

Projet d'étude de Statistiques

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

2024-01-29

Contents

1	Introduction	3
2	Analyse descriptive des données	3
2.1	Analyse unidimensionnelle	3
2.2	Analyse multidimensionnelle	4
3	Classification des EPCI	7
3.1	Clustering	7
3.2	Analyse discriminante linéaire	7
4	EMS	11
4.1	Modèle linéaire	11
4.2	Modèle linéaire généralisé	13
5	Conclusion	13

List of Figures

1	Boxplot des variables nox_kg,co_kg,so2_kg	3
2	Histogramme de la variable co_kg en brute, scale et scale(log())	3
3	Corrélation entre les variables	4
4	Cercle des corrélations	5
5	Pourcentage de variance expliquée par chaque axe	5
6	ACP des variables quantitatives	6
7	MCA avec découpage des données en 3, 4 et 5 intervalles	7
8	LDA sur le taux de méthane	8
9	Prédiction sur le taux de méthane	9
10	LDA en fonction des types EPCI	9

11	Prédiction sur le type d'EPCI	10
12	LDA en fonction des types EPCI	10
13	Prédiction en fonction des types EPCI simplifiés	11

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 unidimensionnelle

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.

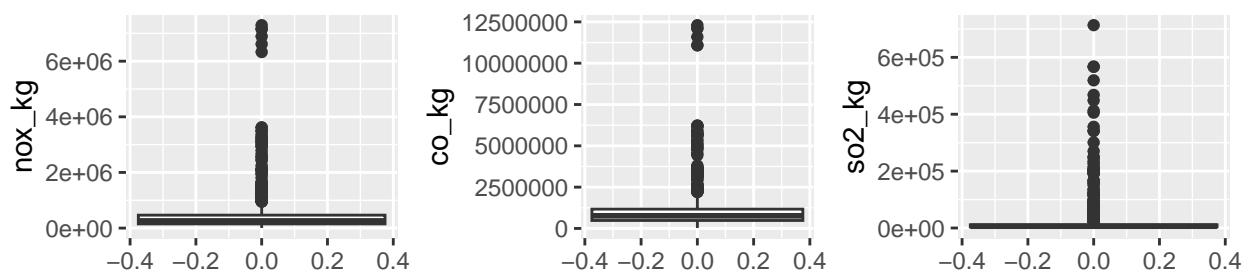


Figure 1: Boxplot des variables `nox_kg`, `co_kg`, `so2_kg`

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.

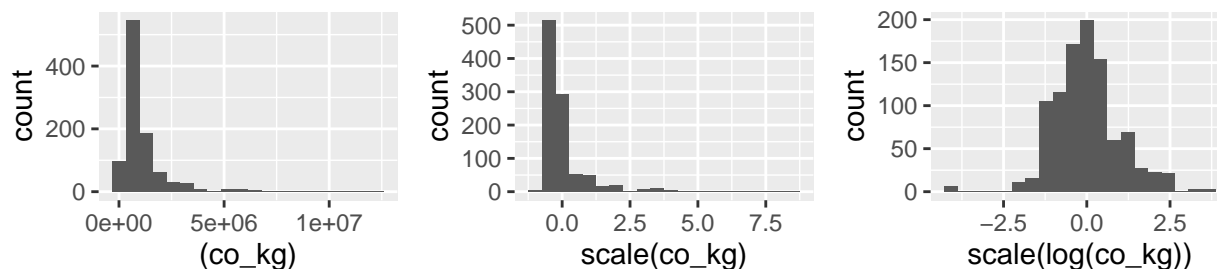


Figure 2: Histogramme de la variable `co_kg` en brute, `scale` et `scale(log())`

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.

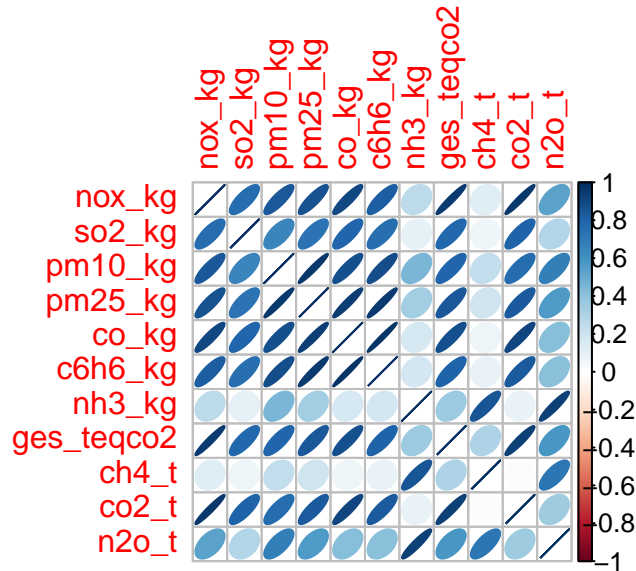


Figure 3: Corrélation entre les variables

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 Composantes Principales (ACP) des variables quantitatives

On s'intéresse 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...

