Projet d'étude : Analyse de données & Éléments de modélisation statistique

Sara ROOL Maxime MOSHFEGHI Minh Duy NGUYEN Jules GOURIO

2024-02-02

Présentation des données et statistiques descriptives

Le jeu de données comprend 984 mesures des émissions de polluants atmosphériques tous secteurs d'activités confondues des EPCI (Etablissements Publics de Coopération Intercommunale) de la région Occitanie de 2014 à 2019.

Chaque mesure est décrite par les variables qualitatives suivantes :

- *lib_epci* : son nom
- *code_epci* : son code d'identification
- *nomdepart* : son (ses) département(s) d'appartenance
- *TypeEPCI* : CC(communauté de commune), CA (communauté d'agglomération), Métrople et CU (communauté urbaine)
- *annee_inv* : l'année de mesure

Et par les variables quantitatives suivantes :

- *nox_kg* : oxyde d'azote en kg
- *so2_kg* : oxyde de soufre en kg
- *pm10_kg* : particules en suspension dans l'air de diamètre inférieur à 10 µm
- *pm25_kg* : particules en suspension dans l'air de diamètre inférieur à 2.5 µm
- *co_kg* : monoxyde de carbone
- *c6h6_kg* : benzène
- *nh3_kg* : ammoniac
- *ges_teqco2* : gaz à effet de serre
- *ch4_t* : méthane
- *co2_t* : dioxyde de carbone
- *n2o_t* : protoxyde d'azote
- *latit* : sa latitude
- *longit* : sa longitude

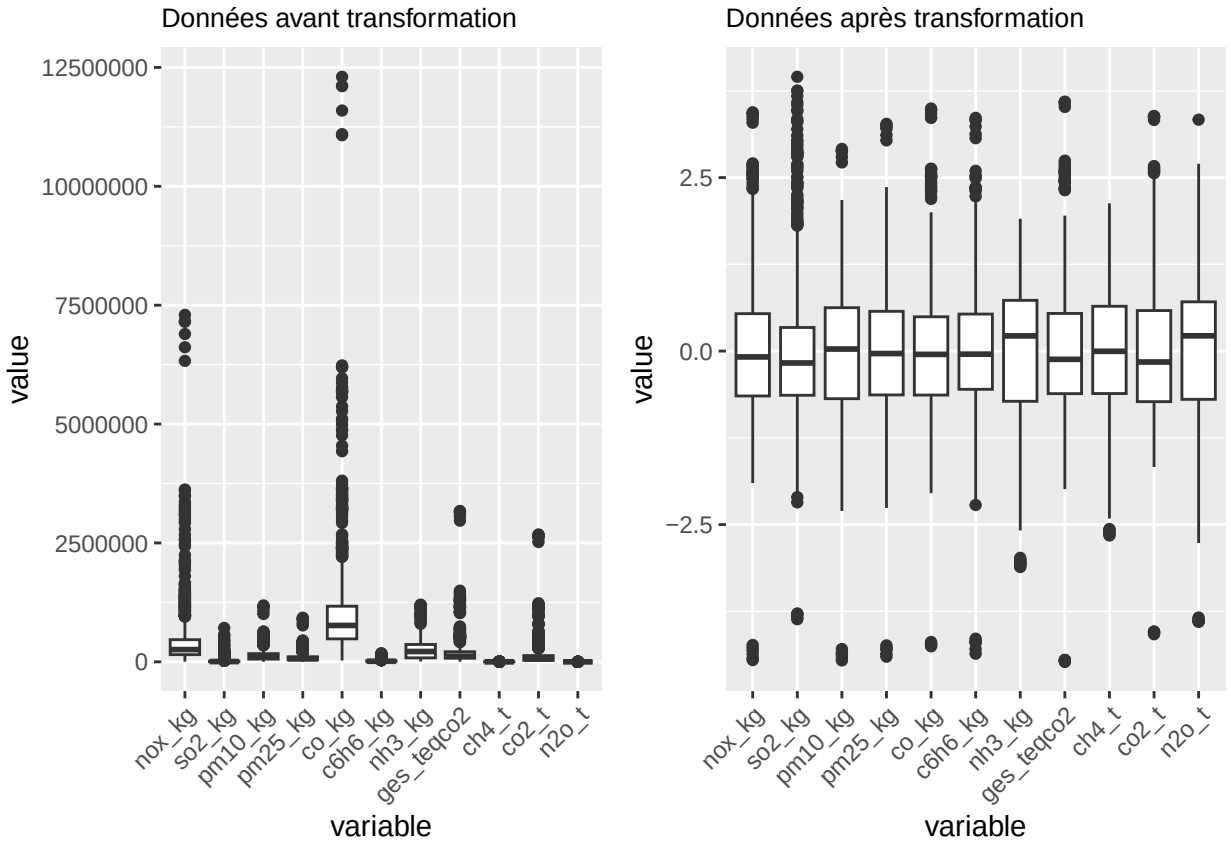
Dans la suite de ce rapport, nous utilisons la notation Δ pour faire référence à des plots qui sont disponibles dans le Rmd mais que nous avons décidé de ne pas inclure dans le rapport.

Analyse unidimentionnelle

Analyse des variables quantitatives

Nous observons avec ce boxplot que nos variables ne sont pas gaussiennes et que les ordres de grandeur sont différents, cela va biaiser nos analyses de variances. Nous appliquons donc une transformation logarithmique

et on centre et réduit nos valeurs quantitatives.



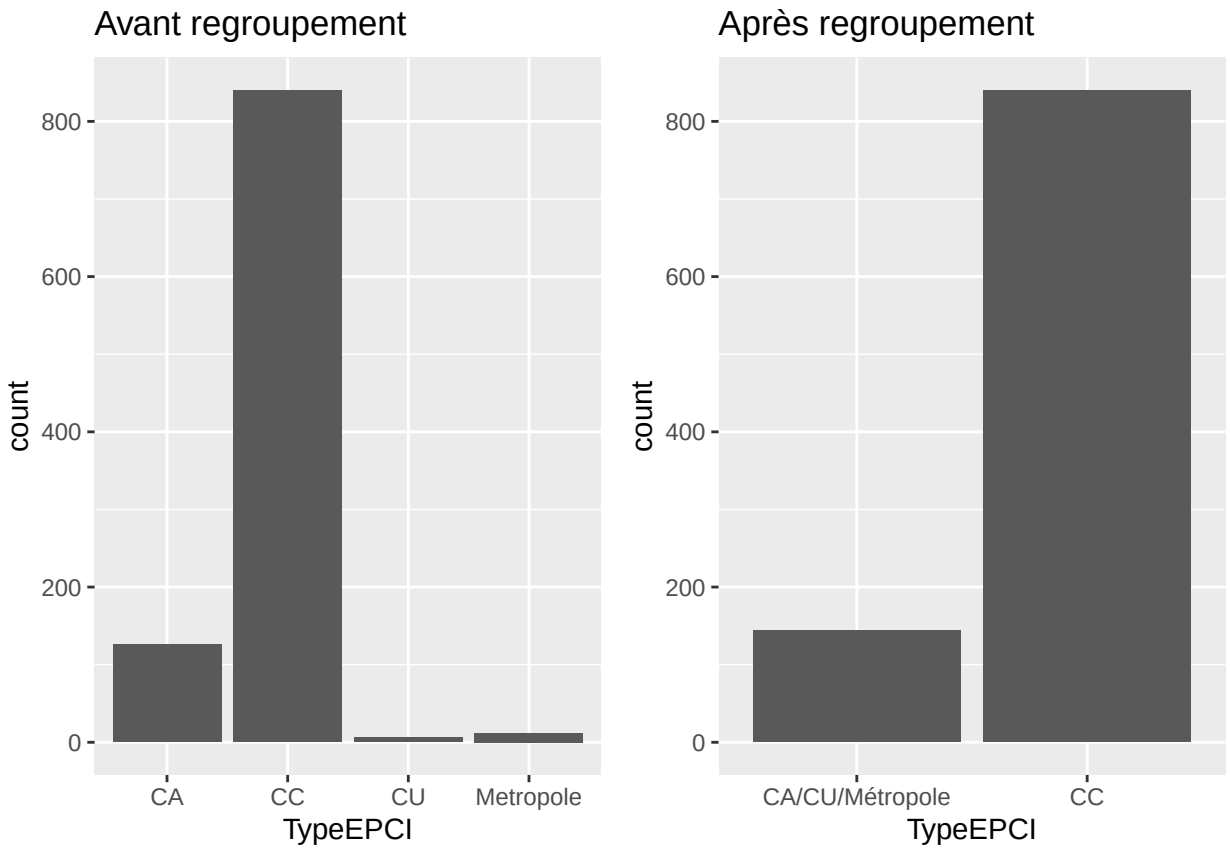
Nous remarquons que ces transformations suffisent à mieux visualiser nos données. Nous conserverons ce jeu de données transformé par la suite. Nous stockons le dataset avec les variables quantitatives transformées dans la variable *Datalog*.

Analyse des variables qualitatives

Nous nous intéressons maintenant aux variables qualitatives : $TypeEPCI$ & $annee\_inv$.

(Δ) Nous remarquons que nous avons le même nombre de prise de mesure pour chaque année, donc aucunes données ne semblent manquer.

Nous affichons la répartition des mesures en fonction de $TypeEPCI$. Nous remarquons que les catégories CA, CU et Métropole ne comprennent pas beaucoup de valeurs par rapport au type CC. Nous choisissons donc de combiner les trois catégories pour avoir un nombre suffisant de valeurs pour chacunes de nos modalités.

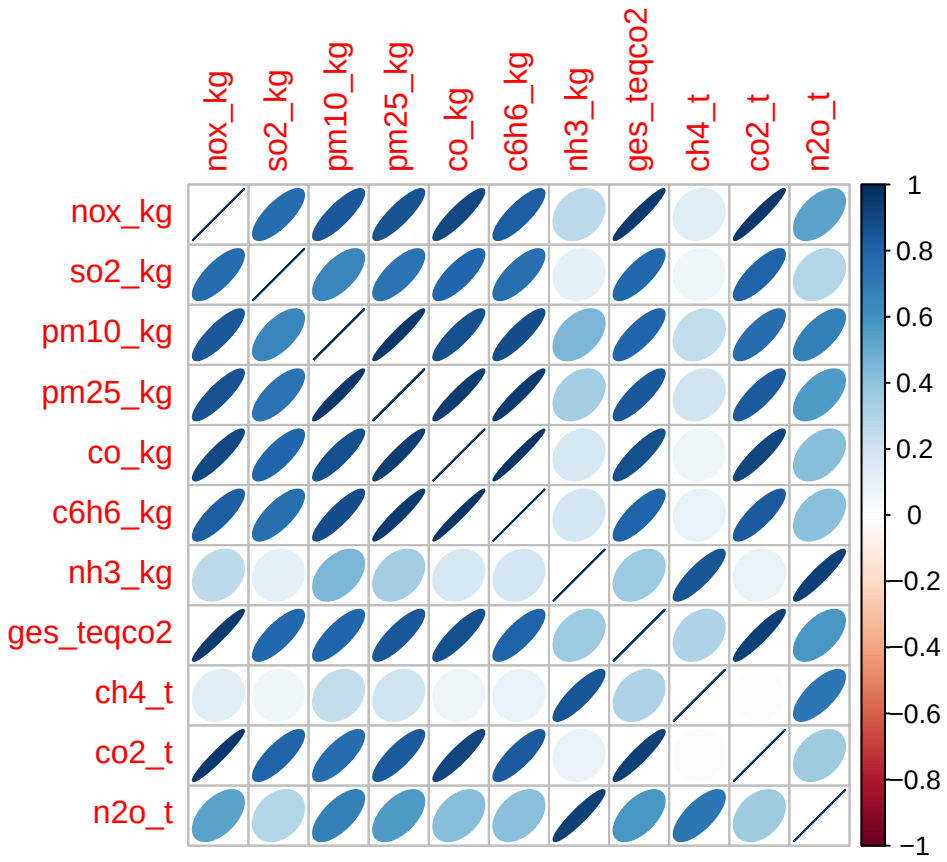


Lorsque nous comparons les années et les types EPCI avec une table de contingence, nous remarquons que les villes gardent le même type.

```
##
##                   2014 2015 2016 2017 2018 2019
##   CA/CU/Métropole   24   24   24   24   24   24
##   CC               140  140  140  140  140  140
```

