Projet d'étude de Statistiques

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

2024-02-02

Contents

List of Figures

1 Introduction

Le but de ce projet est d'étudier différents polluants mesurés par de nombreux EPCI d'Occitanie.
Nous disposons du jeu de données suivant : `Data-projetmodIA-2324.csv`.

Dans la suite de ce rapport, on utilise les notations suivantes :

- a

2 Analyse descriptive des données

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

2.1 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.

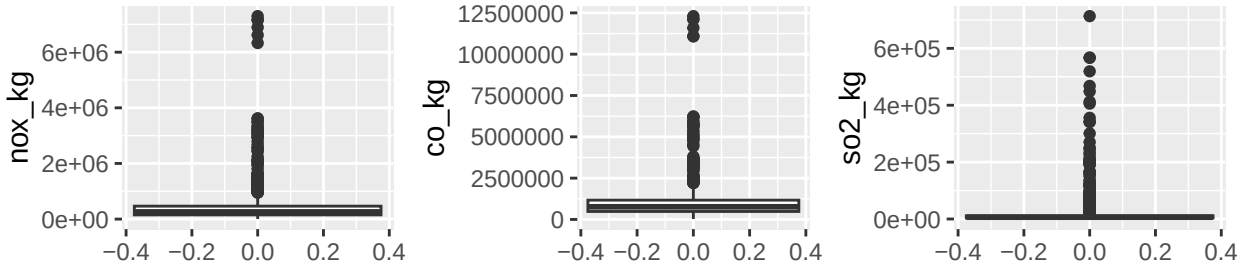


Figure 1: Boxplot des variables nox_kg,co_kg,so2_kg

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 composante principales), nos résultats seront biaisés par la variance et l'asymétrie des données. On transforme donc les données, comme présenté à la figure suivante.

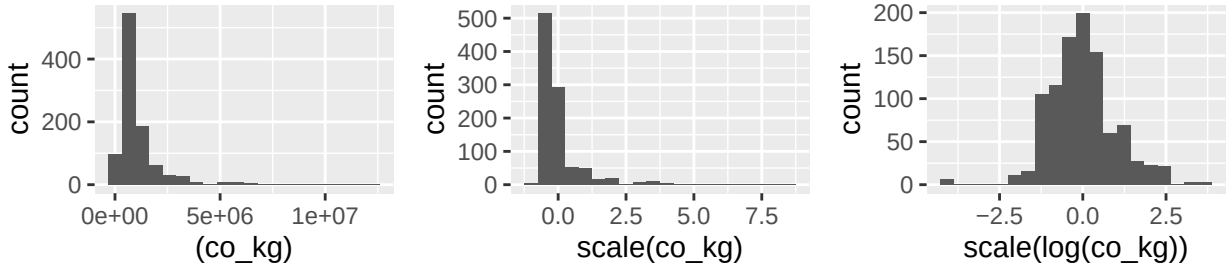


Figure 2: Histogramme de la variable co_kg en brute, scale et scale(log())

La transformation la plus adaptée est la transformation `scale(log())` : Elle 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.

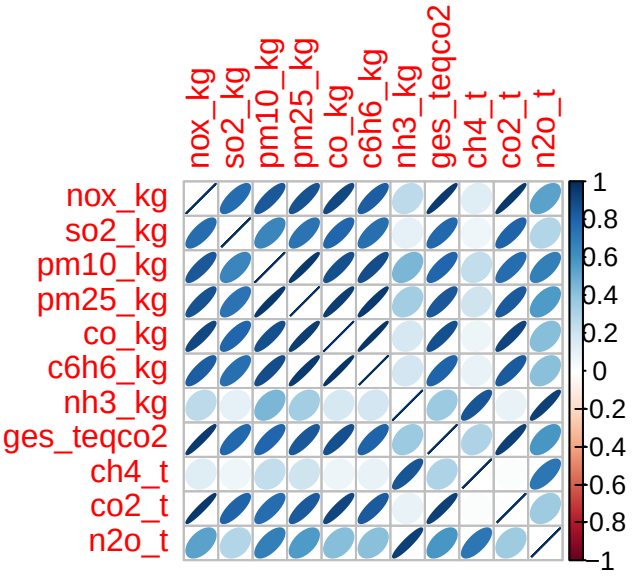


Figure 3: Corrélation entre les variables

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.

2.2 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).

2.2.1 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.

Le premier axe est une combinaison linéaire de... Le deuxième axe est une combinaison linéaire de...

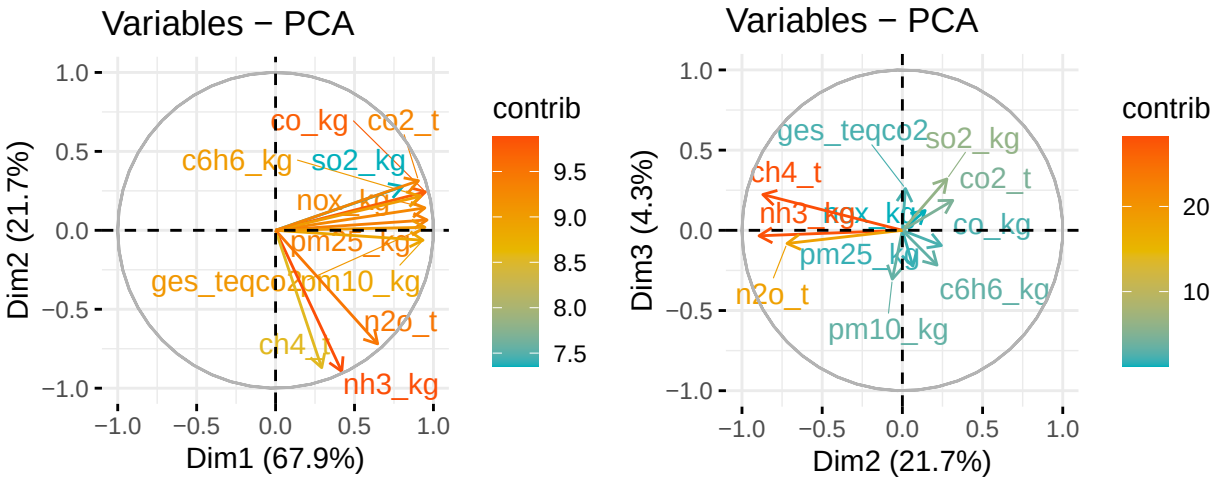


Figure 4: Cercle des corrélations

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

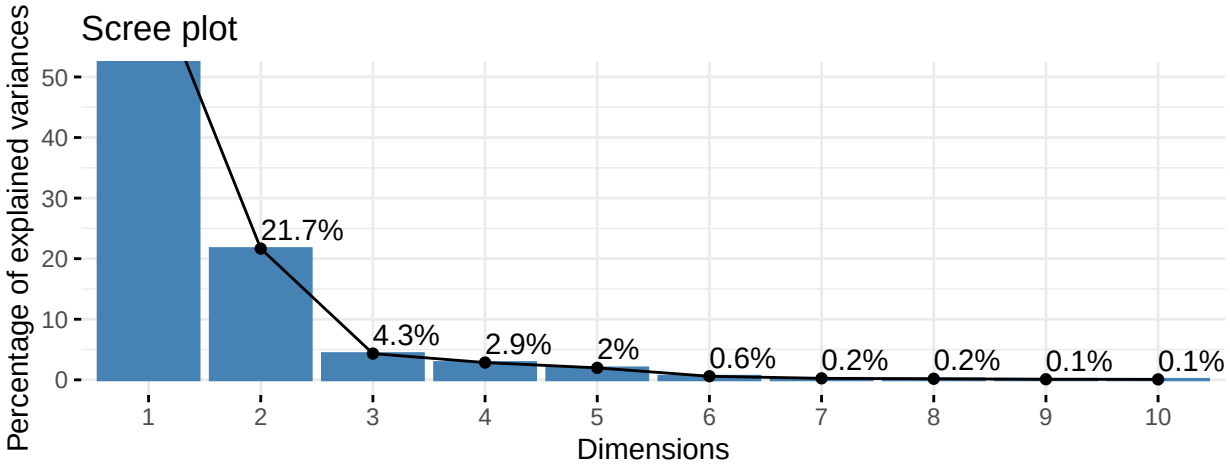


Figure 5: Pourcentage de variance expliquée par chaque axe

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

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

On observe sur la figure 6 que . . .

