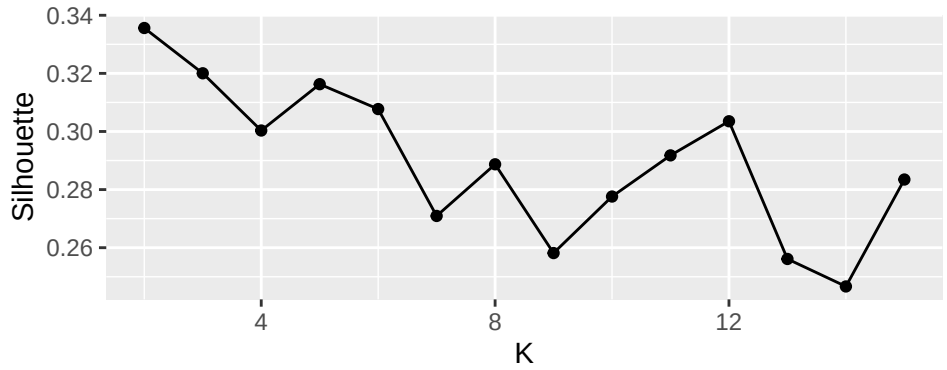


Figure 10: Critère de sélection Silhouette

On choisit le pic du graphe de Silhouette qui est atteint pour $K = 2$. La méthode des K-means proposait $K = 4$ clusters.
Cela peut s'expliquer par le fait que les méthodes de clustering ont des objectifs différents : la méthode des k-means se concentre sur la minimisation de la variance intra-cluster, tandis que Silhouette va se concentrer sur la séparation entre les clusters et l'homogénéité à l'intérieur des clusters.
Ces considérations ne sont cependant pas absolues et dépendent des données en question. C'est pour cela que nous allons exploiter d'autres méthodes de clustering. Voici les résultats graphiques obtenus pour $K = 2$ avec Silhouette :

```
##   cluster size ave.sil.width
## 1       1  600          0.35
## 2       2  359          0.32
```

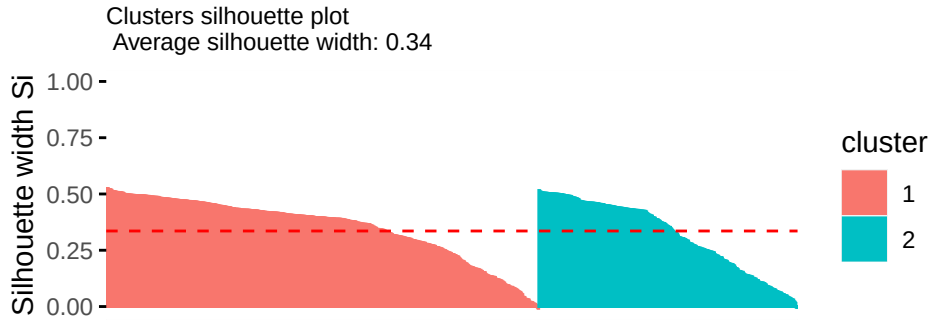


Figure 11: Silhouette avec K=2

Au vu des des silhouettes du graphe, il y a une bonne répartition des clusters. En effet, les classes semblent homogènes en termes d'effectif. De plus, on note qu'il n'y a pas de $s(i)$ négatifs, signifiant que les points de chaques classes sont assez éloignés des autres classes.

2.1.3 Mélanges Gaussiens

Critère de sélection BIC et ICL

Les critères BIC et ICL ne nous ont pas permis d'obtenir de résultats satisfaisants, et leur compilation est très longue en raison du grand nombre de composantes que nous avons dû afficher pour espérer avoir un critère d'arrêt. Nous n'avons donc pas jugé utile d'afficher les différents graphiques obtenus. En effet, avec un premier affichage en faisant varier le nombre de composantes de 2 à 50, aucun critère d'arrêt n'a été trouvé par le critère BIC. Cependant, le graphe nous a permis d'écarter tous les modèles sphériques et diagonaux, qui fournissent des résultats bien moins bons que les autres.

En conservant uniquement les modèles les plus « performants », on remarque que le modèle retenu dans tous les cas est VEV. Cependant, il semblerait que le critère n'arrive toujours pas à trouver un point d'arrêt pour le nombre optimal de clusters. Nous avons donc utilisé le critère ICL sur les coordonnées de l'ACP. Cependant, aucun résultant intéressant n'a été obtenu avec celui-ci aussi, le nombre de classes suggéré étant beaucoup trop grand.

Nous avons donc cherché à savoir quand est-ce que le graphe du critère ICL pour le modèle VEV se stabilisait. Une fois encore, le résultat n'était pas interprétable, compte tenu du fait que le graphe commençait à stagner à partir de $K = 50$ classes, ce qui est beaucoup trop grand.

2.1.4 Dendogrammes

Nous avons décidé d'écarter les mesures d'agrégation single et complete en raison de leurs défauts (sensibilité aux données bruitées, effet de chaînage,...). La mesure d'agrégation de Ward, quant à elle, tend à former des groupes avec des effectifs équilibrés à un niveau hiérarchique donné et c'est pourquoi nous avons décidé de l'utiliser pour la partie CAH.

Avec le critère Pseudo-F (Calinski-Harabasz), on observe un pic sur le graphe atteint pour un nombre de cluster égal à deux. Avec le critère Silouhette, nous obtenons aussi le même résultat.

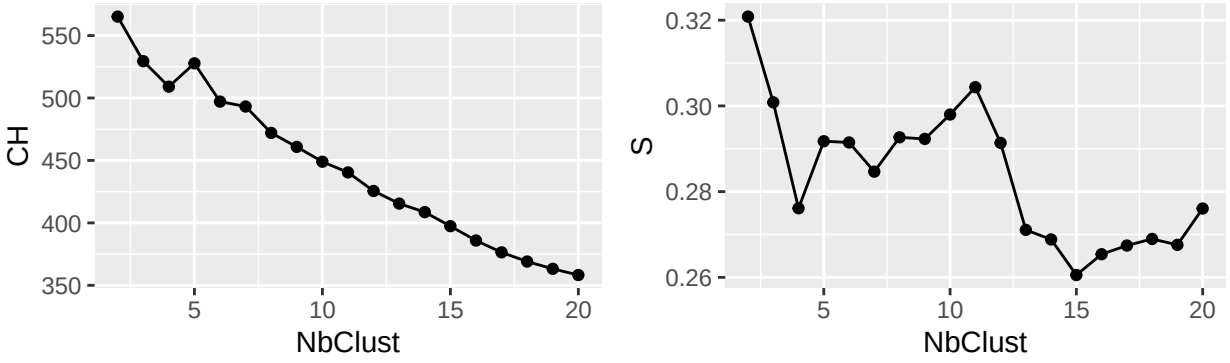
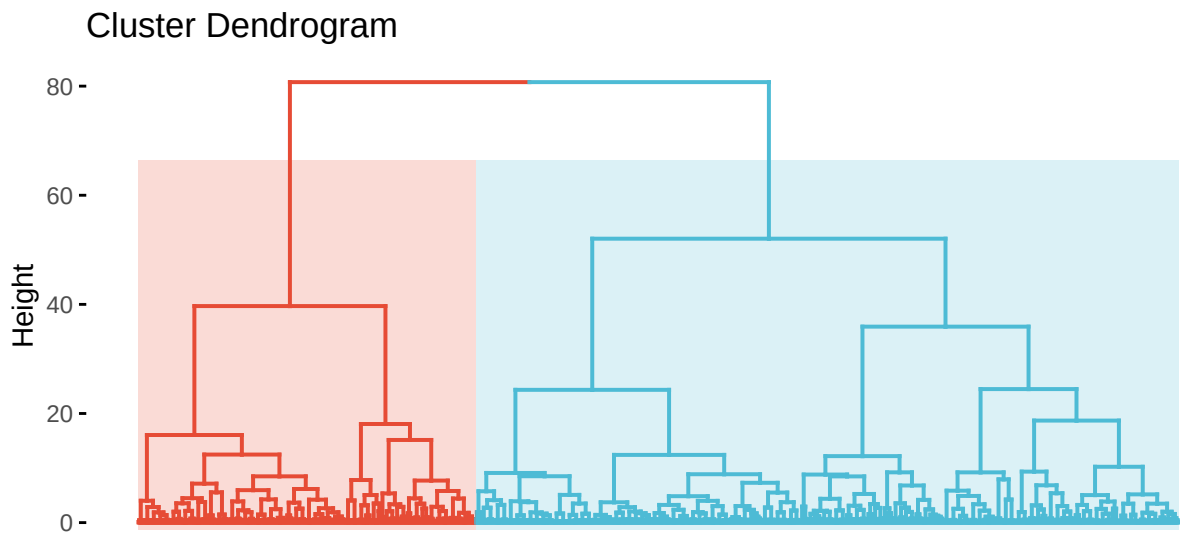