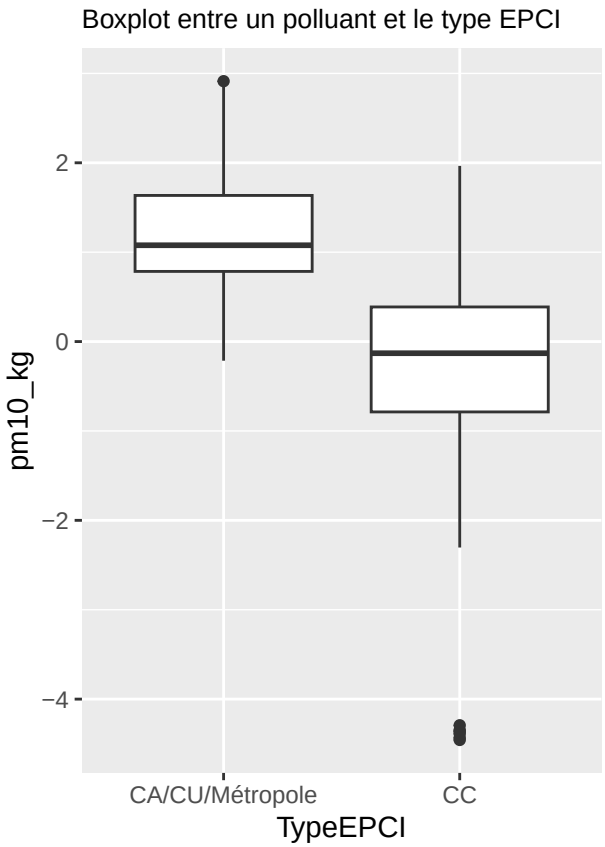
Analyse bidimentionnelle

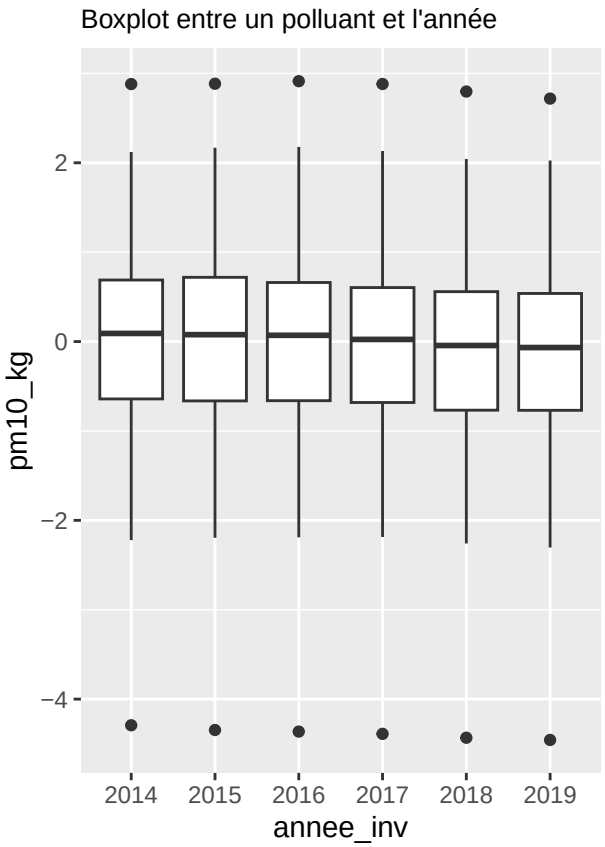
En affichant la matrice de corrélation des variables de notre dataset, nous observons que la plupart des polluants sont corrélés positivement les uns avec les autres à l'exception de $nh3\_kg$, $ch4\_t$ et $n2o\_t$ qui ne semblent être corrélés à aucune autre variable.



Nous constatons avec le boxplot ci-dessous qu'il y a un lien entre le type EPCI et les différents polluants. En effet les boxplots en fonction des types sont à des niveaux différents. Au contraire, l'année ne semble pas avoir d'influence. Nos boxplots sur le second graphique sont au même niveau.

(Δ) D'autres graphiques similaires sont disponibles dans le Rmd.

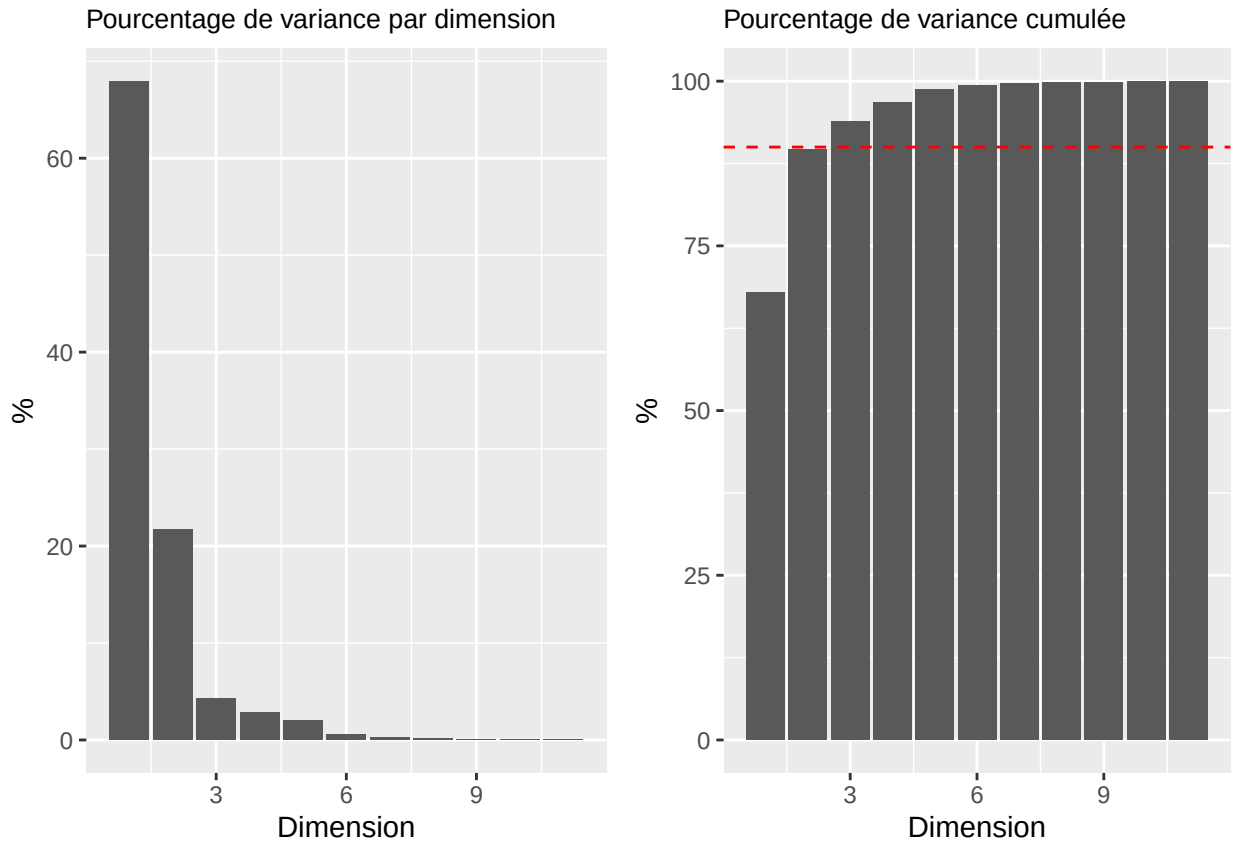




Visualisation des individus (ACP)

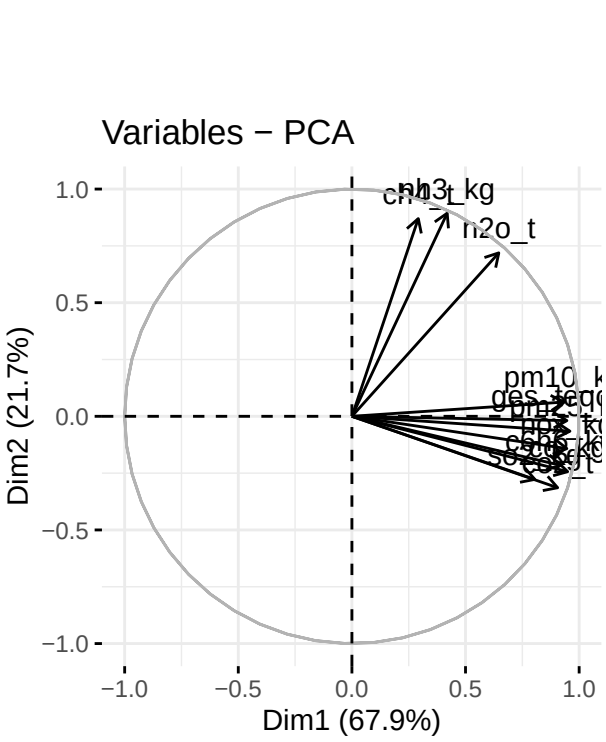
Nous réalisons une ACP pour visualiser nos données dans un espace de dimension inférieure.

Nous remarquons avec les graphes ci-dessous que les deux premières dimensions constituent 90% de la variance. Nous allons poursuivre les analyses par ces deux dimensions.

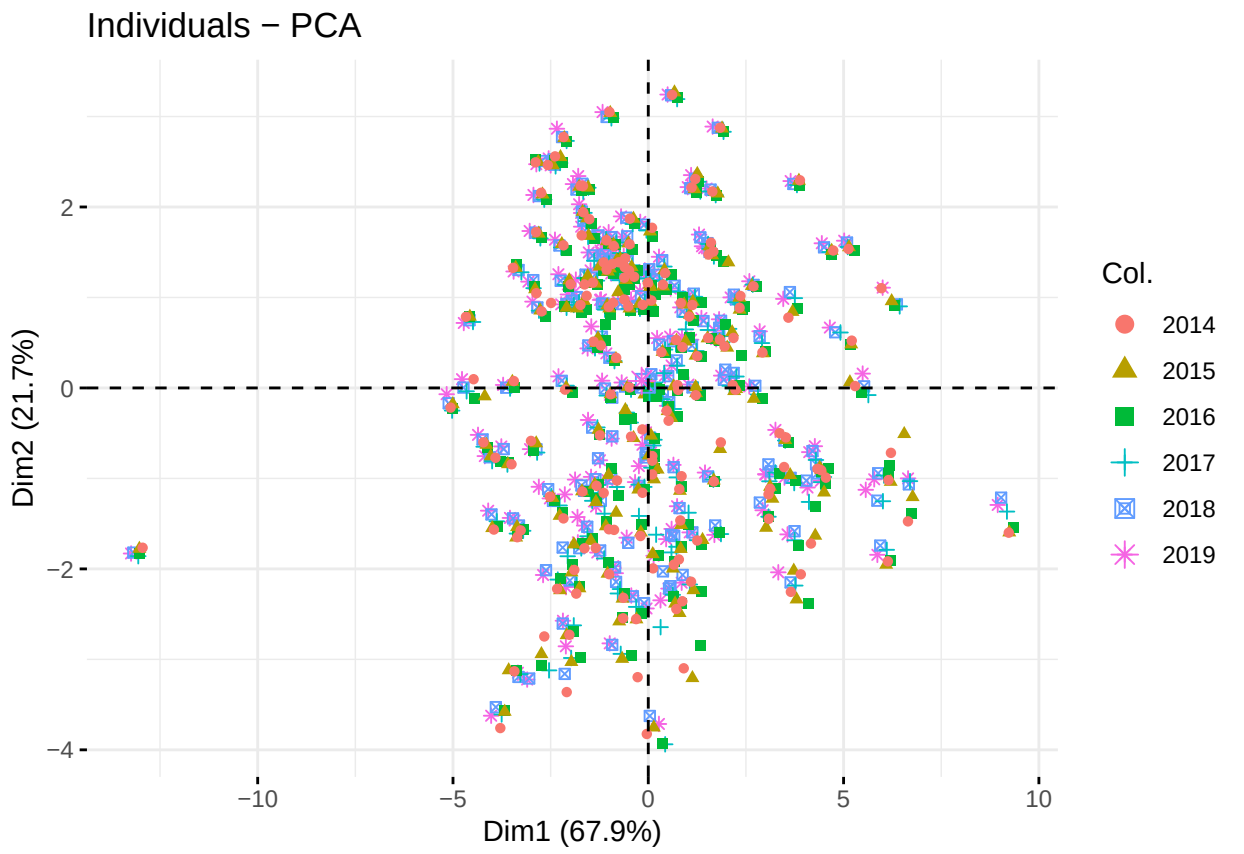


Nous affichons le cercle des corrélations. La première dimension décrit toutes les variables sauf *ch4_t*, *nh3_kg* et *n2o_t* qui décrivent la deuxième dimension. Nous remarquons que les 3 polluants qui ne sont pas corrélés avec les autres polluants (cf corrplot) composent la deuxième dimension de l'ACP.