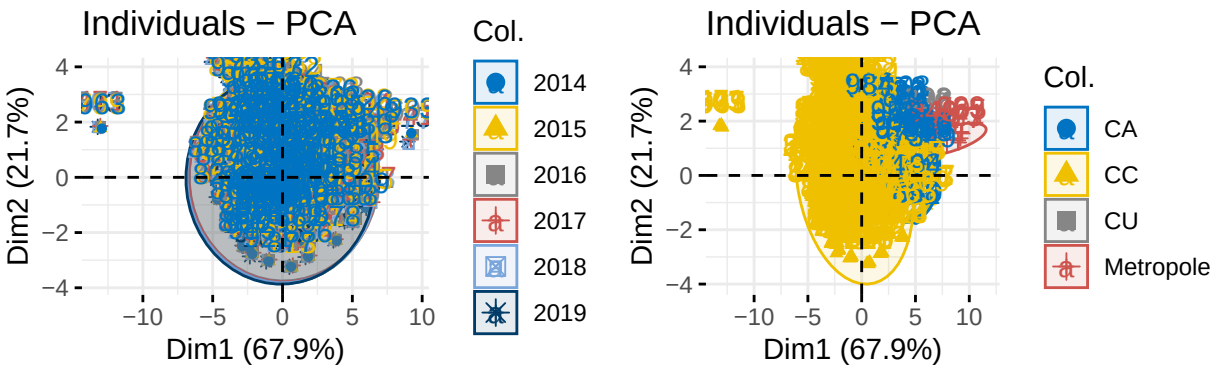


Figure 6: ACP des variables quantitatives

2.2.2 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'intervals qui formeront les modalités des nouvelles variables qualitatives.
Parler des intervalles de discrétisation

Nous allons aussi retirer les valeurs aberrantes c'est-à-dire en-dehors des quantiles (voir boxplot) : 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.

```
## Warning: ggrepel: 34 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

```
## Warning: ggrepel: 46 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

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.

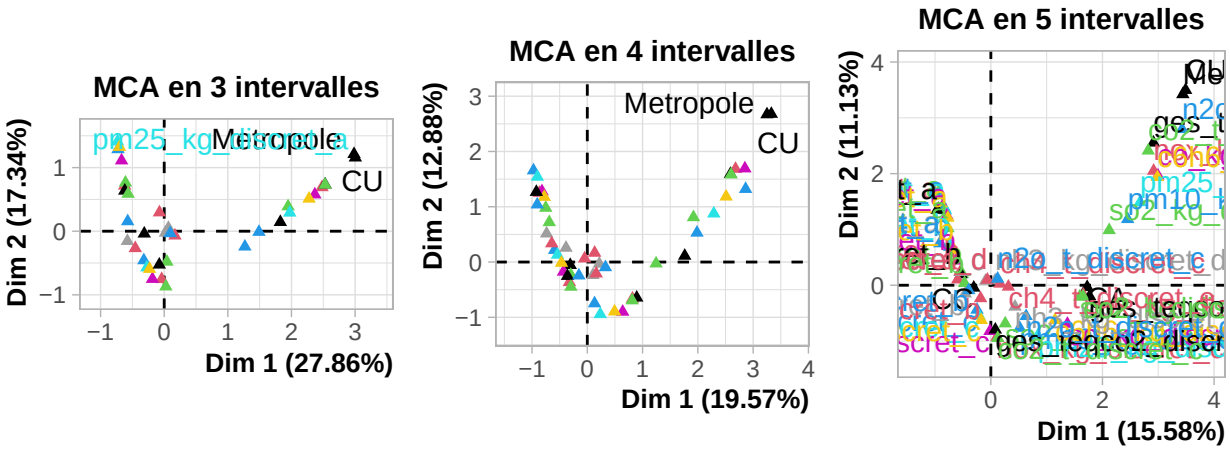


Figure 7: MCA avec découpage des données en 3, 4 et 5 intervalles

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.

3 Classification des EPCI

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

3.1 Clustering

On met en place différents algorithmes de clustering :