Figure 12: Critère de sélection Calinski-Harabasz et Silhouette

Ainsi nous obtenons la répartition des données suivante :



2.2 Analyse discriminante linéaire

Dans la partie précédente, nous avons effectué plusieurs types de clustering pour regrouper les données. Le clustering regroupe les individus de manière non supervisée. Dans cette partie, nous allons essayer de regrouper les différentes EPCI en fonction de critères prédéfinis. Dans un premier temps, nous étudierons le dépassement d'émission de méthane de 1000 tonnes par an, puis nous nous intéresserons au type d'EPCI.

On effectue une analyse linéaire discriminante. Cette méthode consiste à faire une analyse des composantes principales sur les centroïdes des classes, avec la métrique de Mahalanobis. Cette métrique permet de "sphériser" les données. La LDA permet également de trouver la combinaisons linéaires des coordonnées permettant de maximiser la variance inter-classe et de minimiser la variance intra-classe.

2.2.1 Taux d'émission de méthane

Dans notre cas, nous créeons une nouvelle variable binaire, valant 1 si le taux d'émission de méthane dépasse les 1000 tonnes par an, et 0 sinon. Nous effectuons ensuite une LDA, et nous pouvons visualiser les résultats dans la figure **??**.

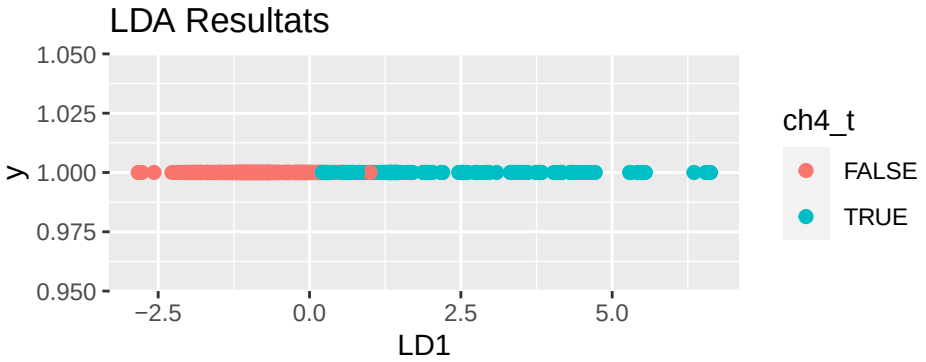


Figure 13: LDA sur le taux de méthane

Premièrement, nous remarquons que la LDA n'a qu'une seule dimension. C'est parce que sa dimension vaut le nombre de modalités moins un. Comme nous avons une variable binaire, le résultat de la LDA ne contient donc qu'une dimension. Deuxièmement, nous remarquons que le taux d'émission de méthane sépare ici plutôt bien les données. En effet, les individus en dessous du seuil ont une coordonnée assez faible (négative ou proche de 0). Tandis que ceux dont le taux de méthane est supérieur au seuil ont une coordonnée grande.

Afin de vérifier la capacité de classification du taux de méthane, nous allons effectuer une prédiction. La LDA précédente a été faite sur 70% des individus, afin de pouvoir faire une prédiction sur les 30% restants. Nous obtenons les résultats sur la figure suivante :

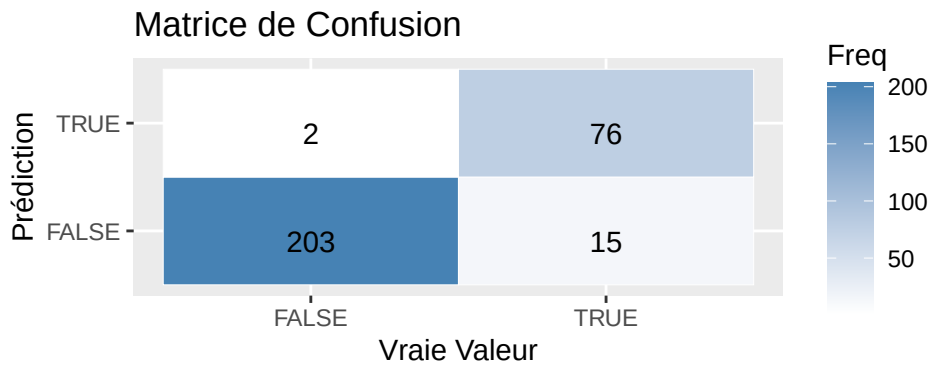


Figure 14: Prédiction sur le taux de méthane

Nous pouvons voir grâce à cette table que les individus sont plutôt bien prédits. En effet, on obtient un taux de précision de 0.943. Ainsi, utiliser le taux de méthane pour classer les individus de façon supervisée semble judicieux, car pratiquement 95% pourcent des individus seraient correctement prédits avec ce procédé.

2.2.2 Type d'EPCI

Nous reprenons le même procédé, mais ici avec la variable qualitative type d'EPCI. Cette variable a 4 modalités, nous allons donc avoir une LDA a trois dimensions. Nous pouvons visualiser le résultat de la LDA dans la figure **??**. Nous pouvons afficher le résultat pour les trois dimensions de la LDA, mais nous avons seulement afficher dans les deux premières dimensions dans la figure 3, car c'est l'affichage le plus parlant.

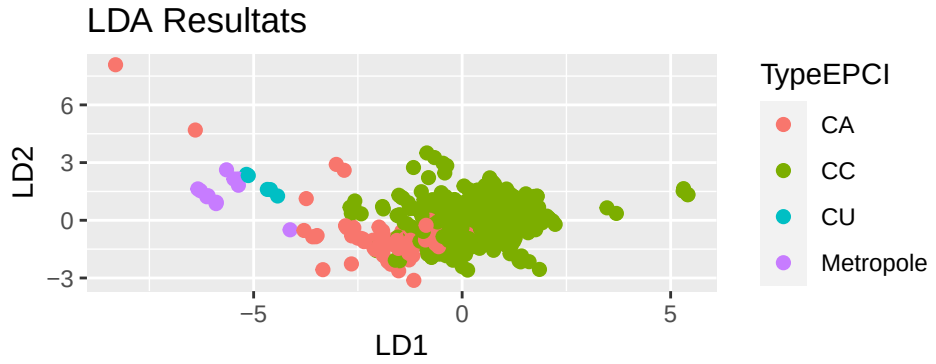


Figure 15: LDA en fonction des types EPCI

Nous pouvons voir que les données semblent bien séparées, chaque type d'EPCI. Le type d'EPCI semble bien séparé les données également, et nous allons confirmer ça par quelques prédictions. Comme pour le taux de

méthane, la LDA a été faite sur 70% des données, et nous allons maintenant faire une prédiction sur les 30% restants.