Nous affichons également le graphe des individus colorés en fonction du Type EPCI. Nous remarquons que les types CA/CU/Metropole ont tendance à avoir en plus grandes quantité les polluants de la première dimension. Nous remarquons aussi deux groupements extrêmes. Un groupement très peu pollué et un autre plus pollué. Après identification, il s'agit de Toulouse Métropole et de CC Pays de Nay. Nous décidons d'enlever ces deux groupements car ils ne représentent pas la tendance général, ce sont des outliers qui créeront une disparité.

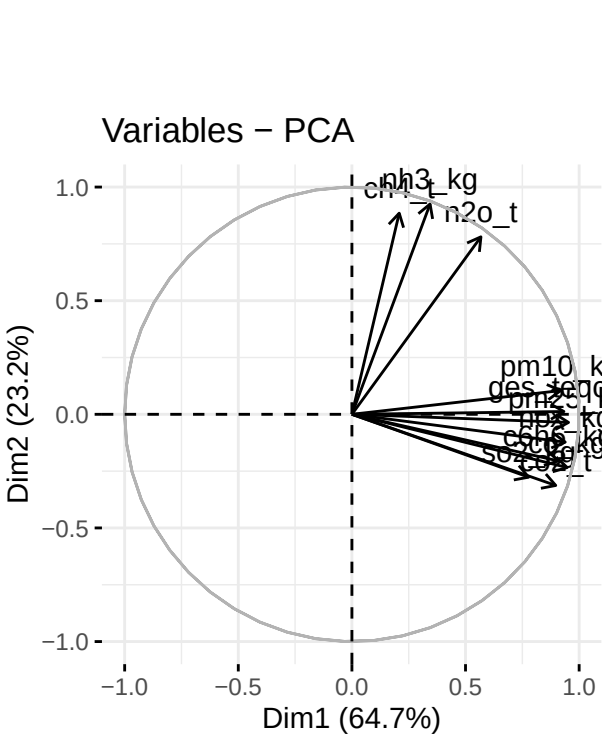


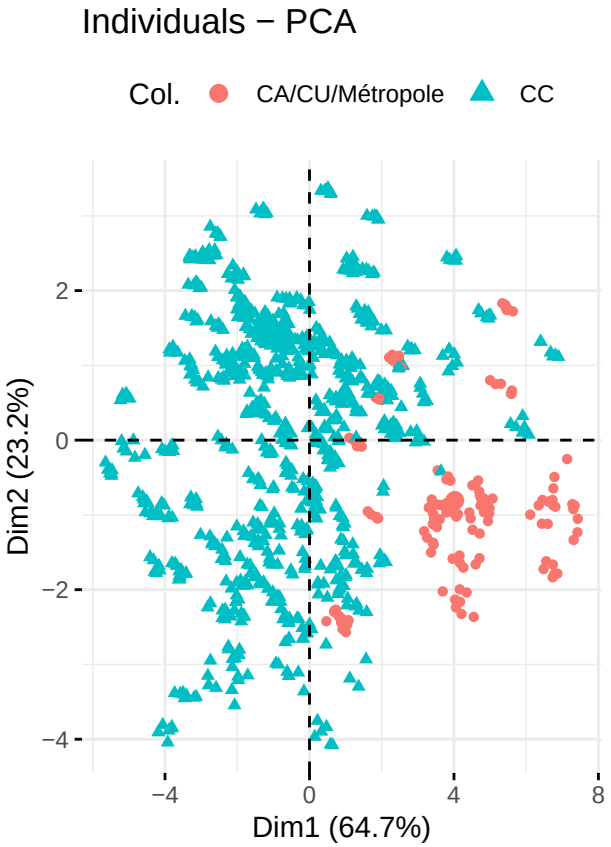
Remarque : Nous constatons, avec le graphique ci dessous, que chaque regroupement corresponds à une ville grâce à l'habillage par année. Cette visualisation nous a permis d'éliminer les outliers.



(Δ) Nous affichons aussi cette représentation sans le groupement des $TypeEPCI$. Cela confirme que nous avons un classement de la pollution en fonction du Type EPCI. Si nous alons du plus au moins pollué, nous avons : Metropole, CU, CA puis CC.

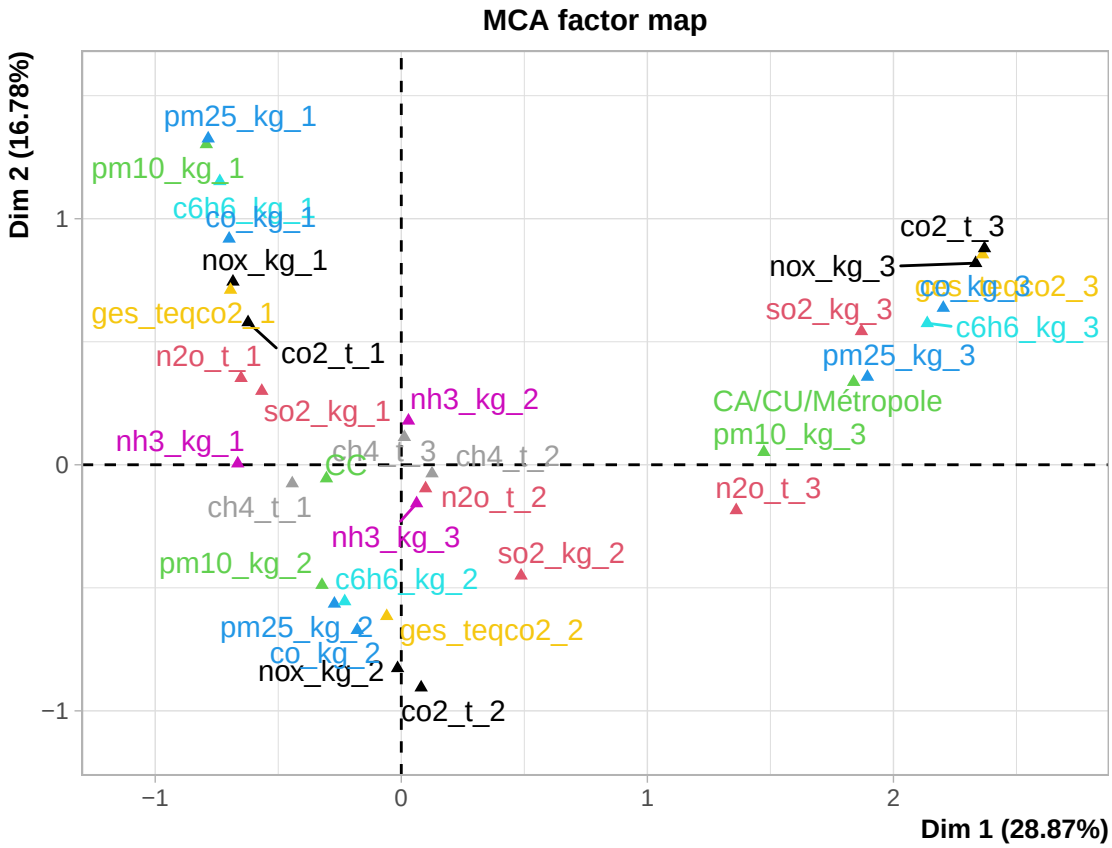
(Δ) Nous réaffichons les mêmes graphes sans les deux outliers.





Analyse multiple des correspondances

Nous réalisons maintenant une MCA à l'aide des données des polluants et du type EPCI, pour cela nous discrétisons les différents polluants afin de créer 3 modalités à chaque fois.



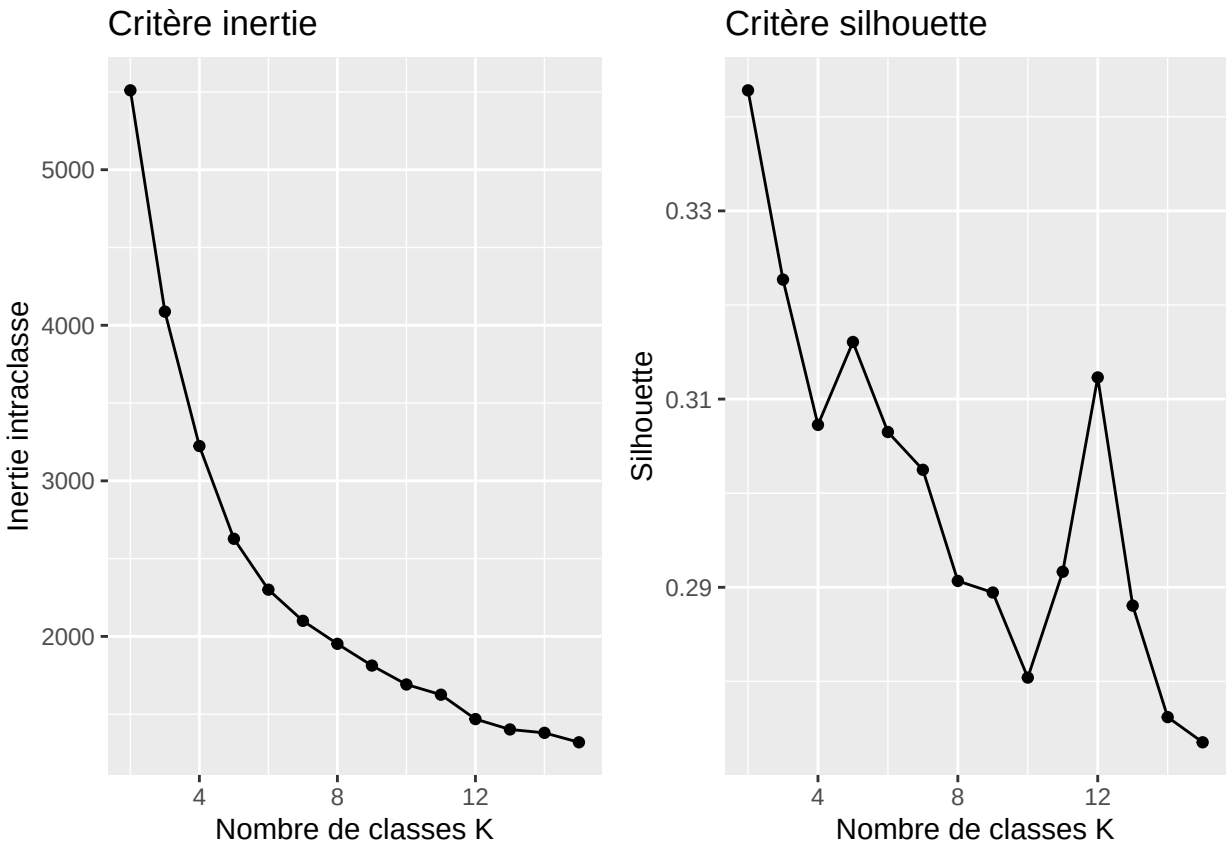
Nous obtenons des résultats plutôt satisfaisant avec 45% de la variance totale qui est expliquée par les deux premières coordonnées.