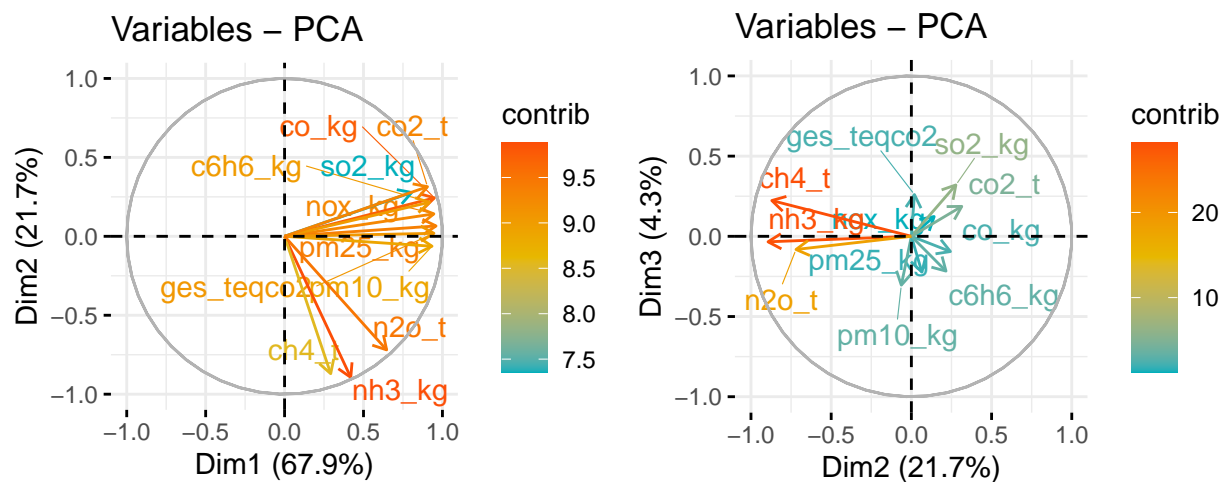


Figure 4: Cercle des corrélations

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

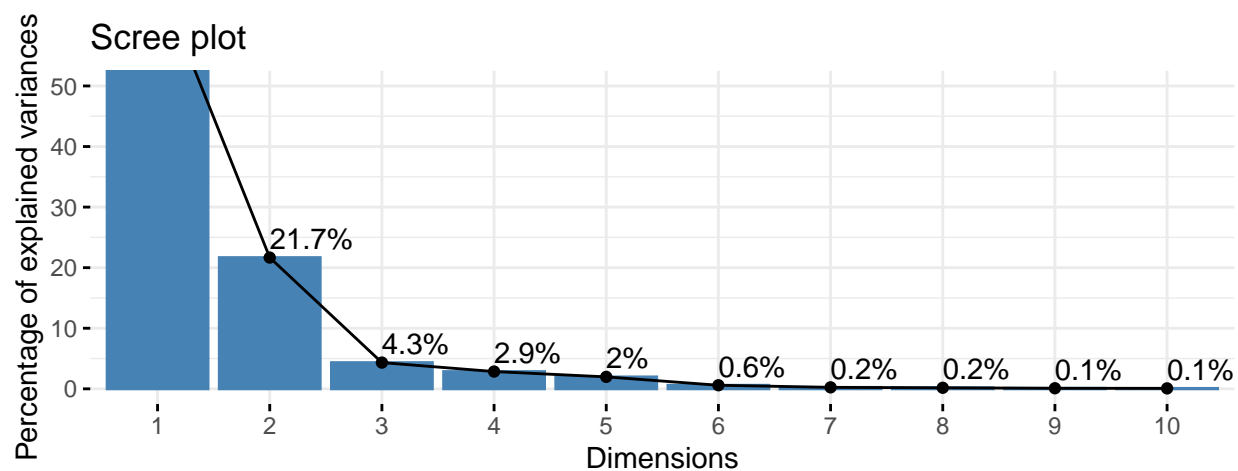


Figure 5: Pourcentage de variance expliquée par chaque axe

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 ...