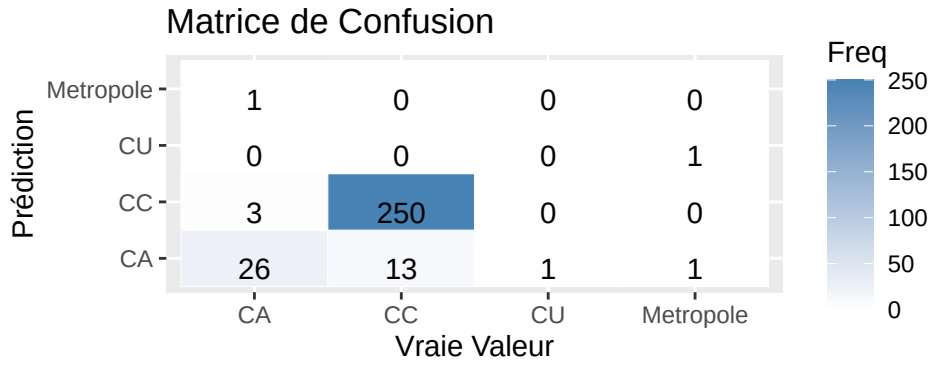


Figure 16: Prédiction sur le type d'EPCI

Nous pouvons voir grâce à la figure 16 table que les individus sont plutôt bien prédits. On obtient un taux de précision de 0.932. Ainsi, le type d'EPCI différencie bien les individus, comme le laisser présager les résultats de l'ACP. et nous obtenons un bon taux de précision. Cependant, il y a une forte dissimilarité entre les nombres d'individus par modalité.

On essaie alors de regrouper les modalités de type d'EPCI. On compare les résultats des LDA appliquées sur les regroupements suivants :

- "CU" et "Métropole"
- "CU", "Métropole", et "CA"

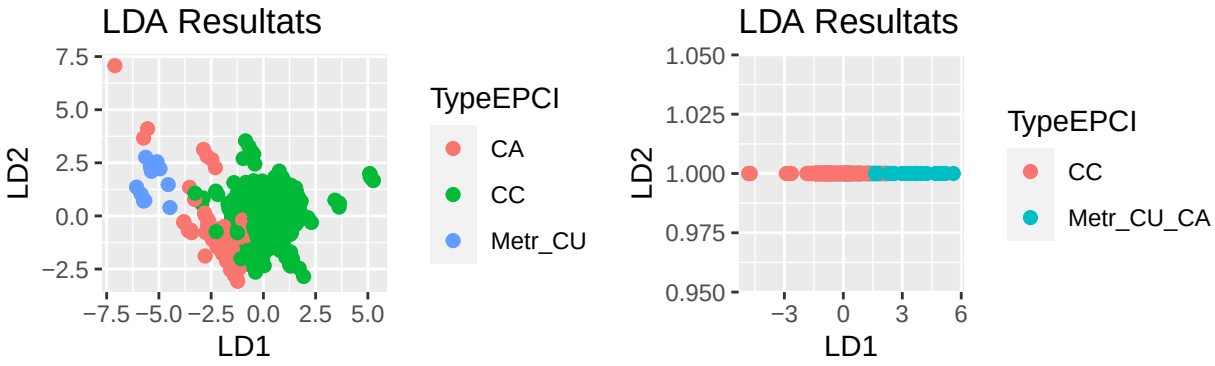


Figure 17: LDA en fonction des types EPCI

Nous remarquons que nous obtenons maintenant des LDA de dimensions 2 et 1. Visuellement, nous ne pouvons pas voir si ces regroupements ont été efficaces. En effet, c'est principalement les classes CA et CC qui sont proches. Ainsi, lors du premier regroupement, nous observons un résultat très similaire au résultat initial. Pour le deuxième regroupement, on semble pouvoir observer que les "CC" ont une coordonnée assez faible, contrairement aux "Metr_CU". Séparer les données à partir de ce regroupement semble plus simple, voyons si les prédictions confirment ceci.

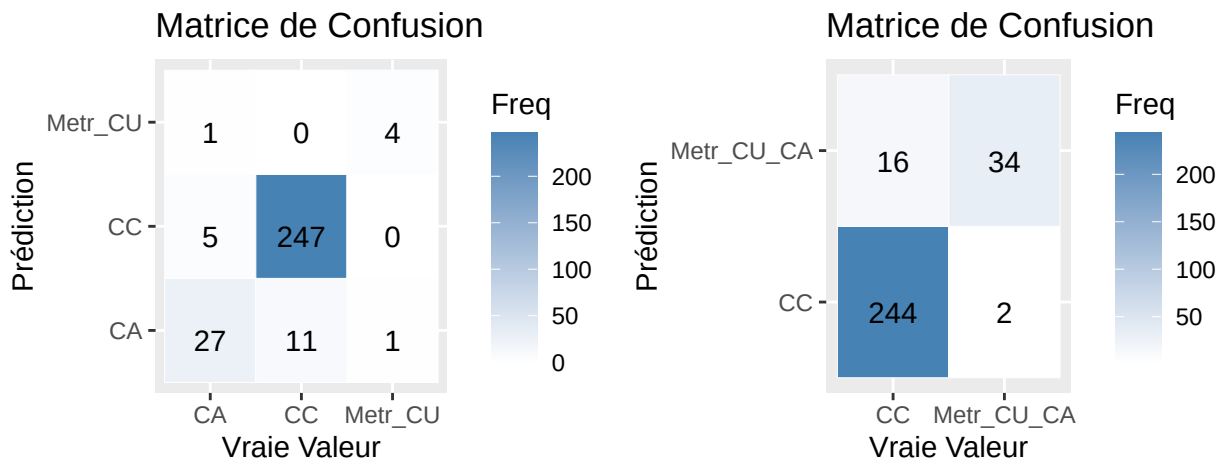


Figure 18: Prédiction en fonction des types EPCI simplifiés

Nous obtenons un taux de précision de 0.939 pour le premier regroupement, et de 0.939 pour le deuxième. Ainsi, contrairement à ce qu'on a pu penser, nous ne gagnons pas largement en précision en faisant des regroupements.

Cela vient probablement du fait que les classes "CA" et "CC" sont les plus proches, et donc l'erreur vient principalement d'une erreur de prédiction entre ces deux classes. Or, nos regroupements n'ont pas agréger ces deux classes, n'améliorant donc pas la précision.

3 EMS

3.1 Modèle linéaire

3.1.1 Modèle d'ANOVA

On explique le gaz à effet de serre en fonction des variables TypeEPCI et années.
Soit $T = \{CC, CA, CU, Metropole\}$, l'ensemble des types d'EPCI et $A = \{2015, 2016, 2017, 2018, 2019\}$ l'ensemble des années.
Soit ges_teqco2_{ij} le gaz à effet de serre de l'EPCI i à l'année j, avec $i \in T$ et $j \in A$.

On utilise le modèle d'ANOVA à deux facteurs avec interaction suivant :

$$\text{(Mod1)} : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta j + \gamma_{ij} + \varepsilon_{ij} \\ \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Sur R : `lm(ges_teqco2 ~TypeEPCI * annee_inv, data=dlog)`

On obtient un R^2 ajusté de $0.416 \approx 0.5$. Cela signifie que le modèle n'explique que partiellement la variance des données.

On essaie de simplifier le modèle en enlevant les interactions avec un test de sous-modèle :

$$\mathcal{H}_0 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad \text{contre} \quad \mathcal{H}_1 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

C'est bien un test de sous-modèle car on enlève des variables explicatives.

Sur R : `anova(modele_avec_interactions,modele_sans_interactions)`

On obtient une p-value de $1 > 0.05$.
On ne rejette pas l'hypothèse de nullité des interactions.
On garde donc le modèle suivant :

$$\text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{avec } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

On essaie de simplifier le modèle en enlevant une des variables explicatives (on fait 2 tests de sous-modèle) :

$$\mathcal{H}_0 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad \text{contre} \quad \mathcal{H}_1 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

et

$$\mathcal{H}_0 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \beta_j + \epsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad \text{contre} \quad \mathcal{H}_1 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Pour le modèle dépendant uniquement du type d'EPCI, on obtient une p-value de $0.599 > 0.05$.
On peut donc enlever l'année dans le modèle :

$$\text{ges\_teqco2}_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{avec } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

On essaie à nouveau de simplifier le modèle en enlevant les variables explicatives :

$$\mathcal{H}_0 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad \text{contre} \quad \mathcal{H}_1 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

On obtient cette fois une p-value de $0 < 0.05$.
On ne peut donc pas enlever le type d'EPCI dans le modèle.