Clustering

K-means

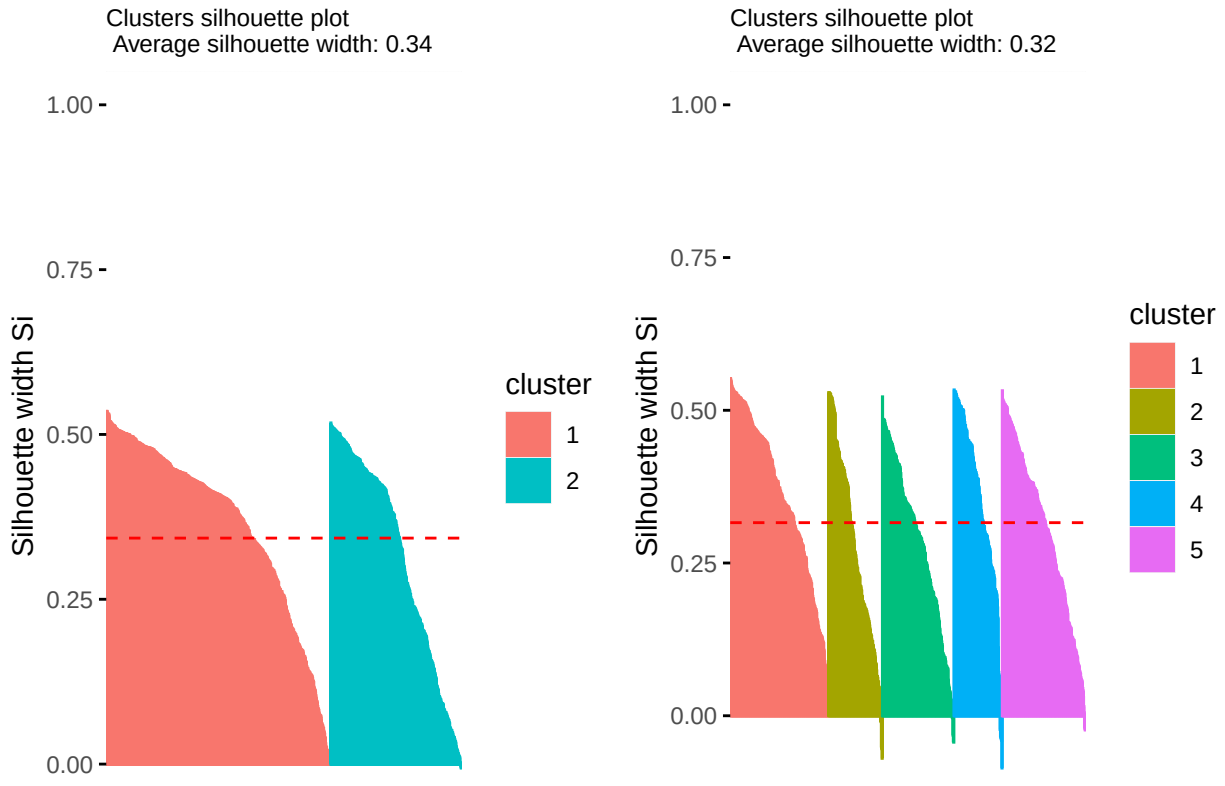
Dans un premier temps, nous allons faire un clustering de type K-means. Nous pouvons utiliser cette méthode car nous avons décidé d'enlever nos outliers, ainsi nos centroïdes ne seront pas influencés par des valeurs extrêmes.

Pour selectionner le nombre de classe, nous affichons la variation de l'inertie intra-classe et le critère silhouette.



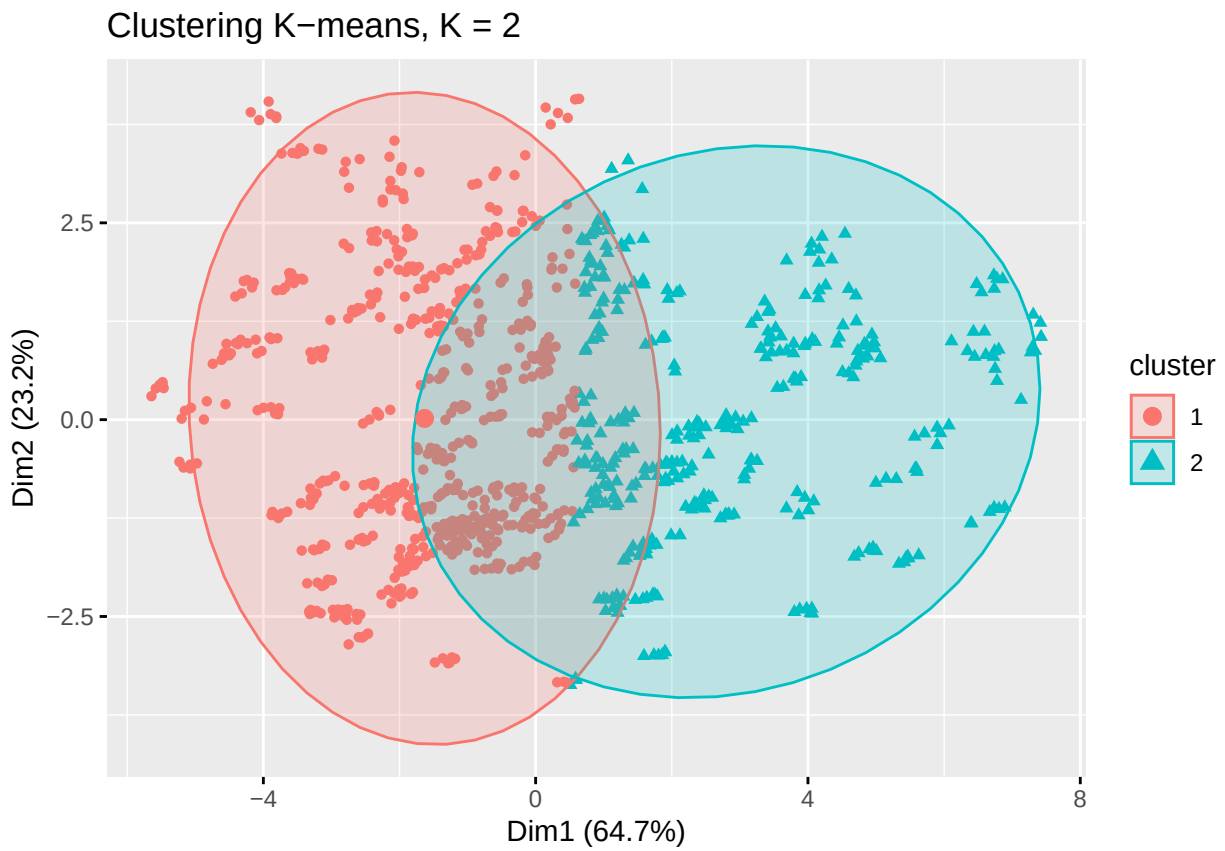
Les différents critères ne s'accordent pas sur le même nombre de classe. Avec silhouette, on choisit 2 classes alors qu'avec l'inertie on choisit 5 classes.

Si on affiche le critère Silhouette pour $K = 2$ et $K = 5$, on constate que les individus sont mal classés en général. Nous avons une hauteur moyenne de 0.34 et 0.32 ce qui confirme la mauvaise classification de nos individus.



Si on affiche la table de contingence entre le clustering à 2 et 5 classes. On constate que la classe 1 pour K=2 contient exclusivement la classe 1 et 3 pour K = 5. La classe 2 pour K=2 contient exclusivement la classe 3 pour K=5. Et les classes 2 et 5 pour K = 5 se séparent. On garde seulement 2 classes finalement.

```
##
##        1   2   3   4   5
##    1  65 133   0 269 147
##    2 163  63 132   0   0
```



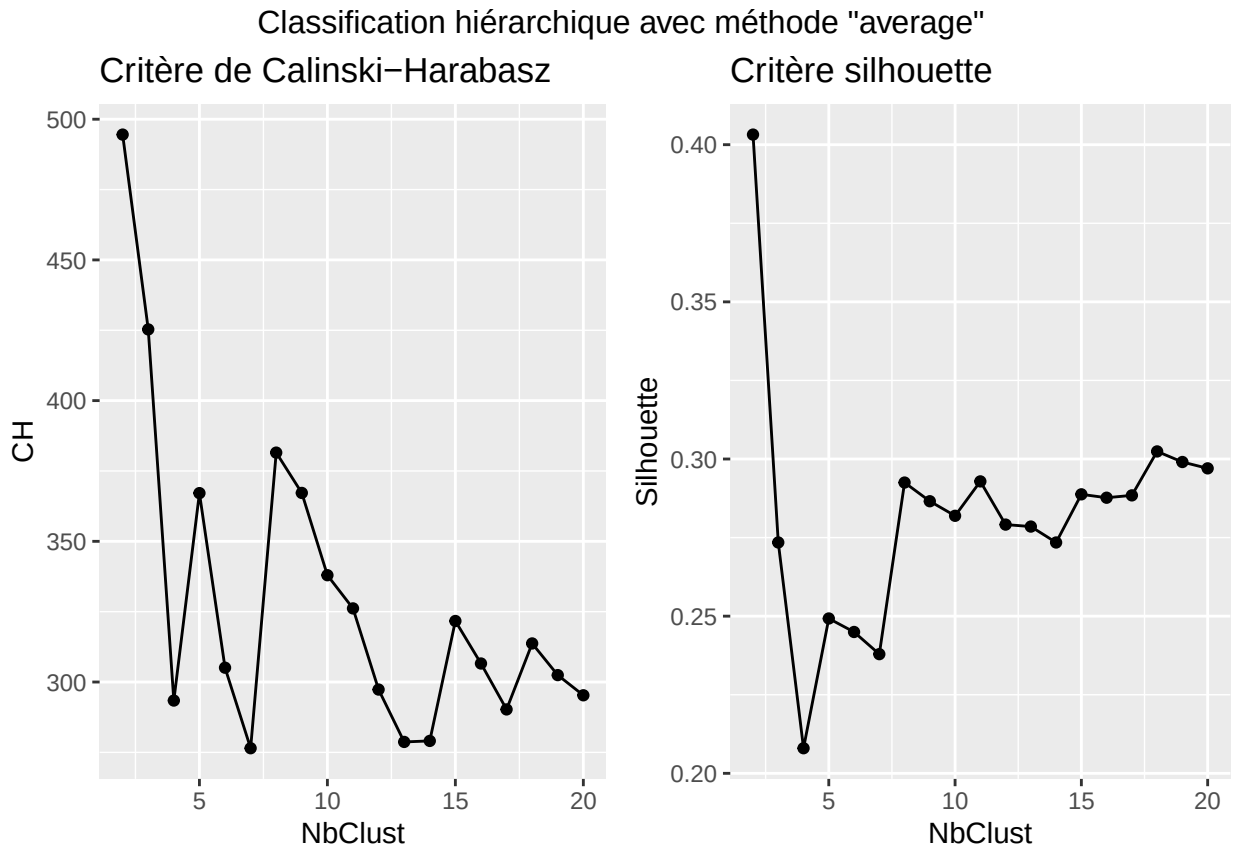
La délimitation entre les 2 classes est une droite. De plus, on compare ce clustering avec les variables qualitatives de notre jeu de donnée. On constate que d'année en année que les villes sont presques toujours classées de la même manière. Lorsqu'on compare avec les types EPCI, on remarque que les types CA/CU/Métropole sont très bien classés dans la deuxième classe mais que les types CC sont répartis dans les 2 classes. Cette classification semble proposer une nouvelle répartition des types CC/CA/CU/Métropole en fonction du niveau de pollution.

```
##
##     2014 2015 2016 2017 2018 2019
##   1  101  100   99  102  104  108
##   2   61   62   63   60   58   54
```