3.1.1 Méthodes des k-means

Explication de la méthode des k-means

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

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

3.1.2 Critère de sélection Silhouette

Explication du critère de sélection Silhouette

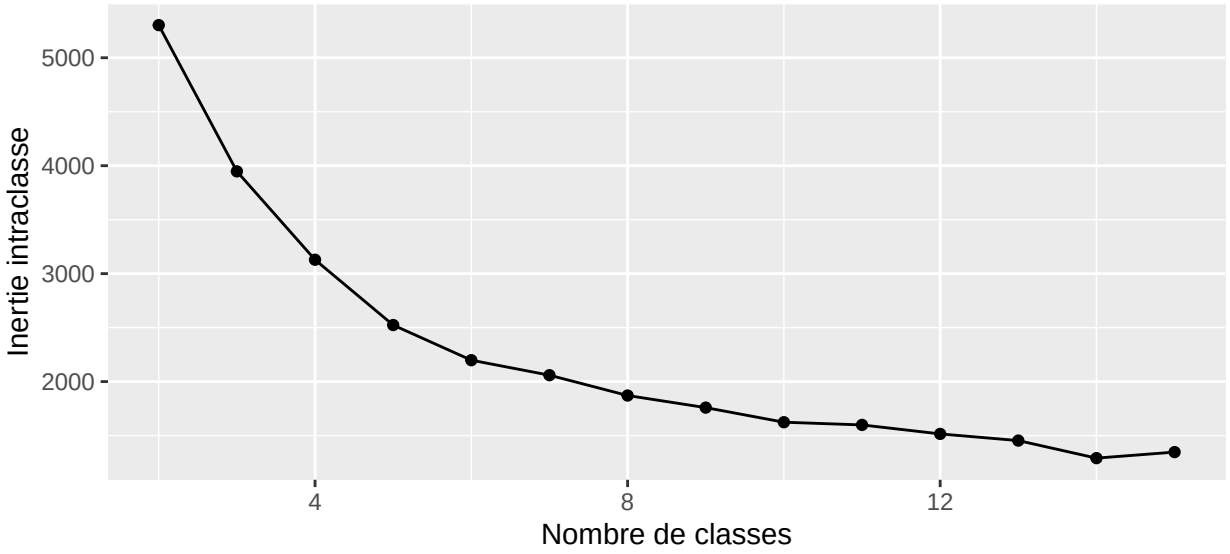


Figure 8: Determination du nombre de clusters optimal

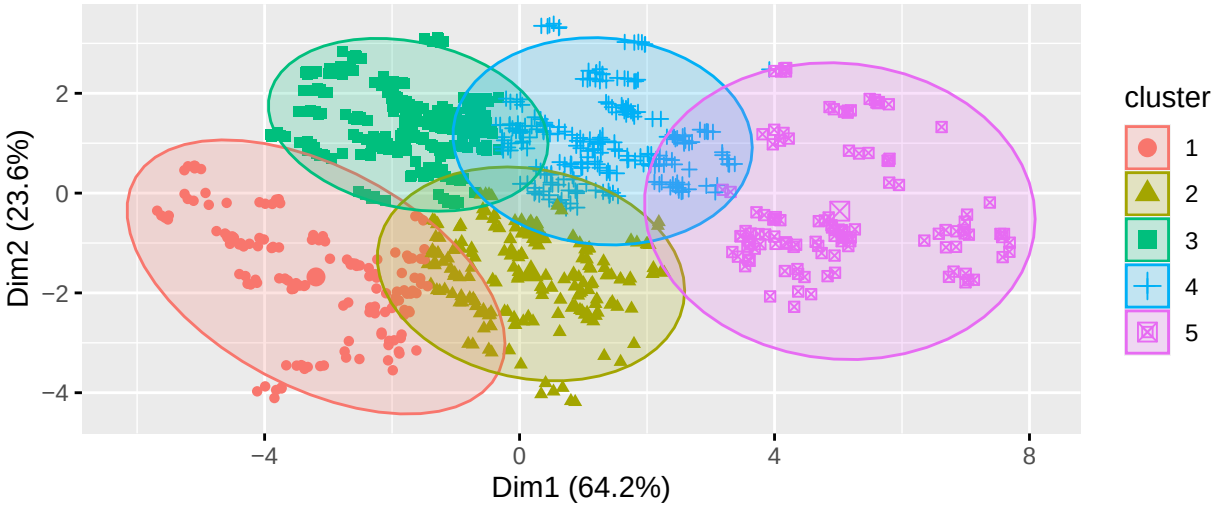


Figure 9: K-means avec K=5

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

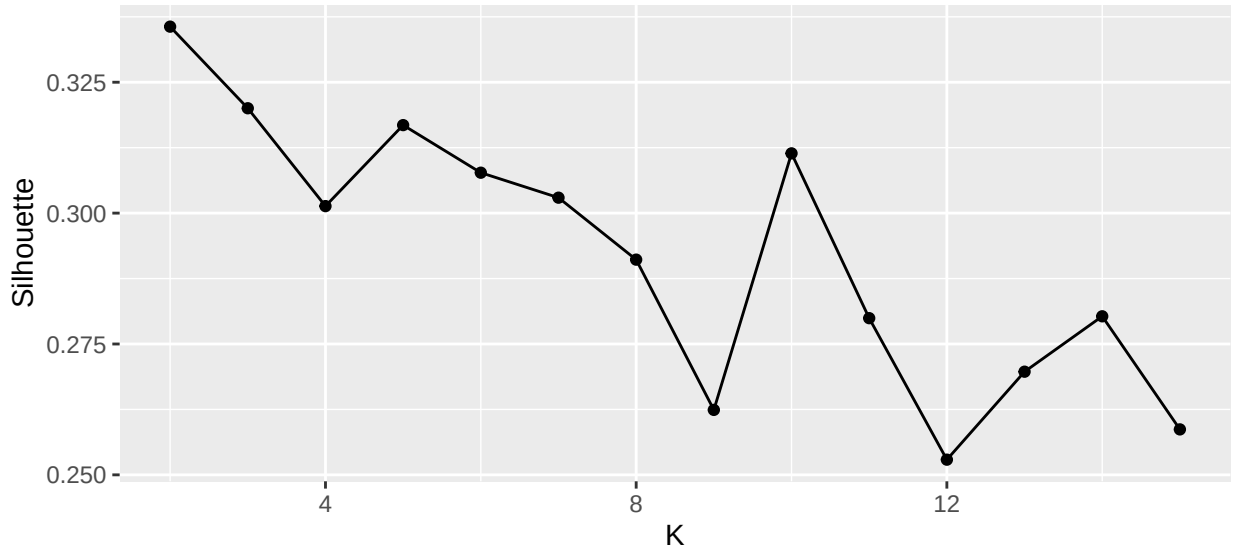


Figure 10: Critère de sélection Silhouette

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 peut se concentrer 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'intérieu 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
```

Au vu des des silhouettes du graphe, il y a une bonne répartition des clusters.

3.1.3 Mélanges Gaussiens

Explication de la méthode des mélanges Gaussiens

3.1.3.1 Critère de sélection BIC Pour commencer, avec un premier affichage en faisant varier le nombre de composantes de 2 à 50, nous n'obtenons pas de résultats satisfaisants, pour pouvoir déterminer le mélange qui s'ajuste le mieux aux données. Cependant, le graphe ci-dessous va nous permettre d'écarter tous les modèles sphériques et diagonaux, qui fournissent des résultats bien moins bons que les autres modèles :