On vérifie finalement la cohérence du modèle retenu :

$$\mathcal{H}_0 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad \text{contre} \quad \mathcal{H}_1 : \begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

On obtient une p-value de $0.99 > 0.05$ donc le modèle est cohérent. On garde donc le modèle :

$$\begin{cases} \text{ges\_teqco2}_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ \\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Nous vérifions le modèle que nous avons obtenu en faisant un peu de prédiction. Pour cela, on crée le modèle sur 70% des données, et on essaye de prédire la valeur du gaz à effet de serre sur les 30% restants.
Les résultats obtenus montrent un écart moyen entre la réalité et la prédiction de 13.14 %.

3.1.2 Régression linéaire

On explique le gaz à effet de serre en fonction de tous les autres polluants.
On considère le modèle de régression linéaire suivant :

$$\begin{cases} \text{ges\_teqco2}_i = \theta_0 + \theta_1 \text{nox\_kg}_i + \theta_2 \text{so2\_kg}_i + \theta_3 \text{pm10\_kg}_i + \theta_4 \text{pm25\_kg}_i + \theta_5 \text{co\_kg}_i \\ \qquad\qquad + \theta_6 \text{c6h6\_kg}_i + \theta_7 \text{nh3\_kg}_i + \theta_8 \text{ch4\_t}_i + \theta_9 \text{co2\_t}_i + \theta_{10} \text{no2\_t}_i + \varepsilon_i \\ \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

```
##
## Call:
## lm(formula = ges_teqco2 ~ ., data = data_scaled_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62828 -0.09113 -0.01756  0.06314  0.75733
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.396e-16  5.098e-03   0.000  1.00000
## nox_kg       3.184e-01  3.167e-02  10.054  < 2e-16 ***
## so2_kg       8.638e-02  9.056e-03   9.538  < 2e-16 ***
## pm10_kg     -2.741e-01  3.260e-02  -8.408  < 2e-16 ***
## pm25_kg      1.841e-01  4.030e-02   4.569 5.54e-06 ***
## co_kg        1.383e-01  5.144e-02   2.688  0.00731 **
## c6h6_kg     -1.263e-01  4.333e-02  -2.914  0.00365 **
## nh3_kg      -2.694e-01  3.065e-02  -8.789  < 2e-16 ***
## ch4_t        2.435e-01  1.308e-02  18.610  < 2e-16 ***
## co2_t        4.838e-01  3.071e-02  15.756  < 2e-16 ***
## n2o_t        3.714e-01  3.235e-02  11.481  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1599 on 973 degrees of freedom
## Multiple R-squared:  0.9747, Adjusted R-squared:  0.9744
## F-statistic:  3746 on 10 and 973 DF,  p-value: < 2.2e-16
```

Sur le résultat affiché, on obtient une p-valeur de 0.004 pour la variable c6h6_kg, et une p-valeur de 0.007 pour la variable co_kg. Ces deux p-valeurs sont supérieures à 0.05. Cela pourrait suggérer la possibilité de les exclure du modèle afin de le simplifier.

Nous allons maintenant simplifier le modèle en selctionnant les variables explicatives pertinentes.

Avec la méthode backward

On obtient les résultats suivants avec la méthode backward.
Sur R : `regsubsets(ges_teqco2~.,data=data_scaled_df,nbest=1,nvmax=10,method="backward")`

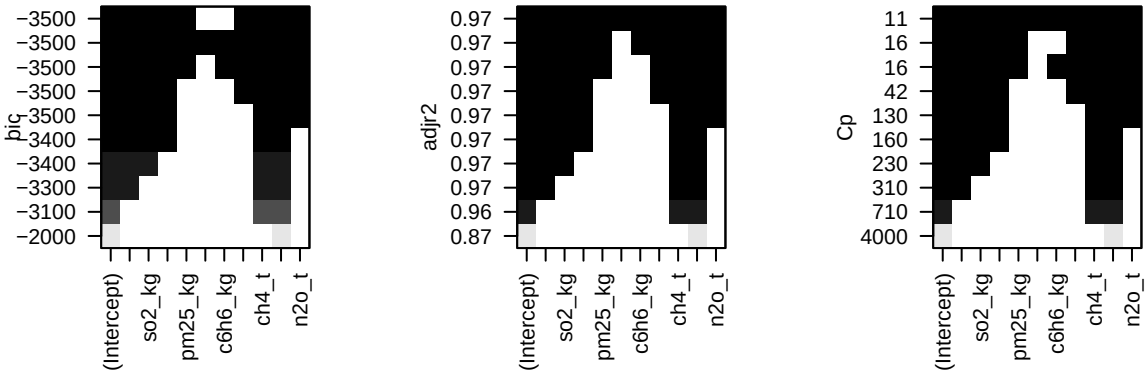


Figure 19: Sélection des variables explicatives en backward

En utilisant la méthode Backward, tous les critères conduisent à la même sélection de variables, celle pour laquelle nous avions formulé l'hypothèse précédemment lors des tests de nullité.

Voici le modèle simplifié proposé avec la méthode backward:

$$\begin{cases} \text{ges\_teqco2}_i = \theta_0 + \theta_1 \text{nox\_kg}_i + \theta_2 \text{so2\_kg}_i + \theta_3 \text{pm10\_kg}_i + \theta_4 \text{pm25\_kg}_i \\ \qquad\qquad + \theta_5 \text{nh3\_kg}_i + \theta_6 \text{ch4\_t}_i + \theta_7 \text{co2\_t}_i + \theta_8 \text{no2\_t}_i + \varepsilon_i \\ \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Avec la méthode forward

Nous allons maintenant effectuer la même selection mais cette fois-ci avec la méthode forward pour vérifier la simplification possible du modèle. Et nous obtenons les résulats suivant avec les critères bic, Cp et R2.

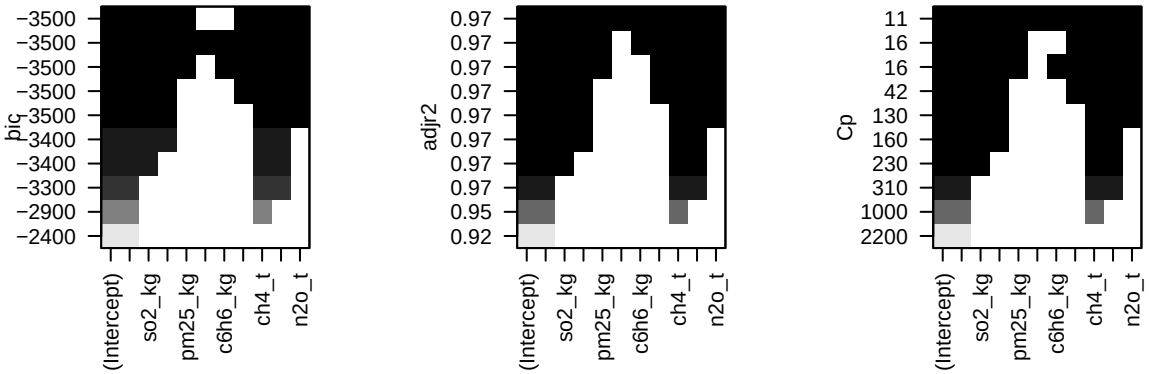


Figure 20: Sélection des variables explicatives en forward

Les résultats obtenus à partir de tous les critères concordent avec ceux de la méthode backward pour simplifier le modèle en éliminant les variables co_kg et c6h6_kg. Il est désormais essentiel de valider ce sous-modèle.