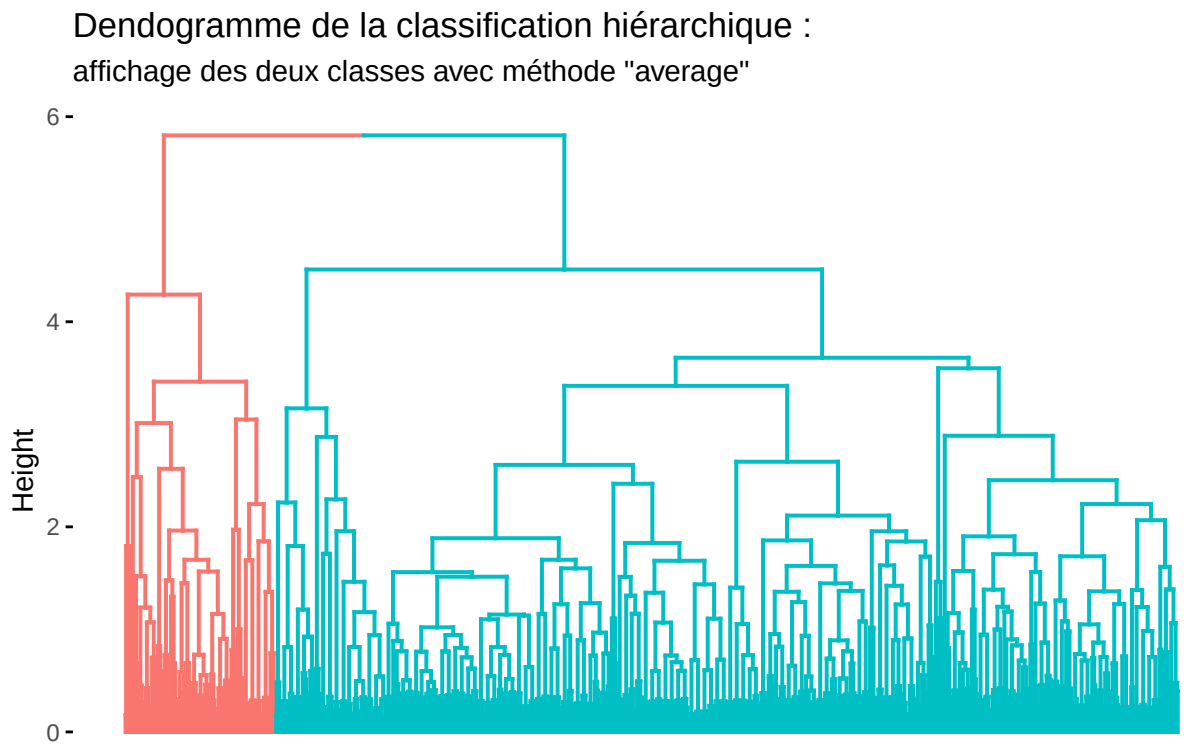
```
##
##     CA/CU/Métropole  CC
##   1               1 613
##   2             137 221
```

Classification hiérarchique ascendante

Nous faisons maintenant une classification hiérarchique ascendante sur notre jeu de données. Nous utilisons le lien moyen (average), car c'est un bon compromis entre lien simple (séparation des classes) et lien complet (diamètre des classes).



En traçant le critère de Calinski-Harabasz et de Silhouette, on remarque que les deux critères atteignent leur maximum en $K = 2$. On décide donc de garder 2 classes pour la classification hiérarchique. On affiche ci-dessous cette classification hiérarchique en deux classes pour la méthode “average”.

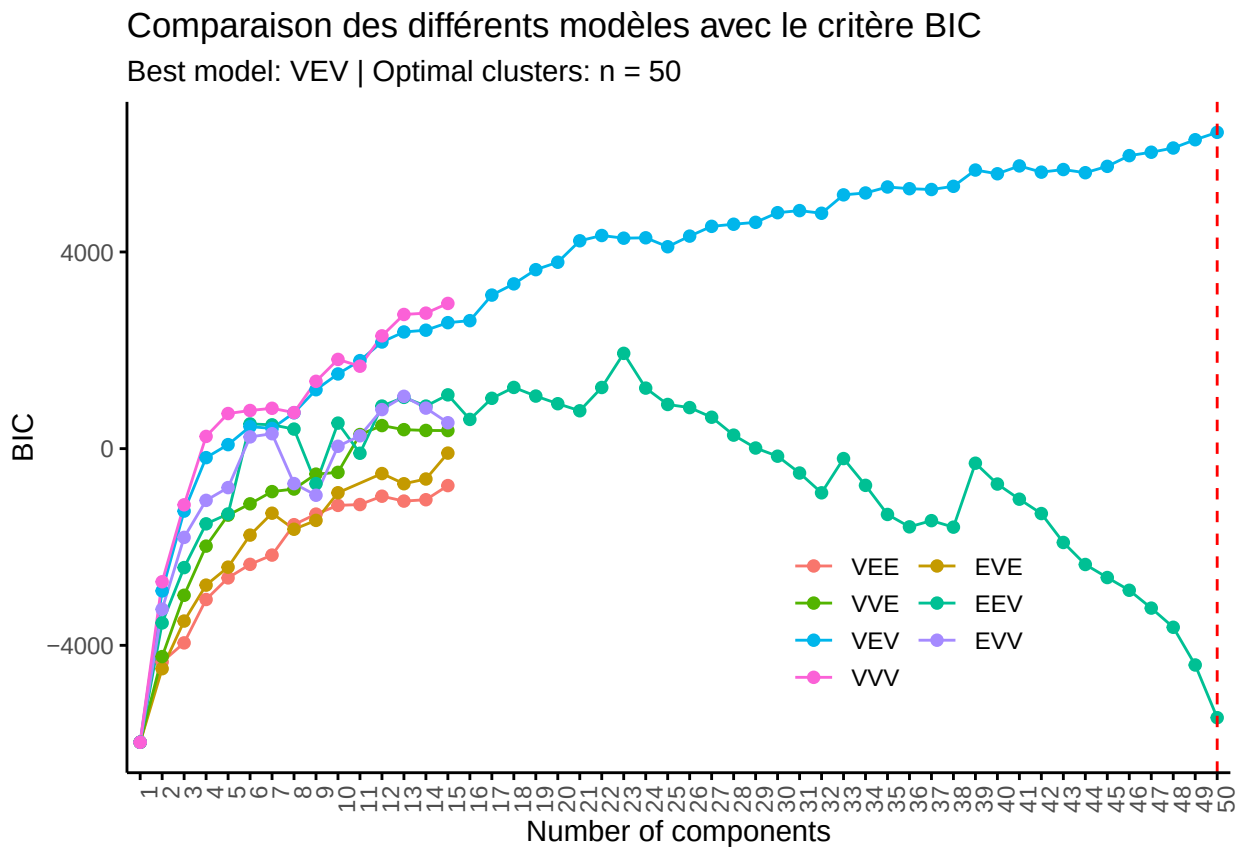


En essayant d'autres méthodes de classification hiérarchique, on remarque que les critères pour la méthode de "Ward" atteignent toutes deux leur maximum en un nombre de classes différents. Il en est de même pour la méthode "single", où l'on perçoit en plus de cela un effet de chaînage. (Δ) Ces sorties sont disponibles dans la Rmd.

Modèles de mélange

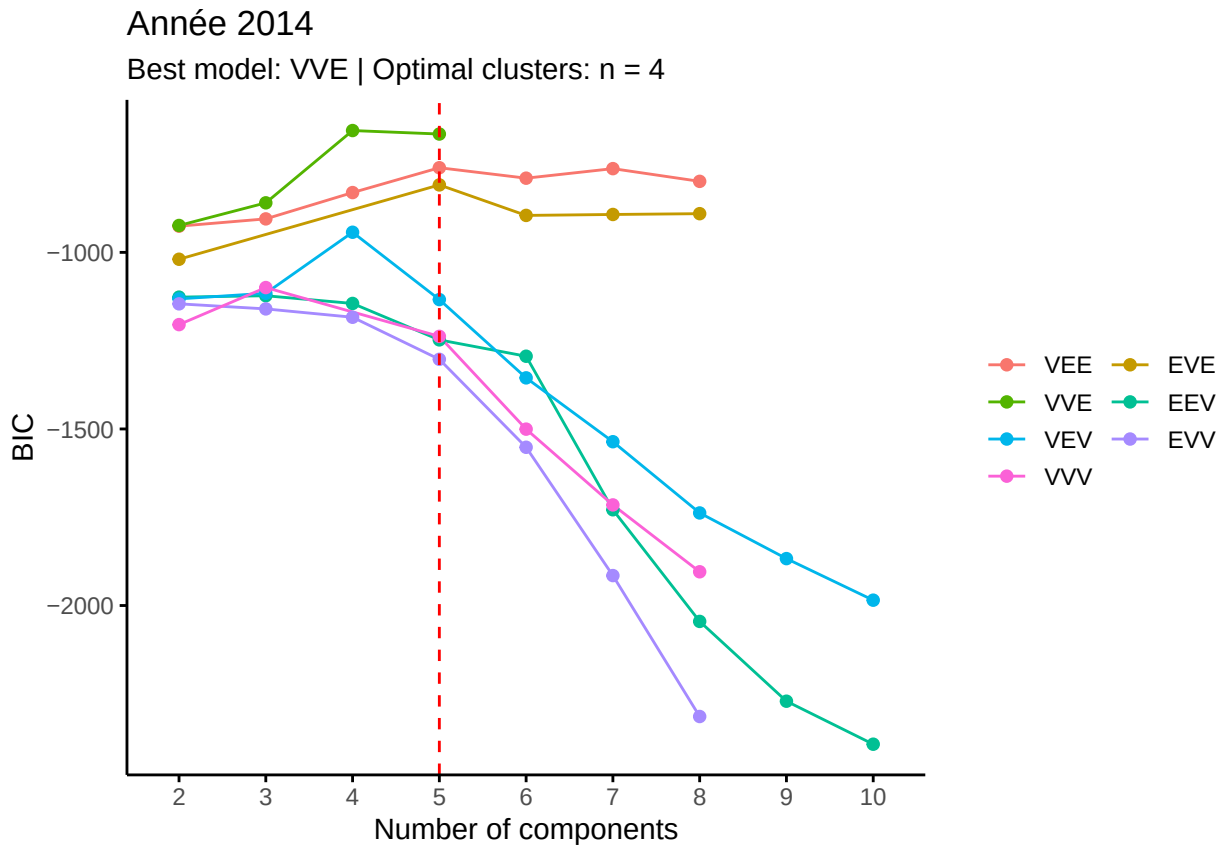
Pour finir, nous utilisons une dernière méthode de clustering : les modèles de mélange. La première étape est de définir la collection de modèle. Nous chosissons uniquement les modèles ellispoïdaux pour laisser un maximum de liberté entre les données. Ce choix est renforcé par les résultats précédents. En effet, la méthode des K-means ne donnant pas de bons résultats, on en conclut que des mélanges sphériques ne seront pas adaptés ici. Pour le nombre de classe, on se laisse une valeur maximale de K assez élevé ($K_{max} = 50$).