```
#Classification HIérarchique CAH

Data_CAH <- data_scaled_df
Data_CAH <- enlever_donnee_aber(Data_CAH, colnames(Data_CAH))
```

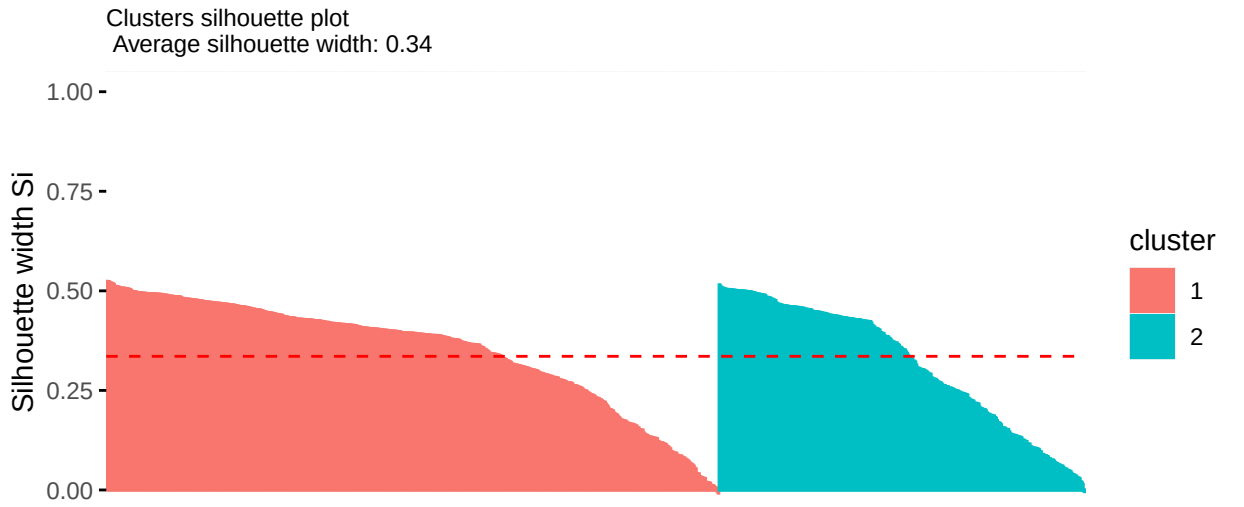


Figure 11: Silhouette avec K=2

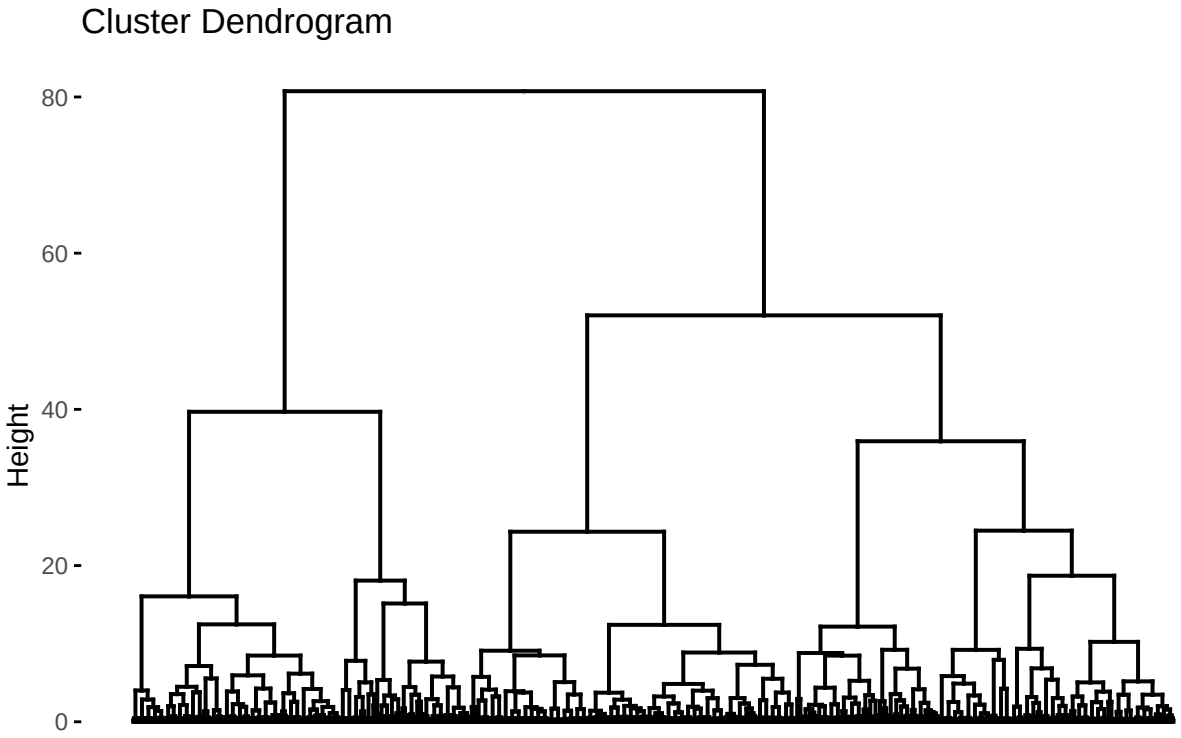
```
d=dist(x = Data_CAH ,method = "euclidian")
#hclustsingle<-hclust(d,method = "single")
#hclustcomplete<-hclust(d,method = "complete")
#hclustaverage<-hclust(d,method = "average")

#fviz_dend(hclustsingle,show_labels=FALSE)
#fviz_dend(hclustcomplete,show_labels=FALSE)
#fviz_dend(hclustaverage,show_labels=FALSE)

#single = dendogramme en escalier, c'est mauvais -> tendance à l'agrégation
#complete = dendogramme équilibré
#average = dendogramme plutôt équilibré mais à tendance

hward<-hclust(d,method = "ward.D2")
fviz_dend(hward,show_labels=FALSE)
```

```
## 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.
```




```
#hward forme des ecaliers ici et tend à l'agrégation
```

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'arrêts pour le nombre optimal de clusters. Nous allons donc utiliser le critère ICL sur les coordonnées de l'ACP.

3.1.4 Critère de sélection ICL

Comme nous l'avons vu précédemment, le modèle qui s'impose le plus est VEV. Nous allons donc nous concentrer sur ce modèle pour le critère ICL. IL est important de noter que le critère ICL n'arrive pas non plus à trouver un critère d'arrêt. Nous avons donc décidé d'afficher l'allure du graphe de VEV pour trouver un point où la progression de la courbe commence à stagner :

On remarque que la courbe commence à stagner à partir de K = 50, ce qui n'est pas un nombre de clusters satisfaisant.

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## i Please use `gather()` instead.
## 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.
```