Nous prévoyons de réaliser un test de sous modèle pour évaluer la performance du modèle simplifié par rapport au modèle complet. Cette comparaison nous permettra de déterminer si le modèle simplifié conserve une précision de prédiction similaire tout en étant plus parcimonieux.

$$\mathcal{H}_0 : \begin{cases} \text{ges\_teqco2}_i = \theta_0 + \theta_1 \text{nox\_kg}_i \\ \qquad\qquad +\theta_2 \text{so2\_kg}_i + \theta_3 \text{pm10\_kg}_i \\ \qquad\qquad +\theta_4 \text{pm25\_kg}_i + \theta_5 \text{nh3\_kg}_i \\ \qquad\qquad +\theta_6 \text{ch4\_t}_i + \theta_7 \text{co2\_t}_i \\ \qquad\qquad +\theta_8 \text{no2\_t}_i + \varepsilon_i \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases} \quad \text{contre} \quad \mathcal{H}_1 : \begin{cases} \text{ges\_teqco2}_i = \theta_0 + \theta_1 \text{nox\_kg}_i \\ \qquad\qquad +\theta_2 \text{so2\_kg}_i + \theta_3 \text{pm10\_kg}_i \\ \qquad\qquad +\theta_4 \text{pm25\_kg}_i + \theta_5 \text{co\_kg}_i \\ \qquad\qquad +\theta_6 \text{c6h6\_kg}_i + \theta_7 \text{nh3\_kg}_i \\ \qquad\qquad +\theta_8 \text{ch4\_t}_i + \theta_9 \text{co2\_t}_i \\ \qquad\qquad +\theta_{10} \text{no2\_t}_i + \varepsilon_i \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

La p-valeur obtenue est de $0.014 > 0.01$. On ne rejette donc pas $\mathcal{H}_0$ au risque 1% et on peut simplifier le modèle additif en un sous-modèle :

$$\begin{cases} \text{ges\_teqco2}_i = \theta_0 + \theta_1 \text{nox\_kg}_i + \theta_2 \text{so2\_kg}_i + \theta_3 \text{pm10\_kg}_i + \theta_4 \text{pm25\_kg}_i \\ \qquad\qquad +\theta_5 \text{nh3\_kg}_i + \theta_6 \text{ch4\_t}_i + \theta_7 \text{co2\_t}_i + \theta_8 \text{no2\_t}_i + \varepsilon_i \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Nous faisons ensuite un autoplot, afin de pouvoir vérifier les différentes hypothèses d'un modèle linéaire sur la figure suivante :

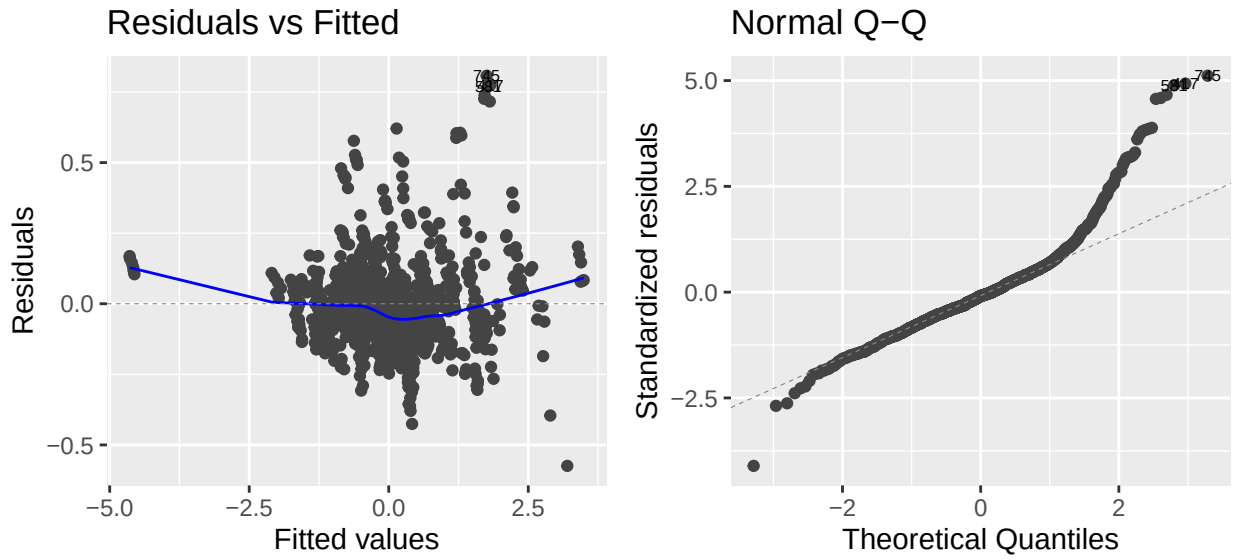


Figure 21: Autoplot modèle régression linéaire

Premièrement, les ϵ_i doivent être centré en 0. Quand on regarde le premier graphe, on remarque que les résidus $\hat{\epsilon_i}$ semblent centrés en 0. La deuxième hypothèse nous dit que tous les ϵ_i ont la même variance. Or, tous les individus semblent contenu dans un tube, nous indiquant que cette hypothèse semble vérifiée. Ensuite la troisième hypothèse est l'indépendance entre les ϵ_i et Y_i. Dans le premier graphe, il n'y a pas de forme particulière, et les ϵ_i et Y_i semblent donc indépendants. La dernière hypothèse est celle de la normalité des Y_i. En regardant le Q-Q plot, les quantiles empiriques sont plutôt proches des théoriques. Ainsi, les 4 hypothèses sont vérifiées, le modèle linéaire est donc adapté pour réprésenter ces données.

3.1.2.1 Régréssion régularisé

Nous allons maintenant effectuer une régression régularisée. Cette méthode consiste à changer la fonction à minimiser pour trouver notre estimateurs des paramètres $\hat{\theta}$. Le but de cette méthode est d'obtenir un estimateur certes biaisé, mais qui a une variance plus petite. Il faudrait résoudre $\hat{\theta} = \text{argmin}_{\theta \in \mathbb{R}_k}(\|Y - X\theta\|^2 - \lambda pen(\theta))$. La fonction $\theta \mapsto \text{pen}(\theta)$ dépend du type de régression régularisée. Nous allons voir la régression de Ridge, de Lasso et Elastic Net.

Ridge

Commençons par faire une régréssion Ridge. Cette méthode consiste à définir $pen(\theta) = \|\theta\|_2^2$. On commence par calculer la valeurs des coefficients de $\hat{\theta}$ minimisant la fonction pour différentes valeurs de λ, réprésentées dans la figure 22. Ensuite, nous faisons une validation croisée afin de trouver le λ optimal. En regardant la figure de droite de 22, on remarque que le λ retenu est 0.002. La droite rouge dans la figure de droite est tracé au niveau du lambda optimal, et nous permet de récupérer les coefficients du $\hat{\theta}$ final.

