# Question 5

2023-12-01

## Contents



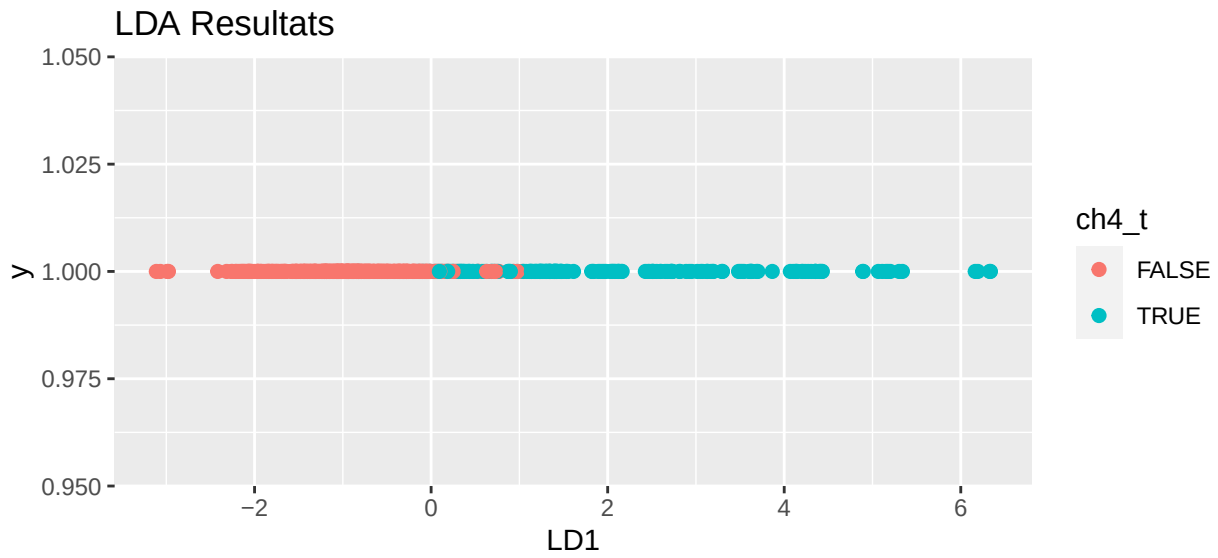
## 0.1 Analyse discriminante linéaire

Nous avons dans la partie précédente effectuer plusieurs types de clustering pour regrouper les données. Le clustering regroupe les individus de manière non supervisée. Dans cette partie, nous allons essayé 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.

### 0.1.1 Taux d'émission de méthane

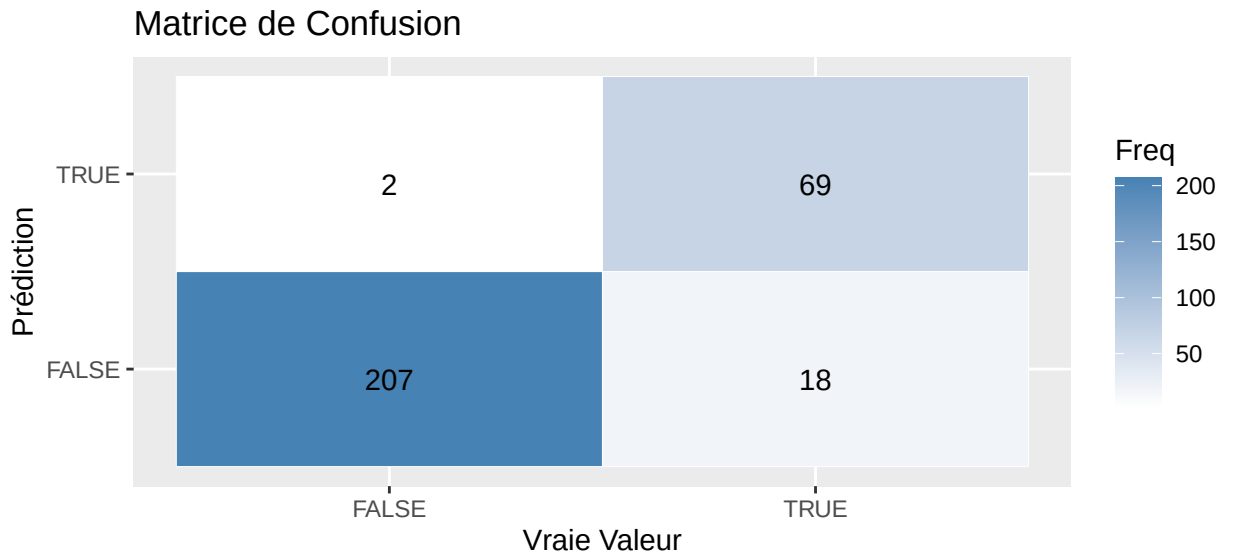
Dans notra cas, nous créeons une nouvelle variable binaire, valant 1 si le taux d'émission de méthane dépasse les 1000 tonnes par an, et 0 sinon. Nous effectuons ensuite une LDA, et nous pouvons visualiser les résultats dans la figure **??**.