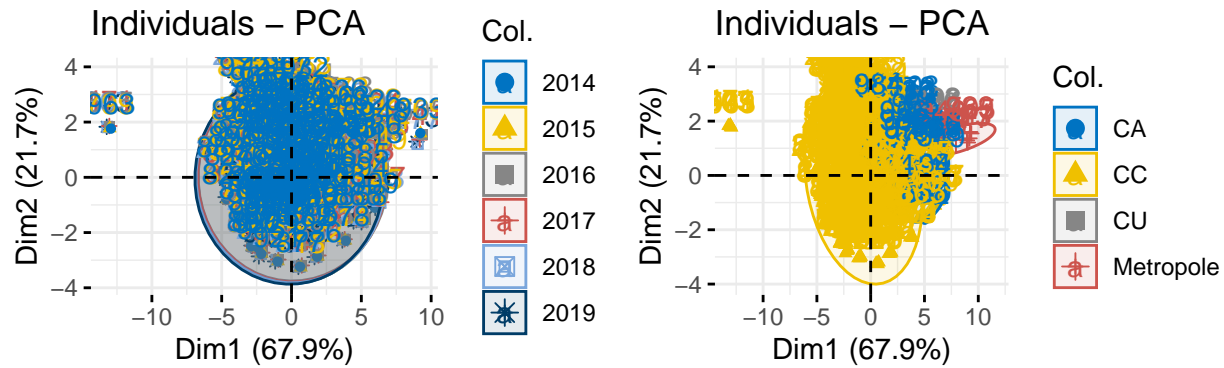


Figure 6: ACP des variables quantitatives

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 quantitatives nous avons donc besoin de discrétiser ces variables. Nous allons former un nombre fini d'intervalles 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.

3.2.1 Taux d'émission de méthane

Dans notre cas, nous créons 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 8.

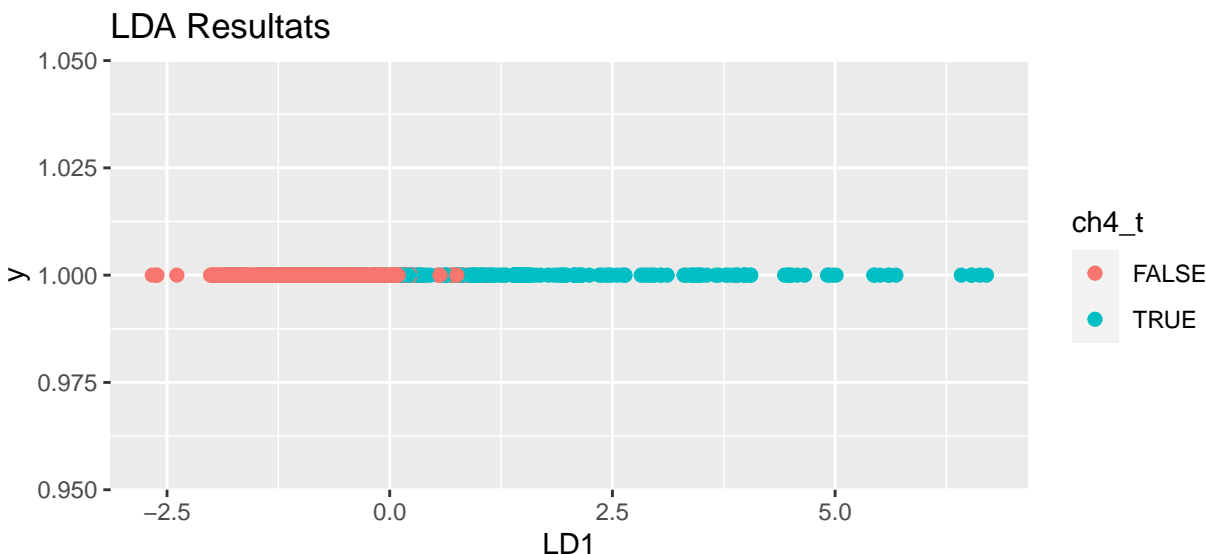


Figure 8: LDA sur le taux de méthane

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 :

Nous pouvons voir grâce à cette table que les individus sont plutôt bien prédits. En effet, on obtient un taux de précision de 0.929. 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 à 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 affiché dans les deux premières dimensions dans la figure 3, car c'est l'affichage le plus parlant.