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

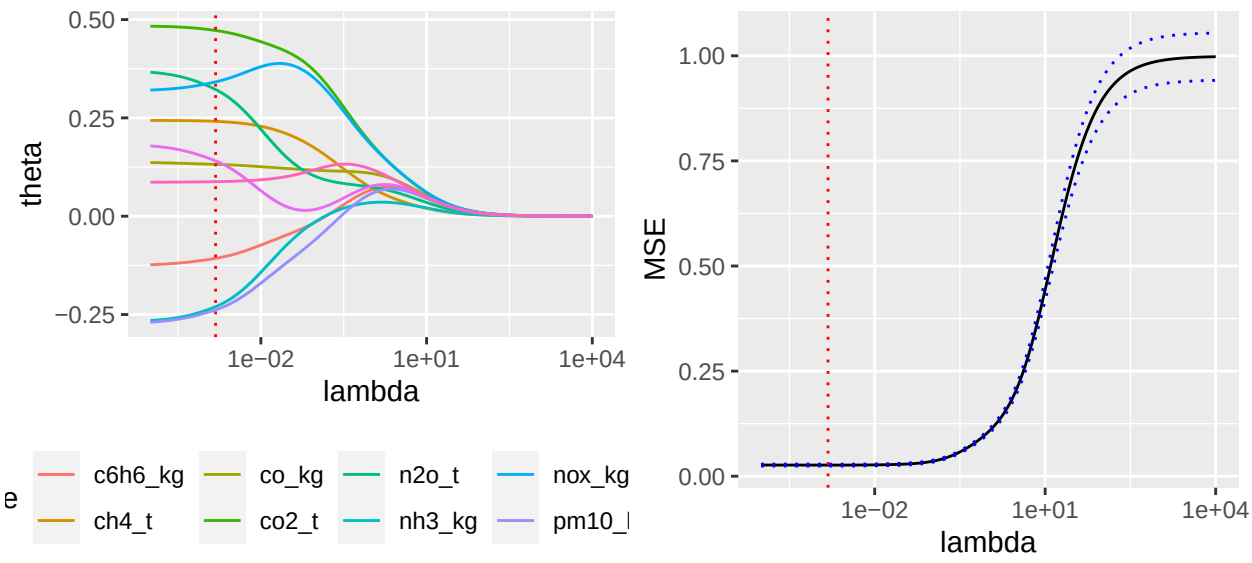


Figure 22: Régularisation Ridge

Lasso

On effectue exactement la même procédure, mais avec une régression de Lasso, qui correspond à : $pen(\theta) = \|\theta\|_1$

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

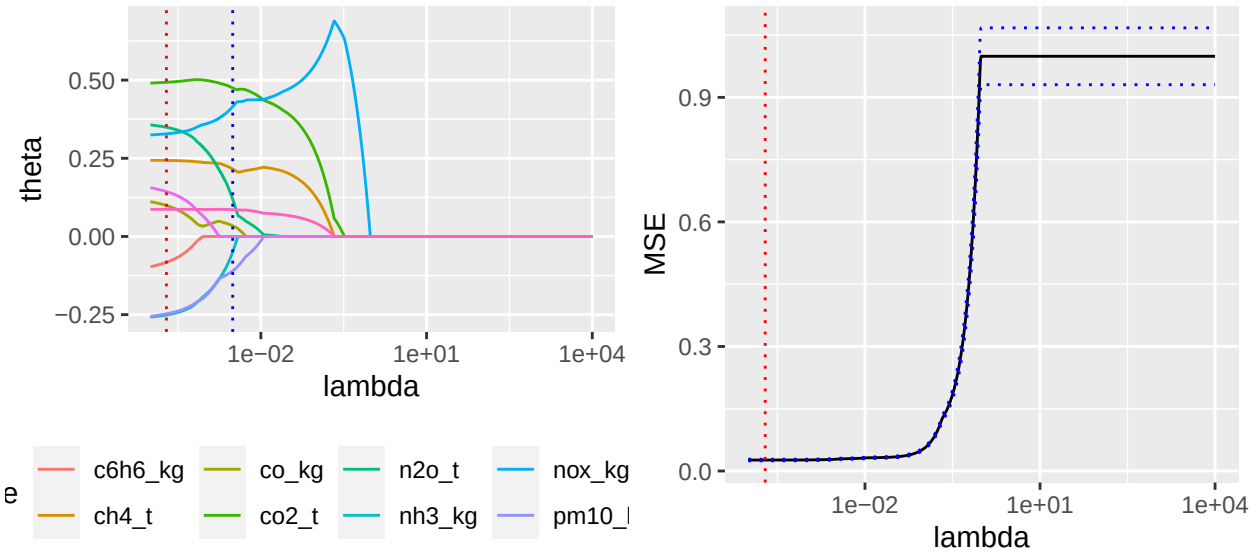


Figure 23: Régularisation Lasso

Notre λ optimal est: 0.

Elastic Net

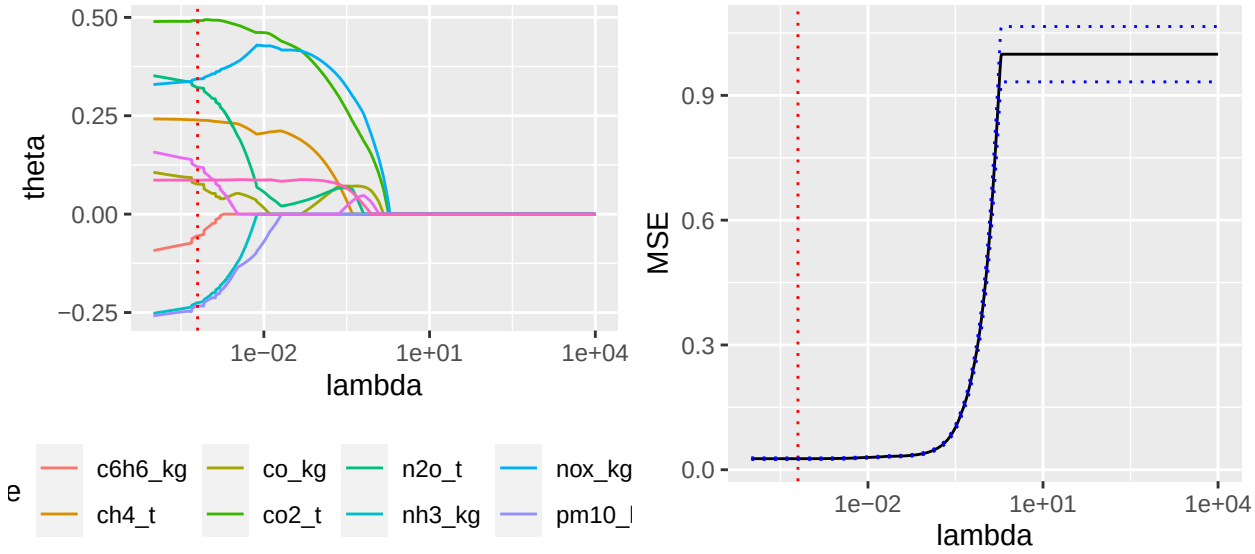


Figure 24: Régularisation Elastic Net

Le λ optimal pour Elastic Net est de : 0.001.

3.1.2.2 Analyse des résultats Pour finir avec cette partie, regardons les valeurs des différents coefficients obtenus à l'aide des méthodes de régréssions.

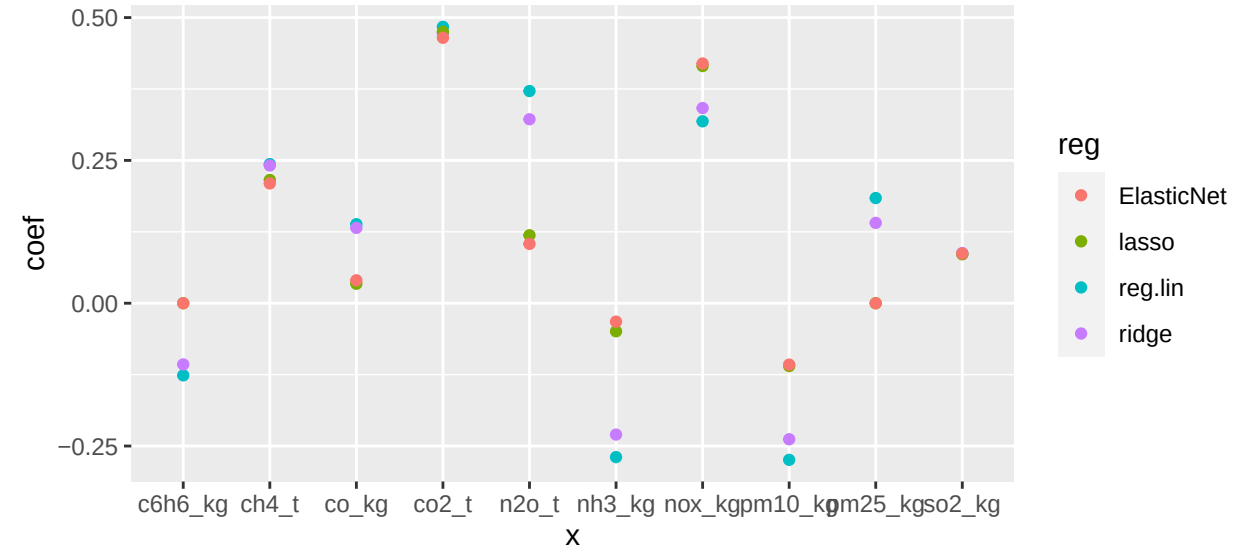


Figure 25: Résultats des différentes régularisations

Le premier point à noter, c'est que les trois régressions régularisées donnent des résultats proches, contrairement à la régression linéaire. Précédemment, nous avions vu que les coefficients de c6h6_kg et co_kg, ce que confirme la figure 25. En effet le point bleu associé à ces variables sont proches de 0.