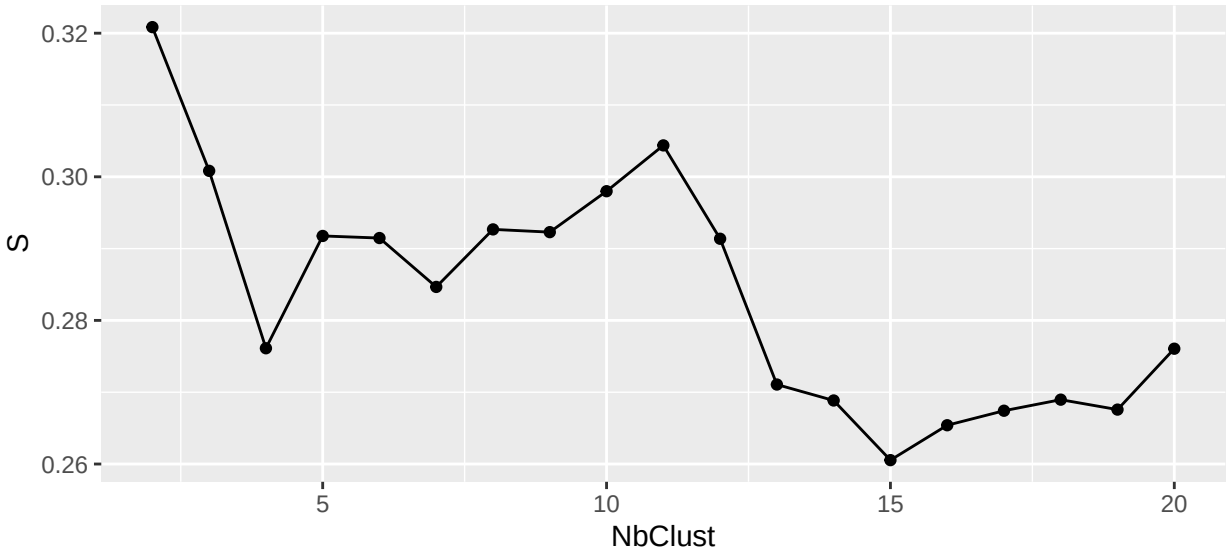


Figure 12: Critère de sélection ICL

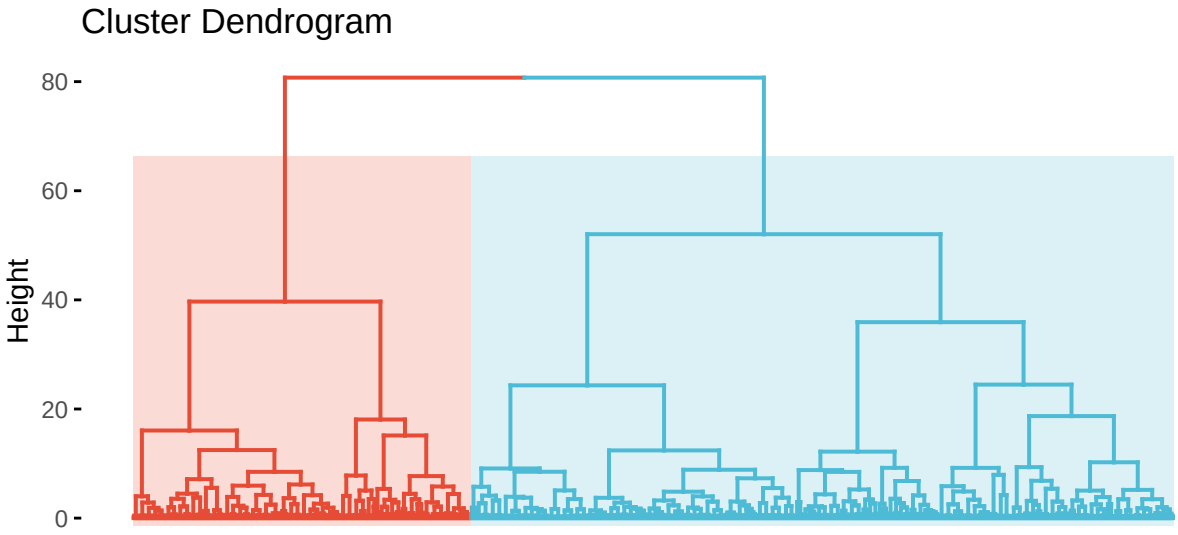


Figure 13: Mélanges Gaussiens avec K=5

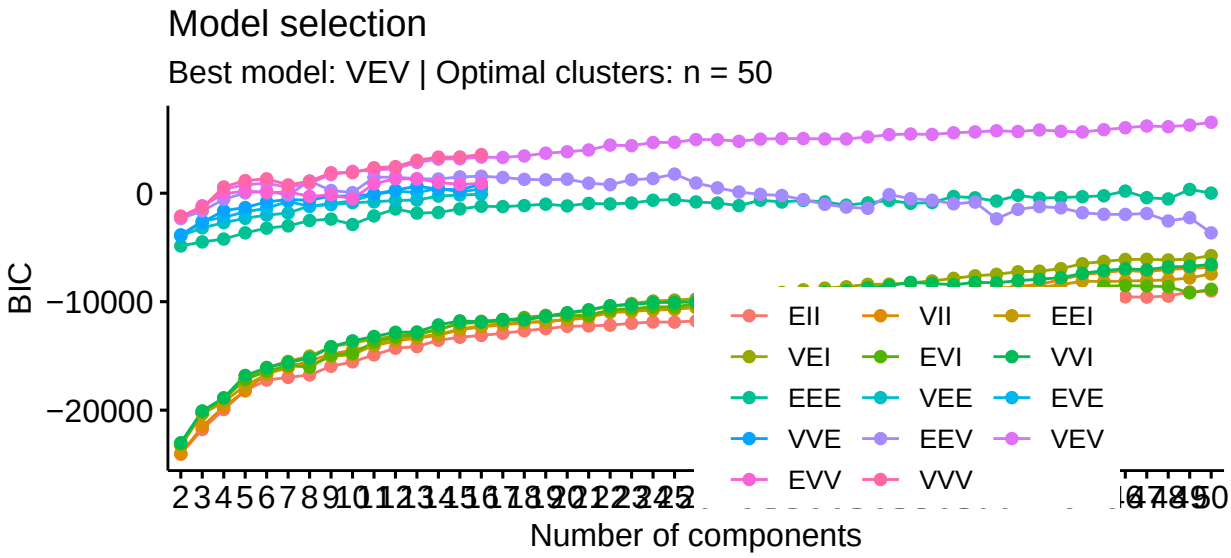


Figure 14: Mélanges Gaussiens avec K=5

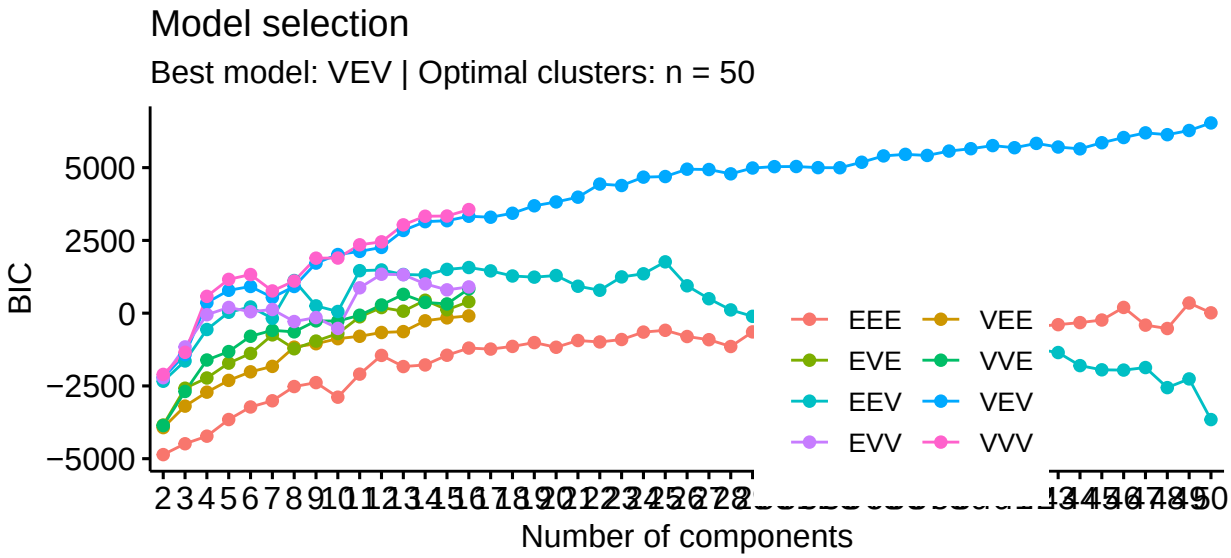


Figure 15: Mélanges Gaussiens avec K=5

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 50 components:
##
##  log-likelihood   n   df      BIC      ICL
##        14970.93 959 3409 6536.034 6536.032
##
## Clustering table:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  6 24 28 48 18  6 42 42 24 18 25 36 41 55  6  6 11 30 60 18 23 41  6  6 11  6
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
## 24 12  6 12 12  6  6 12 17 32 12 17 12 11 18 12 12  6 25 12 18 12  6 10