La comparaison des différents modèles est ici réalisé avec le critère BIC.



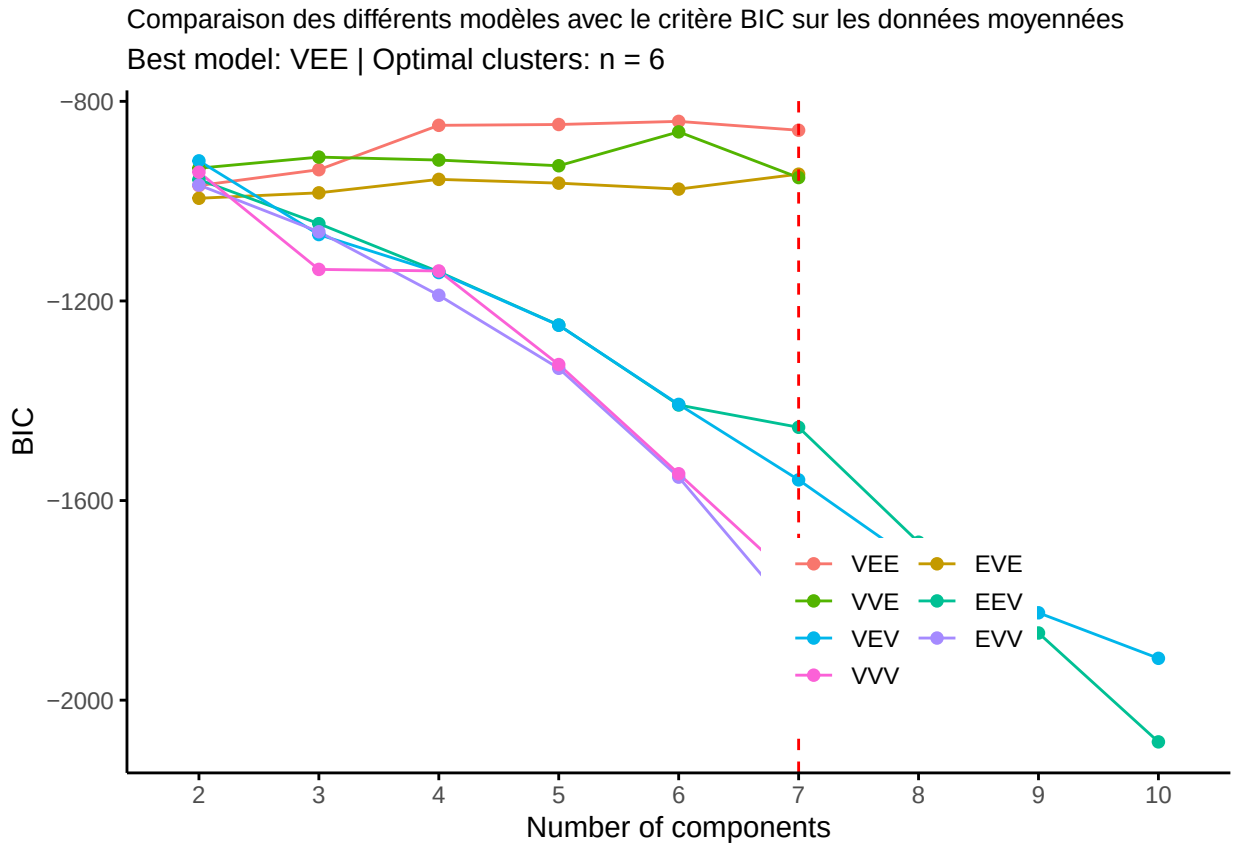
Les résultats obtenus ne sont pas satisfaisants, le critère ne pénalise pas suffisament l'augmentation du nombre de classes (les mêmes résultats sont obtenus avec le critère ICL qui pénalise pourtant plus). Nous réalisons maitenant un clustering en considérant une année. La collection de modèles reste la même. $K_{max} = 10$ est fixé pour une meilleure visualisation.

Nous affichons le clustering pour l'année 2014. (Δ) Les autres années sont disponibles dans le Rmd.



Les résultats sont nettement meilleurs, nous obtenons pour 4 années un clustering à 4 classes. L'année 2015 a un clustering à 8 classes et l'année 2017 à 5 classes. La forme de mélange retenue est VVE : forme ellipsoïdale avec une orientation similaire pour les variables.

Nous considérons maintenant toutes les années du jeu de donnée en réalisant une moyenne des polluants par année et par EPCI.



On remarque que le modèle sélectionné n'est pas le même que celui des 4 années précédentes (VVE et 4 classes). Nous obtenons finalement un modèle VEE à $K = 6$ classes.

Comparaison entre les différents clustering

Nous obtenons par K-means la table suivante :

```
##
##   1   2
## 614 358
```

Nous obtenons par CAH la table suivante :

```
##
##   1   2
## 833 139
```

Nous obtenons par modèle de mélange la table suivante :

```
##
##       1   2
##   1 614   0
##   2 219 139
```

On obtient une classification similaire : la première classe des K-means est séparés dans les deux classes de la CAH. La deuxième classe des K-means est entièrement contenue dans la première classe de la CAH.

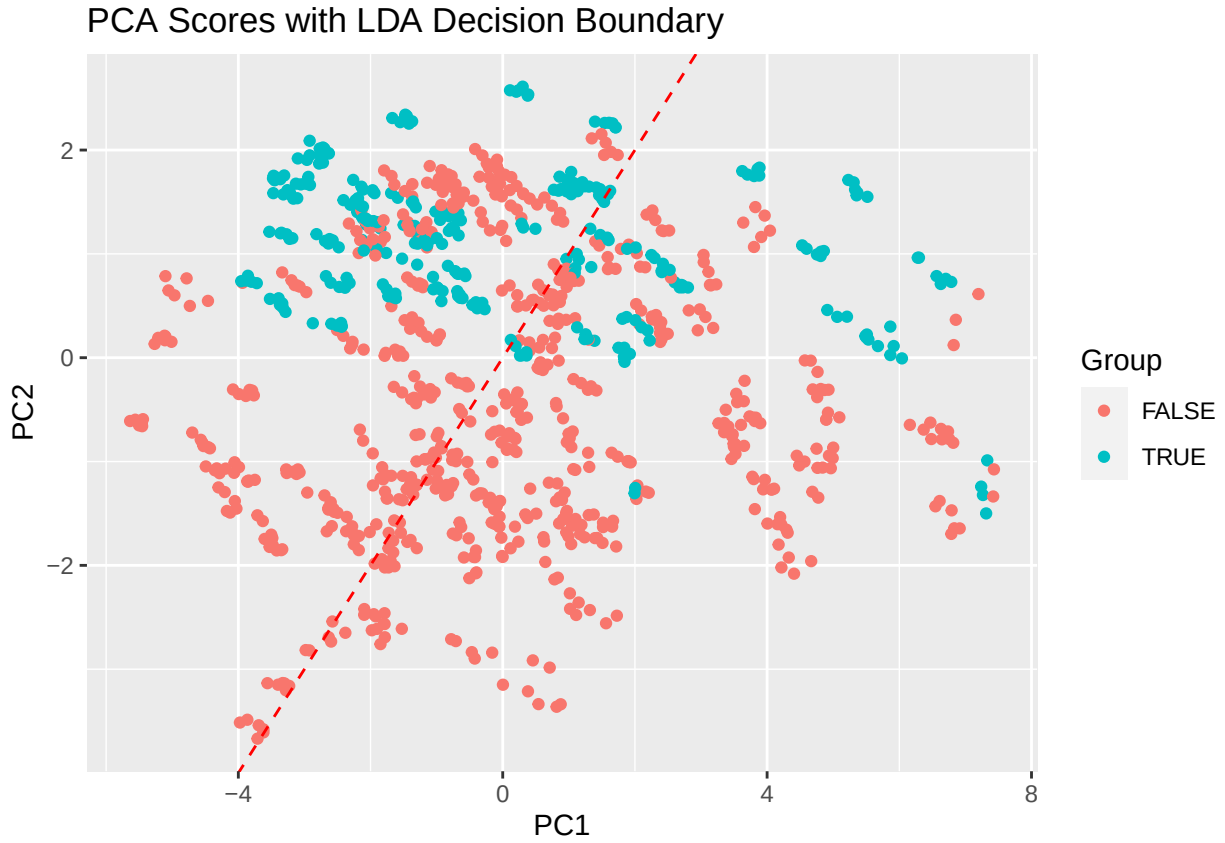
Analyse linéaire discriminante

Exploration et prédiction du dépassement d'émission de méthane de 1000 t par an

Nous voulons prédire si la variable *ch*4_*t* dépasse le seuil de 1000 t par an. Nous avons créé en amont une variable qualitative booléenne *dep_met*_1000 en fonction du dépassement du seuil.

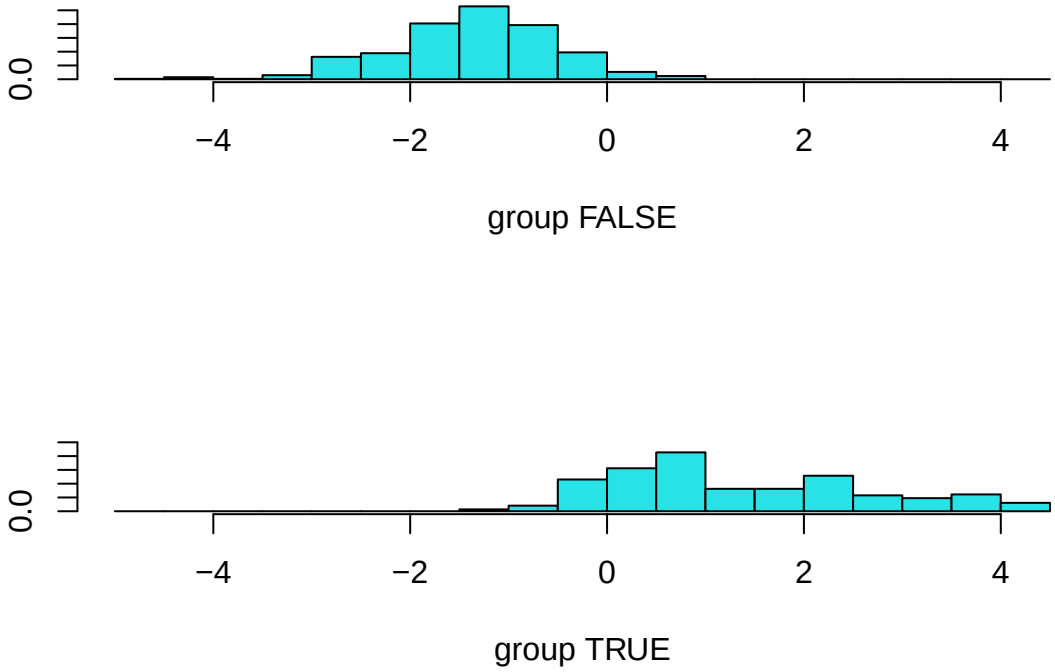
Pour la suite, nous remplaçons la colonne *ch*4_*t*, dans les données modifiées, par y.

Nous traçons la ligne droite séparant les individus en classes. Notons qu'ici, nous n'avons que 2 modalités ainsi la sortie de LDA sera une seule ligne droite.



Nous rappelons que l'hypothèse pour une LDA est que les individus d'une classe sont tirés d'une gaussienne commune avec la même matrice de covariance. Si l'on observe les deux nuages de points sur le graphique de l'ACP, on constate qu'ils ont une forme assez ellipsoïdale et uniforme. Ce qui signifie que les individus sont effectivement tirés de la même gaussienne. De plus, la deuxième dimension est expliquée plus haut comme étant proportionnelle à l'émission de *ch*4_*t*. Ainsi lorsque nous observons les 2 nuages, celui représentant la classe 'au-dessus' est au-dessus (i.e. supérieur) de celui représentant la classe 'en dessous', ce qui est totalement cohérent avec ce que nous avons trouvé dans l'ACP.