Nous pouvons voir que les données semblent bien séparées, chaque type d'EPCI. Le type d'EPCI semble bien séparer 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.

Nous pouvons voir grâce à la figure 11 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, et nous obtenons un bon taux de précision. Cependant, il y a une forte dissimilarité entre les nombres d'individus par modalité.

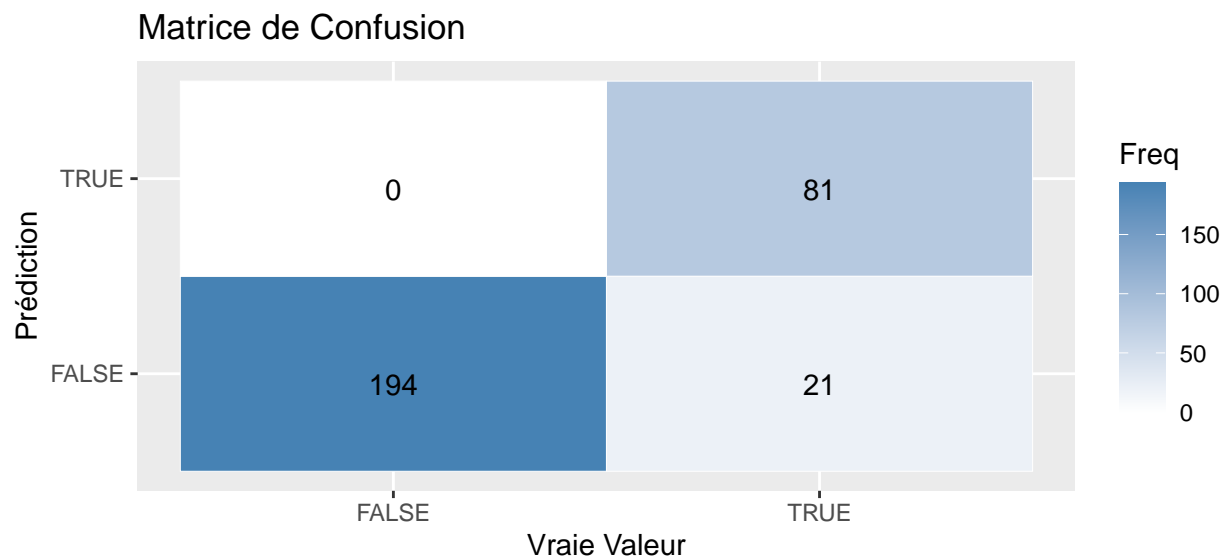


Figure 9: Prédiction sur le taux de méthane

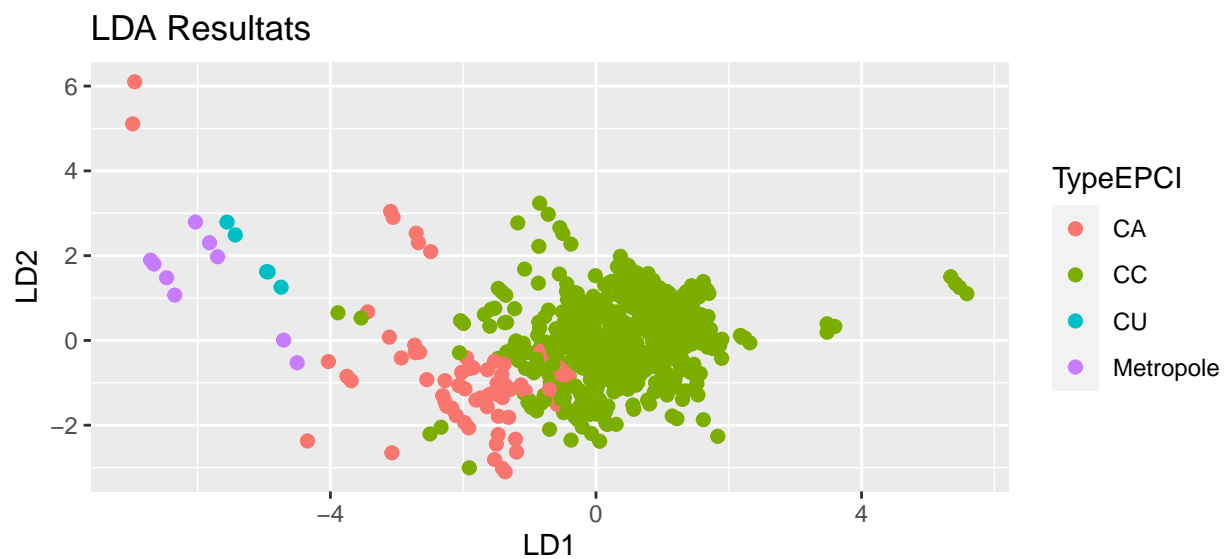


Figure 10: LDA en fonction des types EPCI

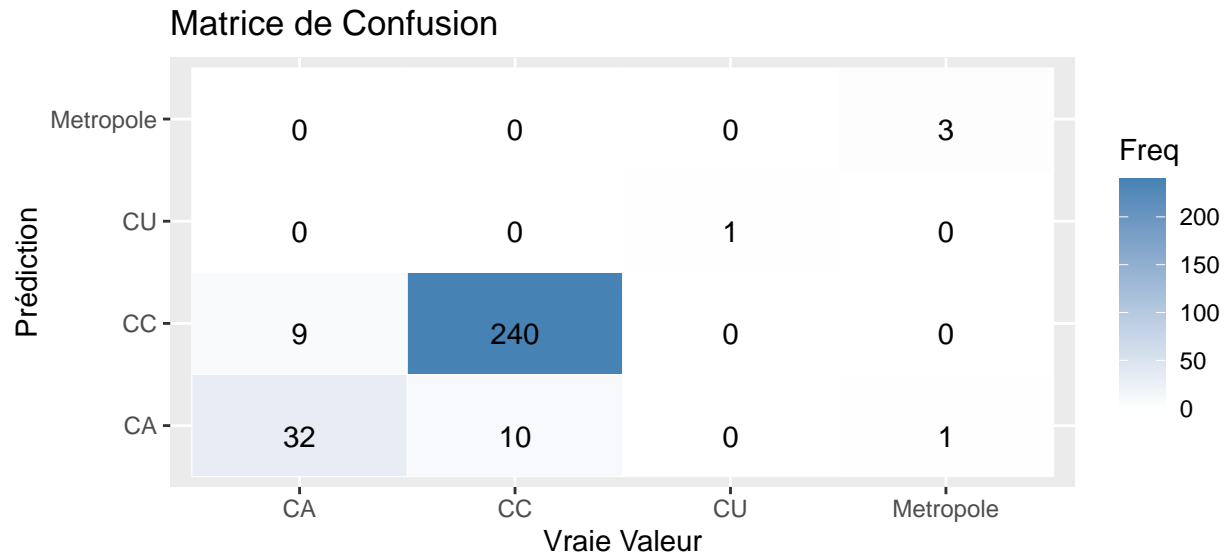


Figure 11: Prédiction sur le type d'EPCI

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”