Les régréssions régularisées annuleraient elles aussi le coefficient associé à c6h6_kg, mais pas celui de co_kg. Ces régularisations, surtout ElasticNet et Lasso, ont trouvé des valeurs proches de 0 plutôt les coefficients associés aux variables : n2o_t, nh3_kg, pm10_kg et pm25_kg.

3.1.3 ANCOVA

Dans cette partie on va chercher à expliquer l'émission de méthane en fonction de l'ammoniac, du protoxyde d'azote, du type d'EPCI et de l'année.

Modèle avec intéraction

Dans un premier temps on va considérer le modèle avec intéraction suivant :

$$\begin{cases} \text{ch4\_t}_i = \theta_0 + \theta_1 \text{nh3\_kg}_i + \theta_2 \text{n2o\_t}_i \\ \qquad\qquad +\theta_3 \mathbb{1}_{\{TypeEPCI_i=CC\}} + \theta_4 \mathbb{1}_{\{TypeEPCI_i=CU\}} + \theta_5 \mathbb{1}_{\{TypeEPCI_i=Metropole\}} \\ \qquad\qquad +\theta_6 \text{annee}_i \\ \qquad\qquad +\gamma_1 \text{nh3\_kg}_i \text{n2o\_t}_i + \gamma_2 \text{nh3\_kg}_i \mathbb{1}_{\{TypeEPCI_i=CC\}} \\ \qquad\qquad +\gamma_3 \text{nh3\_kg}_i \mathbb{1}_{\{TypeEPCI_i=CU\}} + \gamma_4 \text{nh3\_kg}_i \mathbb{1}_{\{TypeEPCI_i=Metropole\}} + \gamma_5 \text{nh3\_kg}_i \text{annee}_i \\ \qquad\qquad +\gamma_6 \text{n2o\_t}_i \mathbb{1}_{\{TypeEPCI_i=CC\}} + \gamma_7 \text{n2o\_t}_i \mathbb{1}_{\{TypeEPCI_i=CU\}} \\ \qquad\qquad +\gamma_8 \text{n2o\_t}_i \mathbb{1}_{\{TypeEPCI_i=Metropole\}} + \gamma_9 \text{n2o\_t}_i \text{annee}_i \\ \qquad\qquad +\gamma_{10} \mathbb{1}_{\{TypeEPCI_i=CC\}} \text{annee}_i + \gamma_{11} \mathbb{1}_{\{TypeEPCI_i=CU\}} \text{annee}_i + \gamma_{12} \mathbb{1}_{\{TypeEPCI_i=Metropole\}} \text{annee}_i + \varepsilon_i \\ \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

On va maintenant comparer ce modèle avec intéraction :

$$\begin{cases} \text{ch4\_t}_i = \theta_0 + \theta_1 \text{nh3\_kg}_i + \theta_2 \text{n2o\_t}_i \\ \qquad\qquad +\theta_3 \mathbb{1}_{\{TypeEPCI_i=CC\}} + \theta_4 \mathbb{1}_{\{TypeEPCI_i=CU\}} + \theta_5 \mathbb{1}_{\{TypeEPCI_i=Metropole\}} \\ \qquad\qquad +\theta_6 \text{annee}_i + \varepsilon_i \\ \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

On retrouve une p-valeur de $0 < 0.05$ donc on ne peut *pas simplifier* le modèle en enlevant les interactions (au risque 5%).

On va maintenant chercher à simplifier le modèle avec intéraction au maximum.

On va utiliser la bibliothèque "MASS" pour effectuer une sélection de modèle basée sur le critère AIC en utilisant la fonction `stepAIC()`. Plus précisément, il effectue une sélection de modèle pas à pas dans le sens inverse (backward), ce qui signifie qu'il commence par un modèle complet (incluant toutes les variables explicatives) puis retire séquentiellement les variables qui n'améliorent pas la qualité du modèle selon le critère AIC. Le résultat est stocké dans l'objet `modselect_aic`, qui contient le modèle sélectionné avec la meilleure performance selon le critère AIC.

```
modselect_aic=stepAIC(ancov,trace=T,direction="backward")
```

```
## Start:  AIC=-1804.81
## ch4_t ~ (nh3_kg + n2o_t + TypeEPCI + annee_inv)^2
##
##                      Df Sum of Sq    RSS     AIC
```

```
## <none>                             151.24 -1804.8
## - TypeEPCI:annee_inv  3    1.6326 152.87 -1800.2
## - nh3_kg:annee_inv    1    1.4239 152.66 -1797.6
## - n2o_t:annee_inv     1    1.5288 152.77 -1796.9
## - nh3_kg:n2o_t        1    2.4011 153.64 -1791.3
## - n2o_t:TypeEPCI      3    7.2447 158.48 -1764.8
## - nh3_kg:TypeEPCI     3   14.2619 165.50 -1722.1
```

On a obtient que le modèle minimisant le AIC est le modèle complet sans selection des variables.

```
modselect_bic=stepAIC(ancov,trace=T,direction="backward",k=log(nrow(dlog)))
```

```
## Start:  AIC=-1711.87
## ch4_t ~ (nh3_kg + n2o_t + TypeEPCI + annee_inv)^2
##
##                       Df Sum of Sq    RSS     AIC
## - TypeEPCI:annee_inv  3    1.6326 152.87 -1722.0
## <none>                             151.24 -1711.9
## - nh3_kg:annee_inv    1    1.4239 152.66 -1709.5
## - n2o_t:annee_inv     1    1.5288 152.77 -1708.9
## - nh3_kg:n2o_t        1    2.4011 153.64 -1703.3
## - n2o_t:TypeEPCI      3    7.2447 158.48 -1686.5
## - nh3_kg:TypeEPCI     3   14.2619 165.50 -1643.9
##
## Step:  AIC=-1721.98
## ch4_t ~ nh3_kg + n2o_t + TypeEPCI + annee_inv + nh3_kg:n2o_t +
##     nh3_kg:TypeEPCI + nh3_kg:annee_inv + n2o_t:TypeEPCI + n2o_t:annee_inv
##
##                     Df Sum of Sq    RSS     AIC
## - nh3_kg:annee_inv  1    0.5143 153.38 -1725.6
## - n2o_t:annee_inv   1    0.5619 153.43 -1725.3
## <none>                           152.87 -1722.0
## - nh3_kg:n2o_t      1    2.3164 155.19 -1714.1
## - n2o_t:TypeEPCI    3    7.6185 160.49 -1694.8
## - nh3_kg:TypeEPCI   3   14.1962 167.07 -1655.3
##
## Step:  AIC=-1725.57
## ch4_t ~ nh3_kg + n2o_t + TypeEPCI + annee_inv + nh3_kg:n2o_t +
##     nh3_kg:TypeEPCI + n2o_t:TypeEPCI + n2o_t:annee_inv
##
##                     Df Sum of Sq    RSS     AIC
## - n2o_t:annee_inv   1    0.0484 153.43 -1732.2
## <none>                           153.38 -1725.6
## - nh3_kg:n2o_t      1    2.2661 155.65 -1718.0
## - n2o_t:TypeEPCI    3    9.3741 162.76 -1687.9
## - nh3_kg:TypeEPCI   3   16.6534 170.04 -1644.8
##
## Step:  AIC=-1732.15
## ch4_t ~ nh3_kg + n2o_t + TypeEPCI + annee_inv + nh3_kg:n2o_t +
##     nh3_kg:TypeEPCI + n2o_t:TypeEPCI
##
##                     Df Sum of Sq    RSS     AIC
## <none>                           153.43 -1732.2
```

```
## - nh3_kg:n2o_t     1    2.2416 155.67 -1724.8
## - n2o_t:TypeEPCI   3    9.6576 163.09 -1692.8
## - annee_inv        1    9.9303 163.36 -1677.3
## - nh3_kg:TypeEPCI  3   17.0173 170.45 -1649.3
```