De plus, la projection des individus dans le sous-espace 2D explique environ 90% de la variance. Nous examinons ensuite l'histogramme des individus projetés sur la droite de décision de LDA .



Nous observons que les deux modalités sont déviées de deux côtés différents lorsqu'elles sont projetées sur l'axe de décision. On remarque cependant une zone commune aux deux classes, ce qui peut rendre difficile la prédiction.

Pour la prédiction, nous divisons l'ensemble de données en deux : un train set avec 70 des données et un test set avec 30 des données.

Nous appliquons la méthode LDA à l'ensemble d'apprentissage et effectuons la prédiction. Nous examinerons donc les tables de confusion et de précision sur l'ensemble de l'apprentissage.

```
##           Actual
## Predicted FALSE TRUE
##     FALSE   448   40
##     TRUE      9  176
```

```
## [1] "La précision sur le trainset : 0.927191679049034"
```

Puis on cherche à observer comment fonctionne le modèle sur le set de test.

```
##           Actual
## Predicted FALSE TRUE
##     FALSE   215   21
##     TRUE      5   58
```

```
## [1] "La précision sur le testset : 0.91304347826087"
```

Nous constatons que les valeurs situées sur la diagonale du tableau sont correctement estimées par le modèle, les autres étant incorrectes. La LDA a prédit avec une grande précision (>90%). Cependant la LDA donne encore de nombreux résultats erronés en raison de la présence du chevauchement dans l'histogramme ci-dessus.

Prédire le type d'EPCI

On veut prédire le type d'EPCI en fonction des autres variables en utilisant l'analyse linéaire discriminante. On affecte à la variable y le $TypeEPCI$.

Comme on a 2 modalités de type d'EPCI, la LDA est ici appliquée sur 2 individus (les 2 centroides des 2 classes : 'CC' et 'CA/CU/Métropole') dans un espace de 12 variables. Nous cherchons à visualiser les 2 classes avec une PCA représentation pour les données log-modifiées.

On prépare les données pour la prédiction. Comme la partie avant, on divise nos données en train set (70% des données) et test set (30% des données). Puis on applique la LDA sur le train set.

Dans le résultat, la partie `Prior probabilities of groups` nous indique que 86% des données sont de type d'EPCI CC. La partie `Coef. of linear discriminants` affichent la combinaison linéaire de variables prédictives utilisées pour former la règle de décision du modèle LDA.

Maintenant on lance la prédiction par le modèle LDA qu'on a ajusté par le train set. Ici on cherche à voir quel pourcentage d'observations le modèle LDA a correctement prédit le type d'EPCI en regardant la table de confusion:

```
##                  Actual
## Predicted         CA/CU/Métropole  CC
##   CA/CU/Métropole              38   5
##   CC                            8 248
```

```
## [1] "La précision sur le testset : 0.956521739130435"
```

Il s'avère que le modèle a correctement prédit l'espèce pour 95.6% des observations du test set. C'est un résultat très bon pour un algorithme appliqué sur les données réelles. Une raison que l'on peut trouver est qu'il a y peu d'individus de type "CA/CU/Métropole" par rapport au type "CC".

En ce qui concerne la partie classification, le modèle a peut-être prédit correctement à un niveau aussi élevé parce que nos données ne sont pas trop compliquées, et qu'il n'y a que deux modalités de type EPCI. Pour la classification du dépassement de $ch4\_t$, les données se chevauchent davantage entre les deux modalités.

Régression linéaire

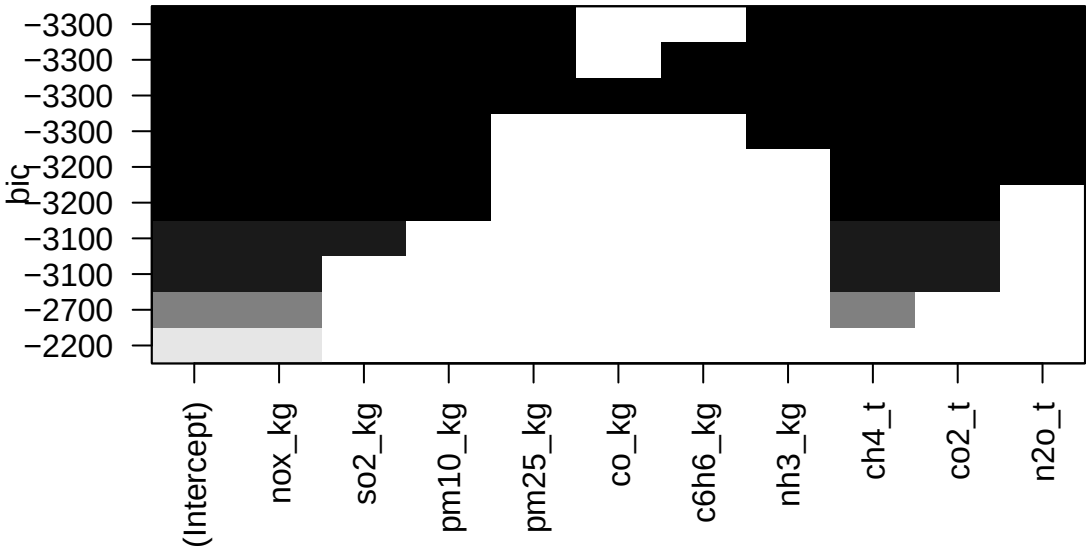
Dans cette section, nous allons expliquer les émissions de gaz à effet de serre ($gesteqco2$) en fonction de tout les autres polluants par un modèle de régression linéaire multiple. Nous utilisons la fonction lm() de R pour effectuer la régression linéaire.

```
##
## Call:
## lm(formula = ges_teqco2 ~ ., data = Data_final[, indices_quantitatives])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59070 -0.09051 -0.01750  0.06126  0.75737
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.001582   0.005143  -0.308  0.75843
## nox_kg       0.323282   0.031857  10.148  < 2e-16 ***
## so2_kg       0.087538   0.009093   9.627  < 2e-16 ***
## pm10_kg     -0.266334   0.032806  -8.119 1.44e-15 ***
## pm25_kg      0.172602   0.040645   4.247 2.38e-05 ***
## co_kg        0.127785   0.051851   2.464  0.01390 *
## c6h6_kg     -0.114365   0.043902  -2.605  0.00933 **
## nh3_kg      -0.255998   0.031609  -8.099 1.67e-15 ***
## ch4_t        0.242289   0.013221  18.327  < 2e-16 ***
## co2_t        0.485027   0.030958  15.667  < 2e-16 ***
## n2o_t        0.358738   0.033340  10.760  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1603 on 961 degrees of freedom
## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9683
## F-statistic:  2968 on 10 and 961 DF,  p-value: < 2.2e-16
```

La sortie de R nous montre des $p - valeurs$ largement inférieures à 0.05 (seuil le plus courant) pour toutes les variables sauf 2 (co_kg et c6h6_kg) mais même ces $p - valeurs$ restent inférieures à 0.05 ce qui veut dire qu'au niveau 5% on conserve toutes les variables.

On décide tout de même d'effectuer un algorithme de sélection de variables avec 3 différents critères (BIC, Adjusted R-Squared et C_p de Mallows) utilisant la méthode pas à pas de type forward (les autres méthodes donnent essentiellement les mêmes résultats).



Parmi les 3 critères de sélection de variables, le critère BIC est le seul critère à éliminer des variables. Il élimine les variables co_kg et c6h6_kg qui étaient les deux variables avec les plus grosses p_{valeur} . On nomme ce modèle M_{RL_1} , voici son expression :

$$M_{RL_1} : \begin{cases} ges\_teqco2_i = b_0 + b_1 \cdot nox\_kg_i + b_2 \cdot so2\_kg_i + b_3 \cdot pm10\_kg_i + b_4 \cdot pm25\_kg_i \\ \quad + b_5 \cdot co\_kg_i + b_6 \cdot c6h6\_kg_i + b_7 \cdot nh3\_kg_i + b_8 \cdot ch4\_t_i \ + b_9 \cdot co2\_t_i + b_{10} \cdot n2o\_t_i + \varepsilon_i \\ \text{où } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \end{cases}$$

On rééstime de nouveau les coefficients avec les variables sélectionnées avec lm() avant de réaliser un test de sous-modèle avec la fonction anova() de R.

On obtient une $p_{valeur} = 0.03193 \geq 0.05$ ce qui veut dire qu'on ne retient pas ce test au seuil 5% (qui est le seuil couramment utilisé). Cela signifie que l'on conserve toutes les variables des polluants pour l'explication du gaz à effet de serre.

On essaye aussi de faire de la régression régularisée, mais on obtient sensiblement les mêmes résultat, on ne parvient pas à simplifier le modèle. On change maintenant le modèle pour rajouter des interactions d'ordre 1 entre toutes les polluants.

$$M_{RLint_0} : \begin{cases} ges\_teqco2_i = c_0 + c_1 \cdot nox\_kg_i + c_2 \cdot so2\_kg_i + c_3 \cdot pm10\_kg_i + c_4 \cdot pm25\_kg_i \\ \quad + c_5 \cdot co\_kg_i + c_6 \cdot c6h6\_kg_i + c_7 \cdot nh3\_kg_i + c_8 \cdot ch4\_t_i + c_9 \cdot co2\_t_i \\ \quad + c_{10} \cdot n2o_{ti} + c_{11} \cdot (nox\_kg_i \times so2\_kg_i) + c_{12} \cdot (pm10\_kg_i \times pm25\_kg_i) \\ \quad + \ldots + c_{55} \cdot (co2\_t_i \times n2o\_t_i) + \varepsilon_i \\ \text{où } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \end{cases}$$

On a tenté de simplifier le modèle avec interactions, cependant les résultats ne sont pas concluants, on ne supprime pas assez de variables ce qui rend le modèle trop compliqué. On préfère donc garder le modèle sans interaction.

Etude du gaz à effet de serre (ANOVA à 2 facteurs)

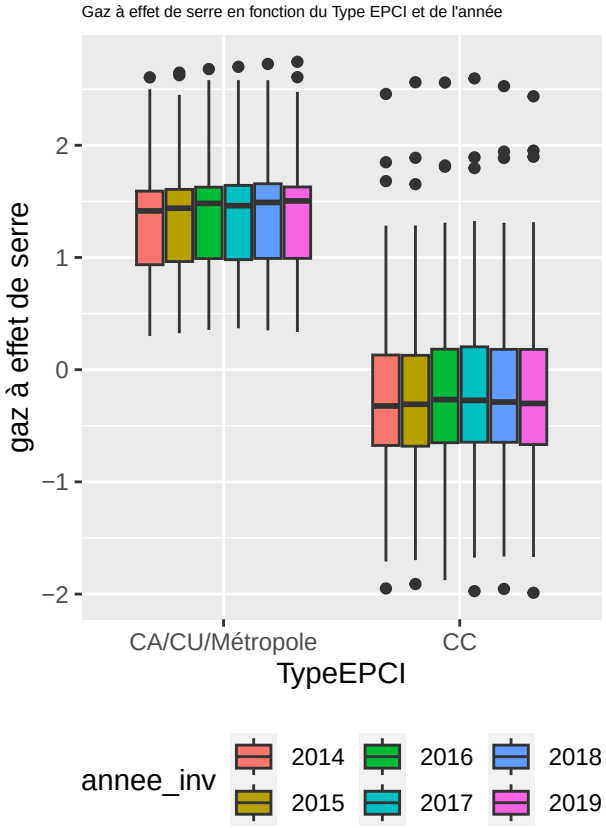
Dans cette partie, nous voulons expliquer l'émission du gaz à effet de serre en fonction du type EPCI et de l'année. Ces deux variables sont qualitatives, nous effectuons une ANOVA à deux facteurs.