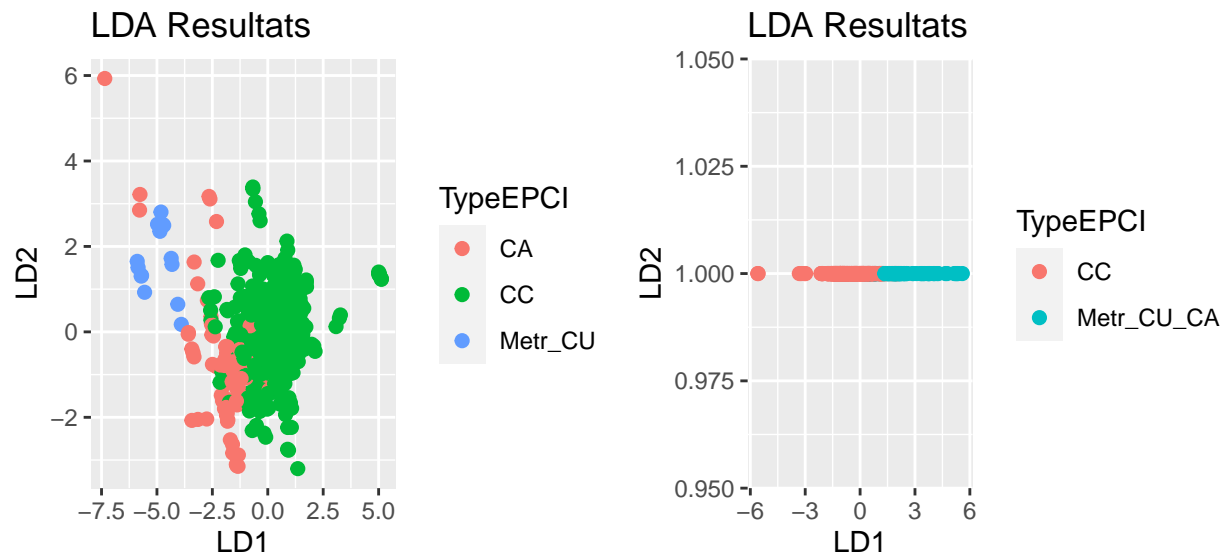


Figure 12: LDA en fonction des types EPCI

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.

Nous obtenons un taux de précision de 0.926 pour le premier regroupement, et de 0.939 pour le deuxième. Ainsi, contrairement à ce qu’on a pu penser, nous ne gagnons pas en précision en faisant des regroupements.

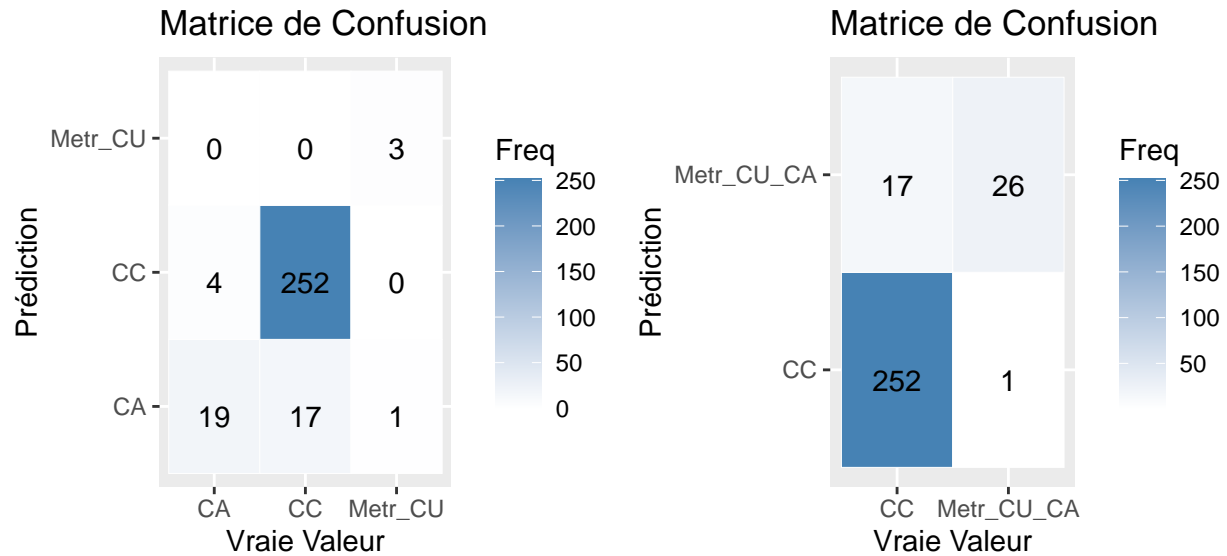


Figure 13: Prédiction en fonction des types EPCI simplifiés

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 agrégé 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 :

$$\begin{aligned}\mathcal{H}_0 : Y_{ij} &= \mu + \alpha_i + \beta_j + \epsilon_{ij} \\ \mathcal{H}_1 : Y_{ij} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}\end{aligned}$$

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 :

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

On essaie de simplifier le modèle en enlevant les variables non significatives (on fait 2 tests de sous-modèle) :

$$\begin{aligned}\mathcal{H}_0 : Y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\ \mathcal{H}_1 : Y_{ij} &= \mu + \alpha_i + \beta_j + \epsilon_{ij}\end{aligned}$$

et

$$\begin{aligned}\mathcal{H}_0 : Y_{ij} &= \mu + \beta_j + \epsilon_{ij} \\ \mathcal{H}_1 : Y_{ij} &= \mu + \alpha_i + \beta_j + \epsilon_{ij}\end{aligned}$$

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 :

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

On essaie à nouveau de simplifier le modèle en enlevant les variables non significatives.

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 :

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

4.1.2 Régression linéaire

4.1.3 ANCOVA

4.2 Modèle linéaire généralisé

5 Conclusion