Avec le critère bic, le modèle simplifié le suivant :

$$\begin{cases} \text{ch4\_t}_i = \theta_0 + \theta_1 \text{nh3\_kg}_i + \theta_2 \text{n2o\_t}_i + \theta_3 \mathbb{1}_{\{TypeEPCI_i = CC\}} + \theta_4 \mathbb{1}_{\{TypeEPCI_i = CU\}} + \theta_5 \mathbb{1}_{\{TypeEPCI_i = Metropole\}} \\ \qquad\qquad +\theta_6 \text{annee}_i \\ \qquad\qquad +\theta_7 \text{nh3\_kg}_i \text{n2o\_t}_i + \theta_8 \text{nh3\_kg}_i \mathbb{1}_{\{TypeEPCI_i = CC\}} + \theta_9 \text{nh3\_kg}_i \mathbb{1}_{\{TypeEPCI_i = CU\}} \\ \qquad\qquad +\theta_{10} \text{nh3\_kg}_i \mathbb{1}_{\{TypeEPCI_i = Metropole\}} + \varepsilon_i \\ \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

On va maintenant comparer l'ajustement du modèle sélectionné (modselect_bic) avec celui du modèle initial (ancov) pour déterminer si la différence dans leur performance est significative.

On obtient une pvaleur de $0.016 < 0.05$ ce qui indique que l'on peut pas simplifier le modèle avec le modèle sélectionné avec le critère bic.

3.2 Modèle linéaire généralisé

Nous allons maintenant modéliser le dépassement d'émission de méthane de 1000 tonnes par an en fonction en fonction de l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année. On exprime une variable binaire donc le modèle à utiliser est une régression logistique.

$\forall i \in \{1, \ldots, n\}$:

- dep_i : variable binaire valant 1 si le taux d'émission de méthane dépasse les 1000 tonnes par an, et 0 sinon.
- $nh3kg_i$: taux d'émission d'ammoniac en kg/hab
- $n2ot_i$: taux d'émission de protoxyde d'azote en kg/hab
- $TypeEPCI_i$: type d'EPCI
- $annee_i$: année
- $T = \{CC, CA, CU, Metropole\}$: ensemble des types d'EPCI
- $A = \{2015, 2016, 2017, 2018, 2019\}$: ensemble des années

On modélise la probabilité de dépassement de 1000 tonnes par an par le modèle suivant :

$$\text{(Mod4)} : \begin{cases} dep_i \sim \mathcal{B}(\pi_i) \\ \\ \pi_i = \theta_0 + \theta_1 nh3kg_i + \theta_2 n2ot_i + \sum_{j \in T} \beta_j \mathbb{1}_{\{TypeEPCI_i = j\}} + \sum_{a \in A} \alpha_a \mathbb{1}_{\{annee_i = a\}} + interactions \end{cases}$$

```
## Analysis of Deviance Table
##
## Model 1: ch4_t ~ nh3_kg + n2o_t + TypeEPCI + annee_inv
## Model 2: ch4_t ~ (nh3_kg + n2o_t + TypeEPCI + annee_inv)^2
```

```
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       977     394.02
## 2       965     329.88 12   64.141 3.928e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nous avons essayé de simplifier ce modèle, en enlevant les intéractions, mais nous avons rejeté l'hypothèse car nous obtenions un p-valeur trop petite. Nous avons également essayé de mettre en place une méthode backward pour trouver un sous-modèle acceptable, mais encore une fois nous avons obtenu une p-valeur trop petite, et nous avons rejeté le sous modèle. Ce modèle ne semble donc pas pouvoir se simplifier, et nous allons tester son efficaité en faisant de la prédiction. Nous prenons 70% de l'échantillon pour faire le modèle, et nous testons sur les 30% restants.

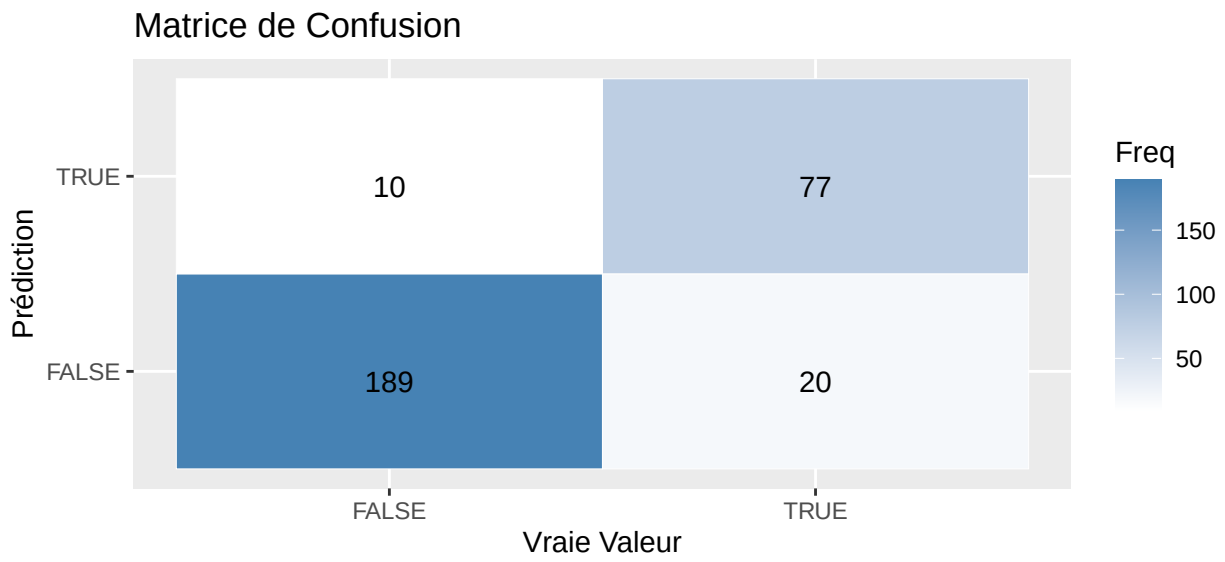


Figure 26: Prédiction sur le taux de méthane

Nous obtenons la figure 26. Ce résultat a un taux de précision de 0.899. C'est très correct, car avec la LDA nous obtenions un taux de précision de 0.943. Ainsi, en ne gardant que certaines variables, nous obtenons un score plutôt proche. On en déduit que l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année explique bien le dépassement d'émission de méthane de 1000 t par an.

4 Conclusion

Ce projet nous a donc permis d'obtenir différents résultats d'analyse de données, afin de comprendre la répartition des données et leurs éventuelles corrélations.

Nous avons vu, par le biais de différentes méthodes de clustering et de certains critères de sélection, que nos données semblaient pouvoir se diviser en deux clusters. Bien que la méthode des k-means en observant l'inertie intraclasses suggère cinq clusters.

Les résultats obtenus sur les différentes régressions réalisées nous ont permis de comprendre que simplifier les modèles était assez dur, car le gaz à effet de serre dépend plus ou moins de toutes les variables.

Ce projet fut l'occasion d'implémenter différentes méthodes et de comparer les résultats obtenus avec chacune de ces méthodes. Au final, elles ne mènent pas toutes aux mêmes conclusions, il a donc été important d'analyser les résultats obtenus à chaque fois afin d'arriver aux bonnes conclusions.