## Best ICL values:
##           VEV,100    VEV,44    VEV,98
## ICL      5415.963 5266.6529 5244.2055
## ICL diff    0.000 -149.3105 -171.7579
```

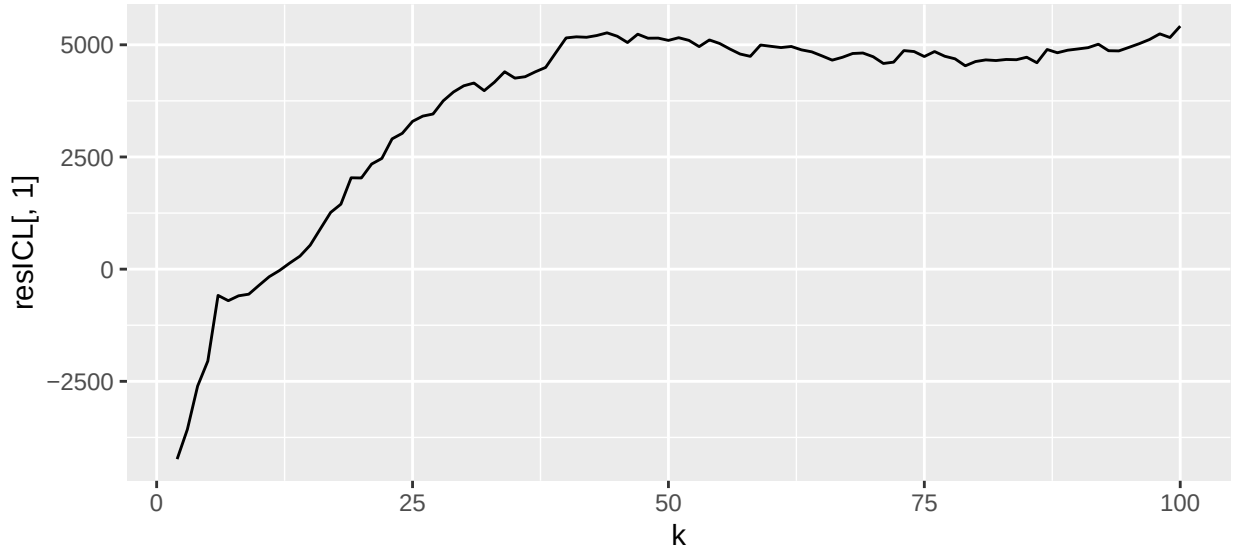


Figure 16: Mélanges Gaussiens avec K=5

3.1.4.1 Dendogrammes A part pour complete, toutes les autres méthodes et silhouettes trouvent deux clusters : on utilise donc hward (à justifier), et on obtient :

3.1.4.2 Leafleet A changer !!!