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 de la figure **??**




Nous pouvons voir grâce à cette table que les individus sont plutôt bien prédits. Nous pouvons même afficher le taux de précision de cette prédiction à partir de la matrice de confusion : 0.9324324 .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.

### 0.1.2 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.

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.

Nous pouvons voir grâce à la figure 2 table que les individus sont plutôt bien prédits. Affichons le taux de précision de cette prédiction : 0.9087838 .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. Essayons de regrouper les plus petites modalités entre elles afin de voir ce que nous obtenons. Dans la suite, nous regrouperons donc "CU" et "Métropole"; et nous essayerons aussi de regrouper "CU","Métropole", et "CA".

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

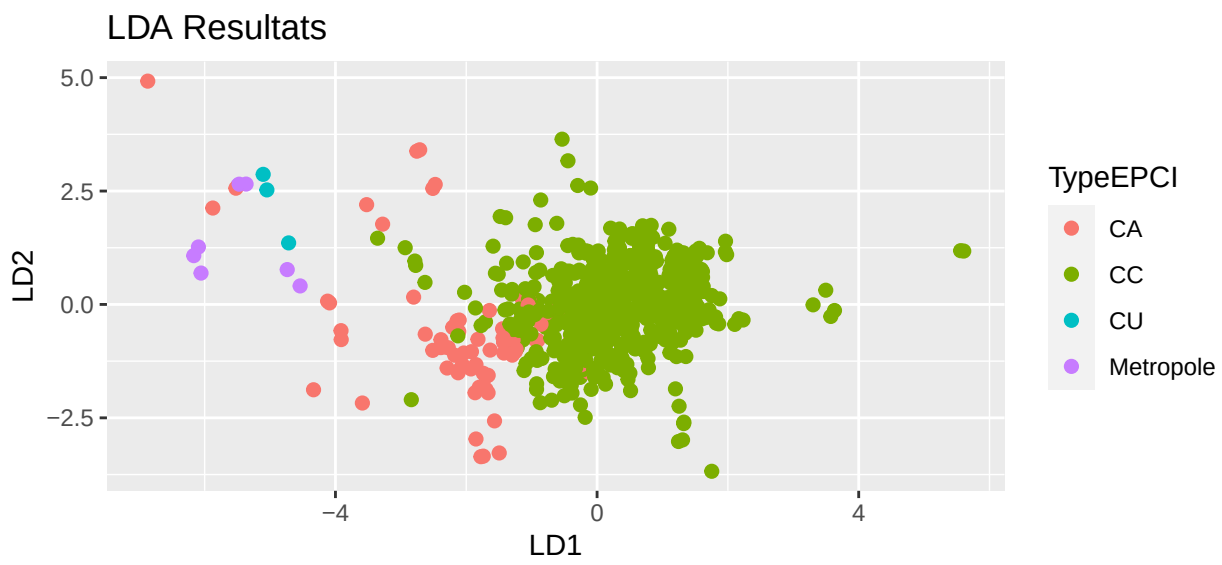



Figure 1: LDA en fonction des types EPCI

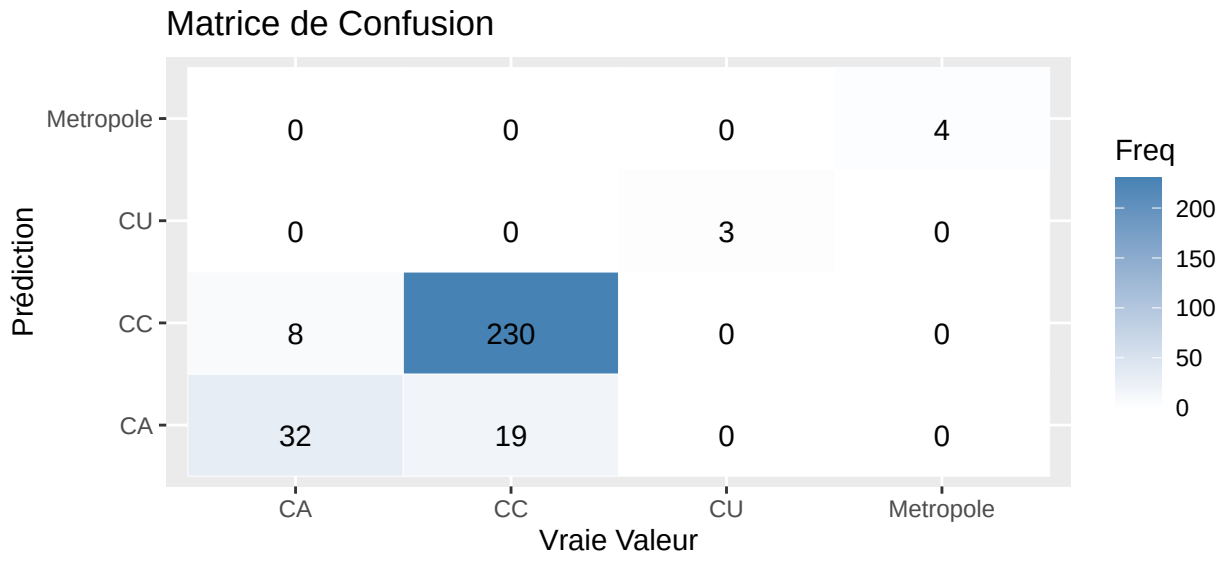



Figure 2: Prédiction sur le type d'EPCI

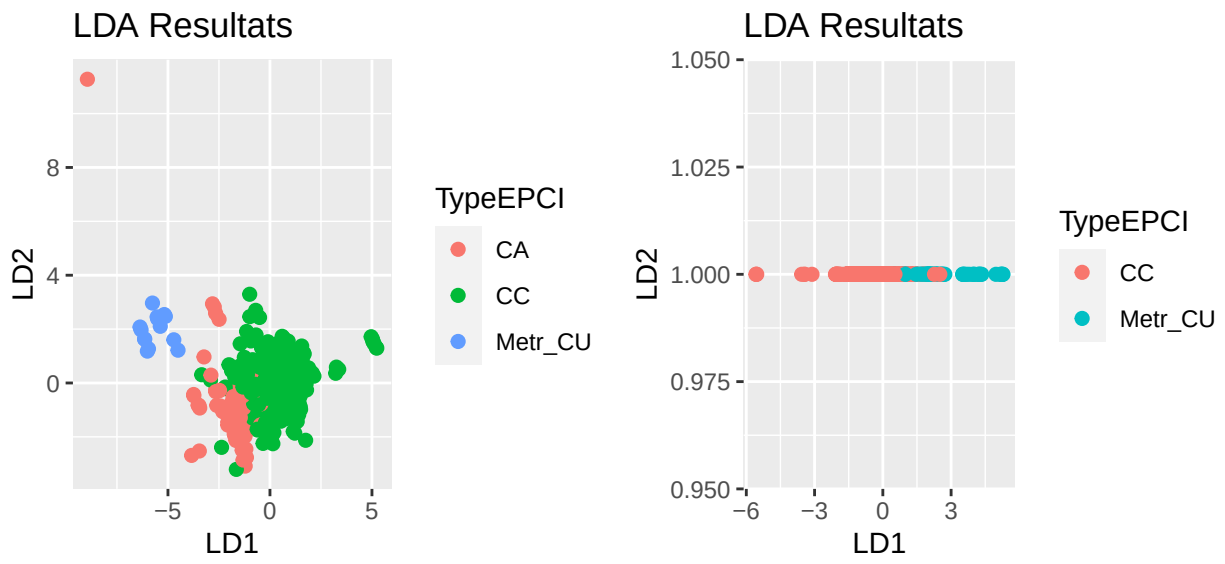
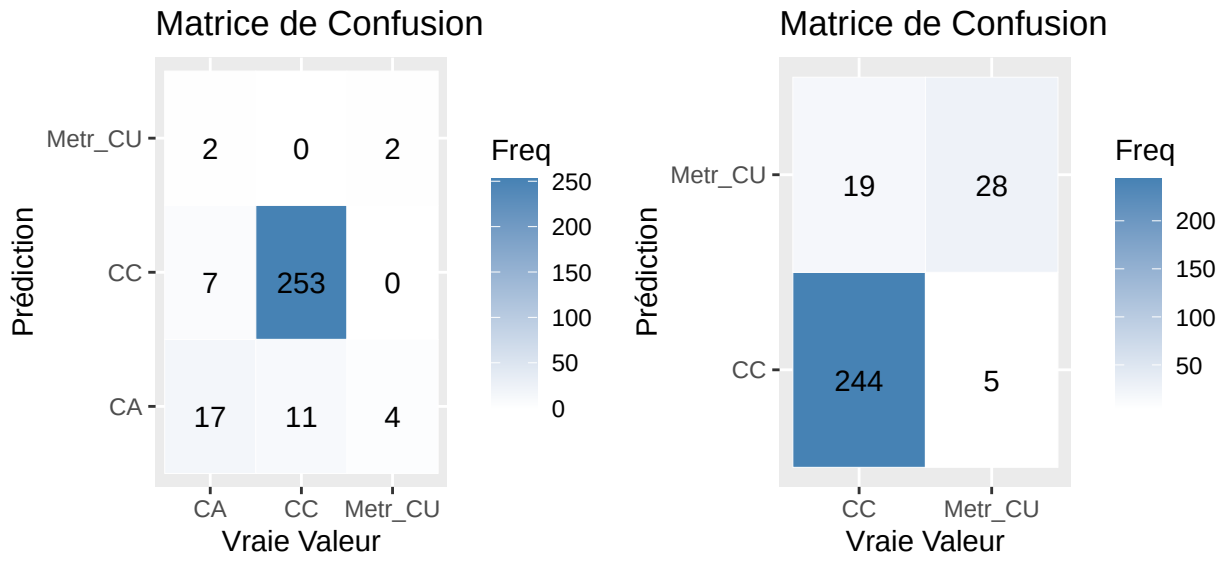



Figure 3: LDA en fonction des types EPCI

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 : 0.9189189 pour le premier regroupement, et de: 0.9189189 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.