```
plotmapquali(pca_coordinates,resBIC_VEV$classification)
```

3.2 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.

3.2.1 Taux d'émission de méthane

Dans notra 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 17.

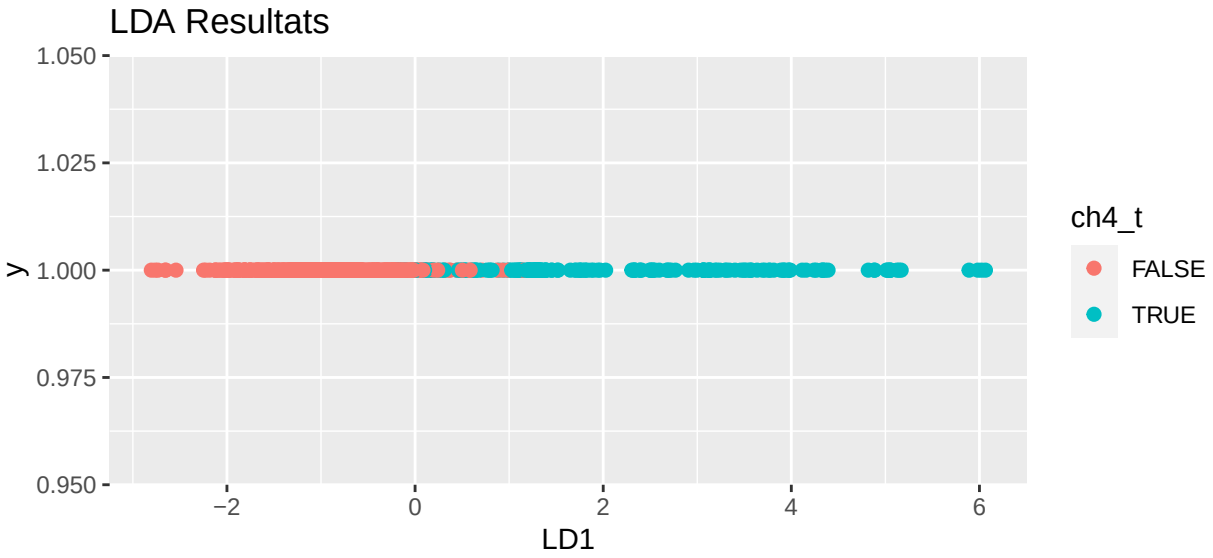


Figure 17: LDA sur le taux de méthane

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 un 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 :

Nous pouvons voir grâce à cette table que les individus sont plutôt bien prédits. En effet, on obtient un taux de précision de 0.936. 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.

3.2.2 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 3. 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.

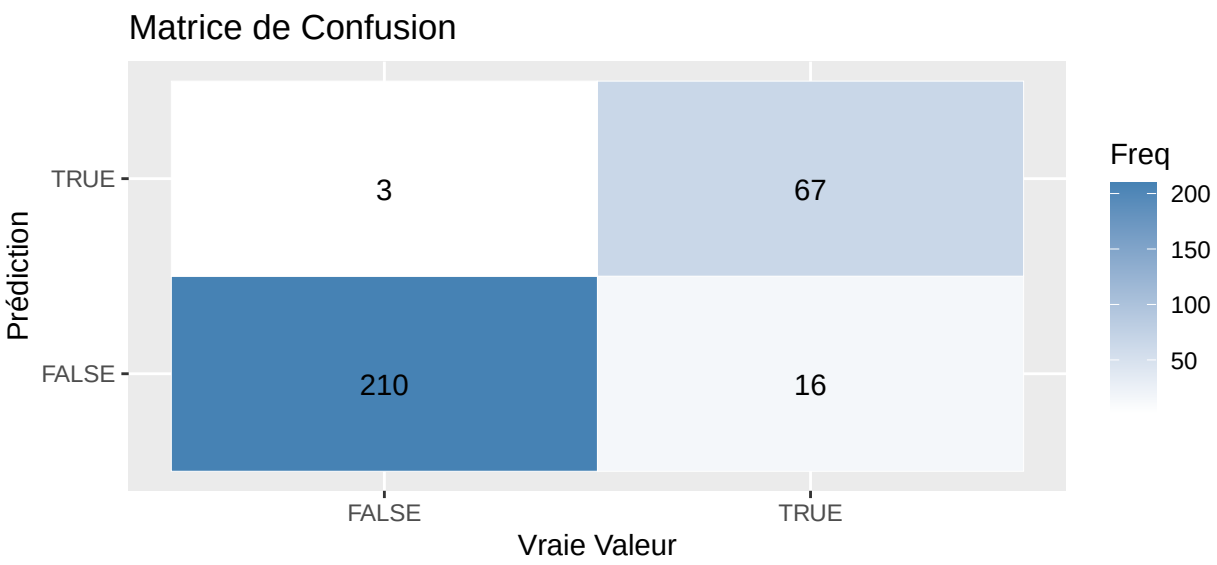


Figure 18: Prédiction sur le taux de méthane

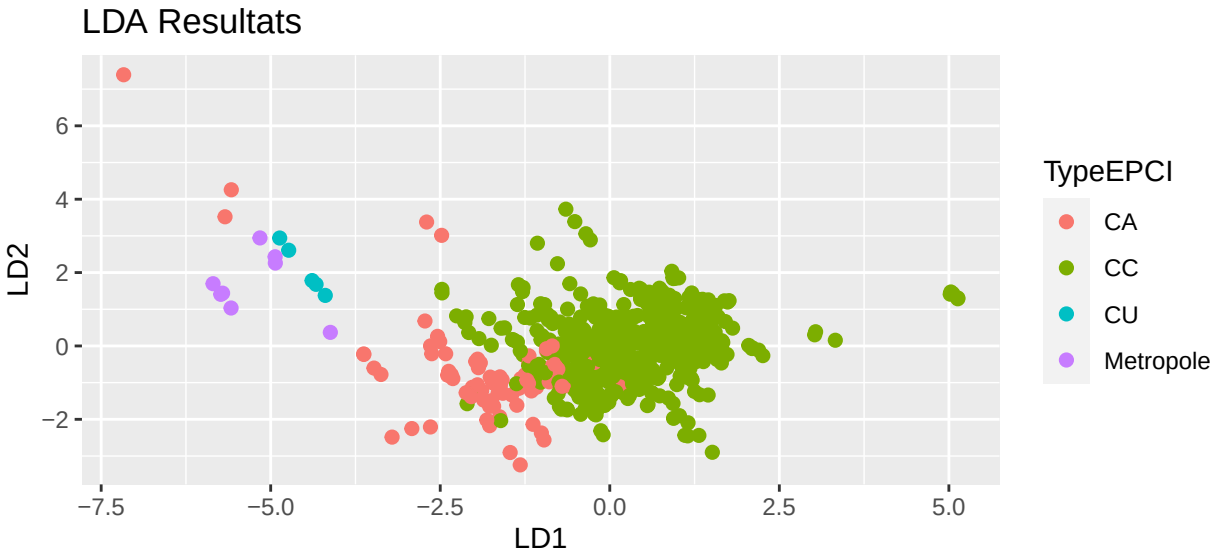


Figure 19: LDA en fonction des types EPCI

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.

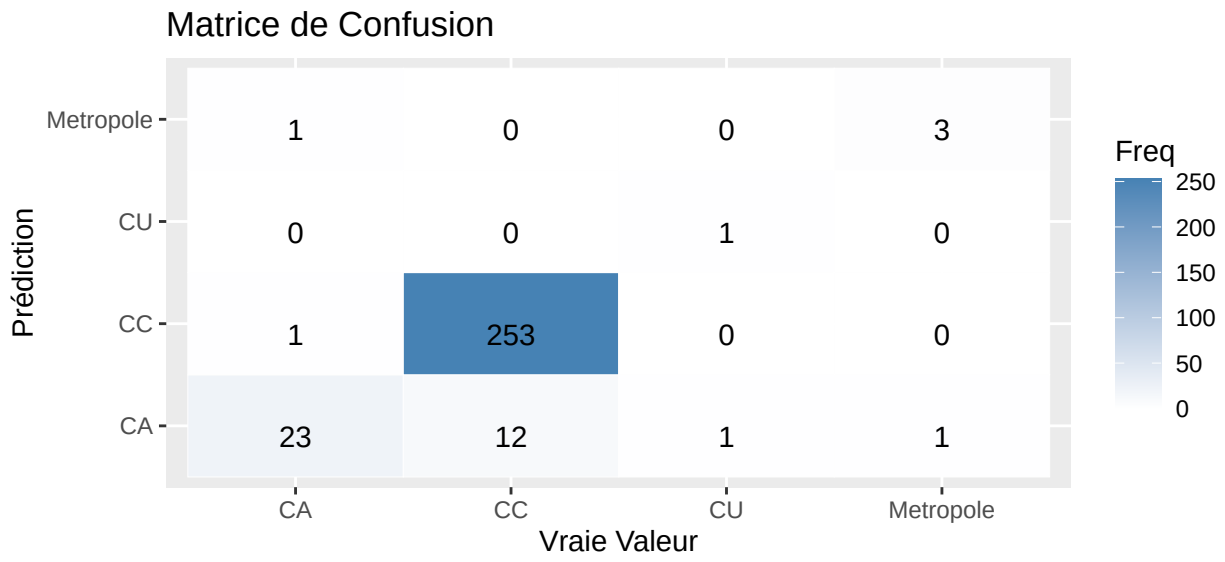


Figure 20: Prédiction sur le type d'EPCI

Nous pouvons voir grâce à la figure 20 table que les individus sont plutôt bien prédits. On obtient un taux de précision de 0.946. Ainsi, le type d'EPCI différencie bien les individus, 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"

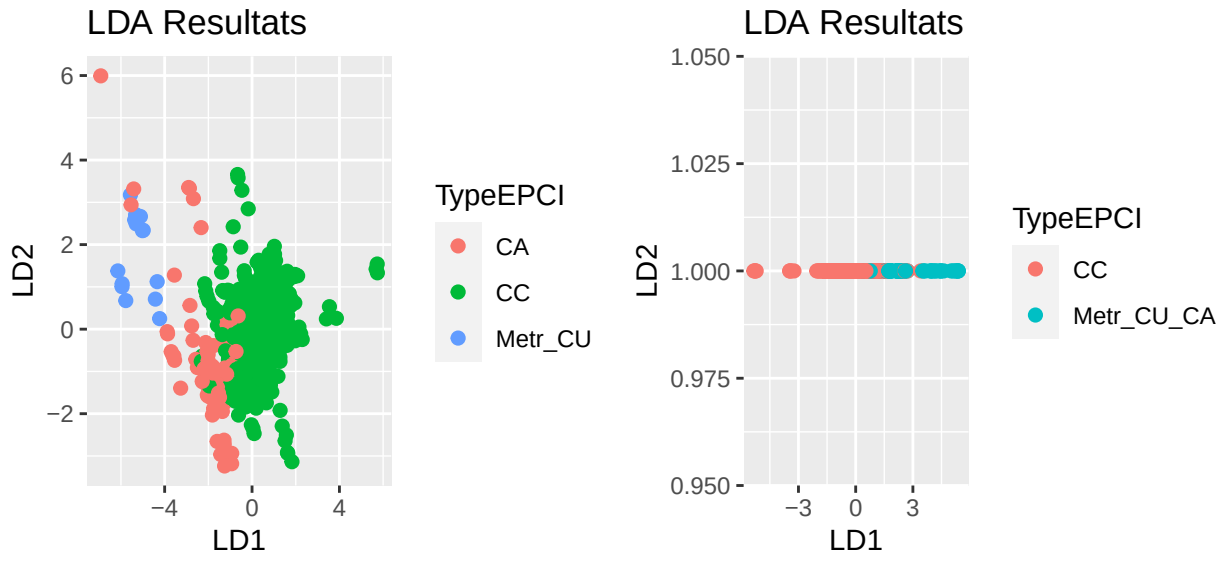


Figure 21: LDA en fonction des types EPCI

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.

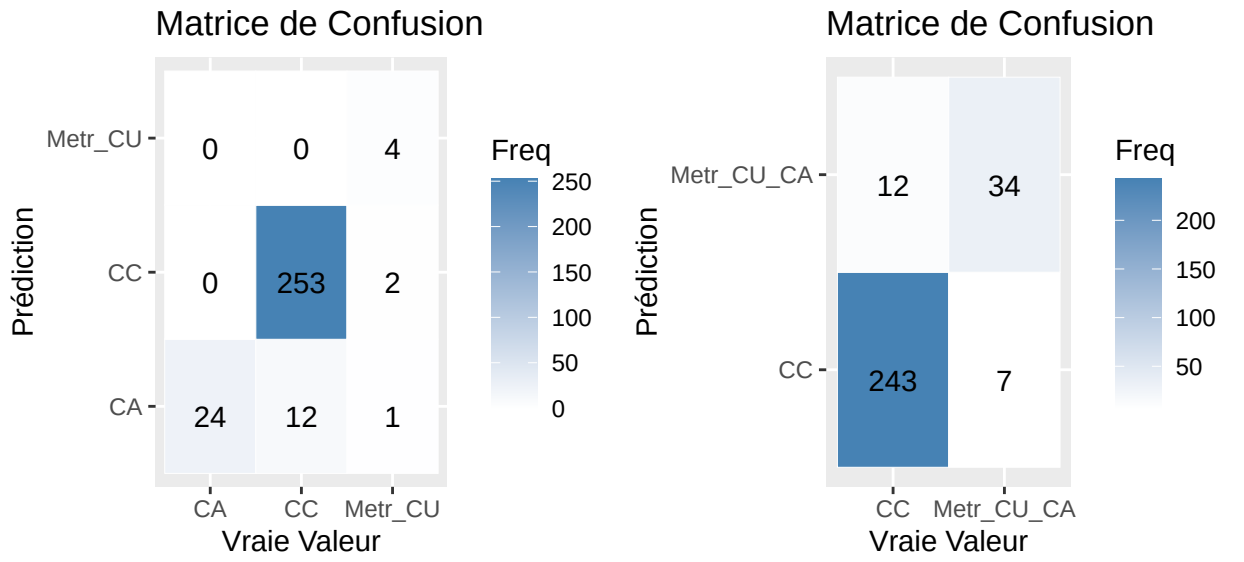


Figure 22: Prédiction en fonction des types EPCI simplifiés

Nous obtenons un taux de précision de 0.949 pour le premier regroupement, et de 0.936 pour le deuxième. Ainsi, contrairement à ce qu'on a pu penser, nous ne gagnons pas 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.

4 EMS

4.1 Modèle linéaire

4.1.1 Modèle d'ANOVA

On explique le gaz à effet de serre en fonction des variables Type et années.

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

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$$

EXPLIQUER LA SIGNIFICATION DES TERMES DU MODELE

```
##
## Call:
## lm(formula = ges_teqco2 ~ TypeEPCI * annee_inv, data = dlog)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2367 -0.4233 -0.0383  0.3863  2.8469
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -18.397236  80.407295  -0.229    0.819
## TypeEPCICC                    4.064479  86.227217   0.047    0.962
## TypeEPCICU                    7.818578 377.143642   0.021    0.983
## TypeEPCIMetropole           -19.949436 272.674402  -0.073    0.942
## annee_inv                     0.009761   0.039875   0.245    0.807
## TypeEPCICC:annee_inv         -0.002779   0.042761  -0.065    0.948
## TypeEPCICU:annee_inv         -0.003354   0.187029  -0.018    0.986
## TypeEPCIMetropole:annee_inv   0.010806   0.135222   0.080    0.936
##
## Residual standard error: 0.7644 on 976 degrees of freedom
## Multiple R-squared:  0.4198, Adjusted R-squared:  0.4157
## F-statistic: 100.9 on 7 and 976 DF,  p-value: < 2.2e-16
```

-> Commentaire sur la valeur de R^2 obtenue.

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

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

Dire pourquoi c'est bien un sous-modèle

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

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 :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

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

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

et

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

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 :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

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

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

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 :

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

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

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, soit un écart moyen de 13.14 %.

4.1.2 Régression linéaire

4.1.3 ANCOVA

4.2 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.

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

- 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 :

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

```
## Warning: glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1
```

```
##
## Call:
## glm(formula = ch4_t ~ .^2, family = binomial(link = "logit"),
##     data = dlog)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  4.309e+03  3.117e+03   1.383 0.166788
## nh3_kg                      -6.202e+03  2.012e+03  -3.082 0.002055 **
## n2o_t                        4.672e+03  1.427e+03   3.274 0.001062 **
## TypeEPCICC                  -1.885e+03  3.084e+03  -0.611 0.541114
## TypeEPCICU                  -1.093e+04  2.968e+07   0.000 0.999706
## TypeEPCIMetropole           -7.385e+04  4.127e+06  -0.018 0.985722
## annee_inv                   -2.157e+00  1.554e+00  -1.388 0.165089
## nh3_kg:n2o_t                -1.483e+00  4.777e-01  -3.105 0.001904 **
## nh3_kg:TypeEPCICC           -1.919e+02  1.388e+02  -1.382 0.166953
## nh3_kg:TypeEPCICU           -2.128e+02  2.114e+05  -0.001 0.999197
## nh3_kg:TypeEPCIMetropole    -2.306e+02  4.875e+03  -0.047 0.962273
## nh3_kg:annee_inv             3.182e+00  9.968e-01   3.192 0.001413 **
## n2o_t:TypeEPCICC             1.516e+02  1.088e+02   1.394 0.163326
## n2o_t:TypeEPCICU             1.567e+02  4.811e+05   0.000 0.999740
## n2o_t:TypeEPCIMetropole     -4.268e+02  3.758e+04  -0.011 0.990938
## n2o_t:annee_inv             -2.399e+00  7.063e-01  -3.396 0.000683 ***
## TypeEPCICC:annee_inv         9.522e-01  1.538e+00   0.619 0.535813
## TypeEPCICU:annee_inv         5.431e+00  1.467e+04   0.000 0.999705
## TypeEPCIMetropole:annee_inv  3.688e+01  2.060e+03   0.018 0.985713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1201.84  on 983  degrees of freedom
## Residual deviance:  329.88  on 965  degrees of freedom
## AIC: 367.88
##
## Number of Fisher Scoring iterations: 20

## Warning: glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1

## 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 un 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.

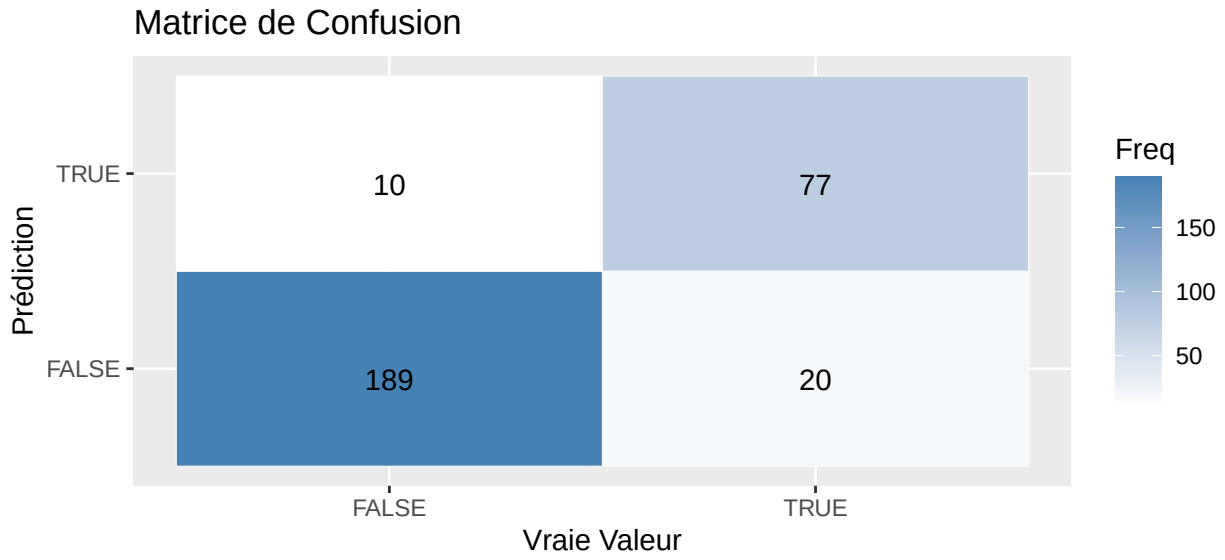


Figure 23: Prédiction sur le taux de méthane

Nous obtenons la figure 23. 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 0.936. 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.

5 Conclusion