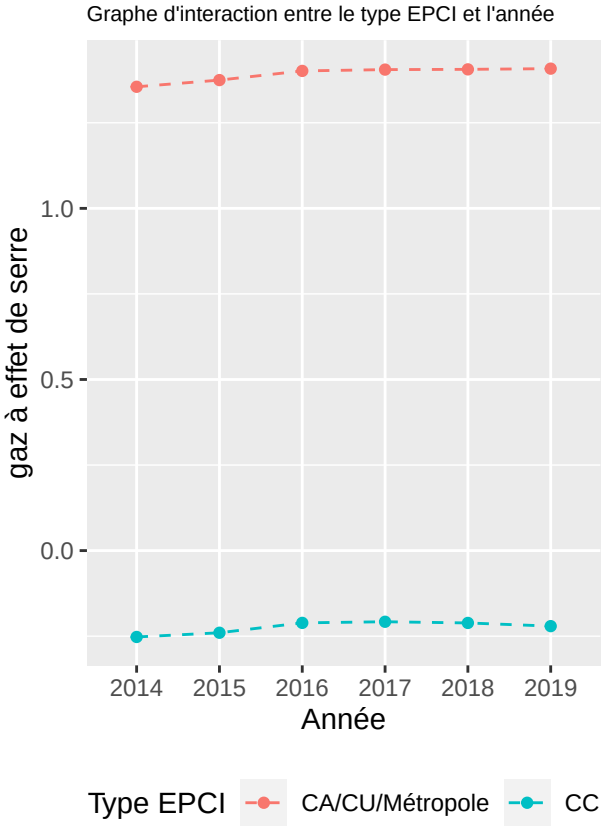
Nous considérons le modèle complet avec interactions suivant :

$$(M0_{anova}) \begin{cases} i \in \{\text{CC, CA/CU/Metropole}\}, j \in \{2014, ..., 2019\}, k \in \{1, ..., n_{ij}\} \\ \text{gesteqco2}_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

avec gesteqco2_{ijk} l'émission de gaz à effet de serre du k-ième individu avec un type EPCI i et d'année k.

Par des analyses exploratoires, nous observons une influence du type EPCI sur l'émission du gaz à effet de serre mais nous ne constatons pas de grande influence de l'année. De plus avec le graphe d'interaction, nous observons qu'il n'y a pas d'interaction entre le type EPCI et l'année. En effet, nos droites sont parallèles.





Ces deux graphiques et des sorties de test de student (Δ) nous laissent supposer que l'on peut réduire le modèle. Nous allons confirmer ces intuitions avec des tests de sous modèle.

Nous considérons le modèle additif (sans interactions) et nous effectuons le test de sous modèle avec le modèle complet ce qui revient à tester la nullité de l'interaction. La p-valeur de ce test est environ égale à 1 donc nous ne rejettons pas $H0$ et on enlève les effets d'interactions.

Nous testons ensuite la nullité de chaque variable individuellement ($Année_inv$ et $TypeEPCI$) par rapport au modèle additif.

Avec les sorties des tests (Δ), nous constatons que nous ne pouvons pas annuler l'effet des types EPCI cependant nous pouvons enlever la variable $annee_inv$. En effet, nous avons une p-valeur de 0,9895.

Finalement, nous tester le modèle sans interaction et sans la variable année par rapport au modèle complet. Nous avons un p-valeur trèc proche de 1.

Ainsi, l'émission de gaz à effet de serre peut s'expliquer qu'en fonction des $TypeEPCI$. Nous conservons alors le modèle suivant :

$$(M1_{anova}) \begin{cases} i \in \{\text{CC, CA/CU/Metropole}\}, k \in \{1, ..., n_{ij}\} \\ \text{gesteqco2}_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \\ \varepsilon_{ijk} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

Etude de l'émission de méthane (ANCOVA)

Nous souhaitons étudier l'émission de méthane $ch4\_t$ en fonction de l'ammoniac $nh3\_kg$, du protoxyde d'azote $n2o\_t$, du type d'EPCI $TypeEPCI$ et de l'année $annee\_inv$. Nous avons des variables quantitatives et qualitatives, nous allons donc faire une ANCOVA. Nous effectuons l'ANCOVA sur les données modifiées.

Dans un premier temps, on considère le modèle complet avec interactions.

$$(M_1) \begin{cases} ch4\_t_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + (a1 + a2_i + a3_j) \times nh3\_kg_{ijk} + (b1 + b2_i + b3_j) \times n2o\_t_{ijk} + \nu \times n2o\_t_{ijk} \times nh3\_k \\ i = 1, \ldots, I = 6,\ j = 1 \ldots, J = 2,\ k = 1, \ldots, n_{ij}. \\ (\varepsilon_{ijk})_{i,j,k} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

$ch4\_t_{ijk}$ représentes la valeur de la $k^{ème}$ mesure du $ch4\_t$ pour la $i^{ème}$ année et pour le $j^{ème}$ type d'EPCI.

Au vue des sorties des tests de student pour chaque coefficient, il semblerait que nous puissions supprimer certaines de ces interactions. Nous allons donc essayer de réduire notre modèle. Pour cela, nous supposons d'abord le modèle suivant sans interactions :

$$(M_2) \begin{cases} ch4\_t_{ijk} = \mu + \alpha_i + \beta_j + \theta \times nh3\_kg_{ijk} + \gamma \times n2o\_t_{ijk} + \varepsilon_{ijk}, \\ i = 1, \ldots, I = 6,\ j = 1 \ldots, J = 2,\ k = 1, \ldots, n_{ij}. \\ (\varepsilon_{ijk})_{i,j,k} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

Nous effectuons alors un test de sous modèle de Fisher pour voir si nous pouvons enlever les interactions.

Notre p-valeur est très faible, nous ne pouvons pas supprimer toutes les interactions. Nous allons à la place utiliser un algorithme de sélection de variable pour réduire notre modèle.

Nous choissisons la méthode backward avec les critères BIC et AIC.

Les deux algorithmes nous donne le même sous modèle :

$$(M_3) \begin{cases} ch4\_t_{ijk} = \mu + \alpha_i + \beta_j + (a1 + a3_j) \times nh3\_kg_{ijk} + (b1 + b3_j) \times n2o\_t_{ijk} + \nu \times n2o\_t_{ijk} \times nh3\_kg_{ijk} + \varepsilon_{ijk}, \\ i = 1, \ldots, I = 6,\ j = 1 \ldots, J = 2,\ k = 1, \ldots, n_{ij}. \\ (\varepsilon_{ijk})_{i,j,k} \text{ i.i.d } \mathcal{N}(0, \sigma^2) \end{cases}$$

Nous avons supprimé 3 coefficients. Nous vérifions ce resultat avec un test de sous modèle.

```
## Analysis of Variance Table
##
## Model 1: ch4_t ~ annee_inv + nh3_kg + n2o_t + TypeEPCI + nh3_kg:n2o_t +
##     nh3_kg:TypeEPCI + n2o_t:TypeEPCI
## Model 2: ch4_t ~ (annee_inv + nh3_kg + n2o_t + TypeEPCI)^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    960 141.83
## 2    945 140.41 15     1.423 0.6385 0.8444
```

Notre p-valeur est bien supérieur à 0.05 donc on ne rejette pas ce sous modèle.

Nous pourrions alors considérer le modèle final (M_3)

Dépassement d'émission de méthane (Modèle linéaire généralisé)

Dans cette section, nous allons expliquer le dépassement d'émission de méthane ($ch4$)de $1000t$ par an en fonction de l'ammoniac ($nh3$), du protoxyde d'azote ($n2o$), du type d'EPCI et de l'année par un modèle linéaire généralisé.

Nous créons la nouvelle variable booléenne dep_met_1000, valant 1 si le taux de méthane $ch4\_t$ est supérieur à $1000t$, 0 sinon.

La variable dep_met_1000 étant binaire, nous allons utiliser la régression logistique.

Dans un premier temps, nous créons le modèle complet avec interactions, modèle que l'on note (M_{GL_1}) :

$$(\mathrm{M}_{GL_1}) : \begin{cases} dep\_met\_1000_i \sim \mathcal{B}(\pi(x_i)),\ dep\_met\_1000_1,\ ...,\ dep\_met\_1000_n \text{ indépendant.} \\ \\ \text{logit}[\pi(x_i)] = \ln(\dfrac{\pi(x_i)}{1-\pi(x_i)}) = \mu + \theta_1 \cdot nh3_k g_i + \theta_2 \cdot n2o_{ti} + \gamma \cdot nh3_k g_i \cdot n2o_{ti} \\ \qquad\qquad + (\beta_1 + \beta_2 \cdot nh3_k g_i \ + \beta_3 \cdot n2o_{ti}) \mathbf{1}_{TypeEPCI_i = CC} \\ \qquad\qquad + \displaystyle\sum_{k=1}^{5} (\delta_{1k} + \delta_{2k} \cdot nh3_k g_i \ + \delta_{3k} \cdot n2o_{ti}) \mathbf{1}_{annee_i = 2014+k} \\ \qquad\qquad + \displaystyle\sum_{k=1}^{5} (\kappa_{1k} \cdot \mathbf{1}_{\mathrm{TypeEPCI}_i = \mathrm{CC}} \cdot \mathbf{1}_{\mathrm{annee}_i = 2014+k}) \end{cases}$$

Nous observons que certaines variables du modèle complet ont une p-valeur > 0.05. Nous allons nous intéresser à la simplification du modèle. Nous commencons par faire un test de sous-modèle entre le modèle complet et le modèle sans interactions.

Nous obtenons une p-valeur $<< 0.05$. On rejette donc l'hypothèse $\mathcal{H}_0$, et nous devons conserver dans un premier temps le modèle complet, avec toutes les interactions. Nous allons maintenant appliquer des algorithmes de selection de variables pour réduire le modèle complet.

Nous utilisons 3 méthodes d'algorithme de selection de variable en méthode backward et les 3 s'accordent sur le même modèle.

$$(\mathrm{M}_{GL_2}) : \begin{cases} dep\_met\_1000_i \sim \mathcal{B}(\pi(x_i)),\ dep\_met\_1000_1,\ ...,\ dep\_met\_1000_n \text{ indépendant.} \\ \\ \text{logit}[\pi(x_i)] = \ln(\dfrac{\pi(x_i)}{1-\pi(x_i)}) = \mu + \theta_1 \cdot nh3_k g_i + \theta_2 \cdot n2o_{ti} + \gamma \cdot nh3\_kg_i \cdot n2o\_t_i \\ \qquad\qquad + (\beta_1 + \beta_2 \cdot nh3\_kg_i \ + \beta_3 \cdot n2o\_t_i) \mathbf{1}_{TypeEPCI_i = CC} \end{cases}$$

Nous effectuons un test de sous-modèle pour le valider. La p-valeur est de $0.5528 > 0.05$. Nous ne rejetons alors pas le modèle réduit.

```
## Analysis of Deviance Table
##
## Model 1: dep_met_1000 ~ annee_inv + nh3_kg + n2o_t + TypeEPCI + nh3_kg:n2o_t +
##     nh3_kg:TypeEPCI + n2o_t:TypeEPCI
## Model 2: dep_met_1000 ~ (annee_inv + nh3_kg + n2o_t + TypeEPCI)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       960     418.78
## 2       945     405.14 15   13.643   0.5528
```

On a : $T := \mathcal{D}(M_0) - \mathcal{D}(M_1) \sim \chi^2(k_1 - k_0)$, avec $\mathcal{D}(M) = -2\{l(Y, \hat{\theta}) - l(Y, \hat{\theta}_{sat})\}$. Avec les sorties R, on constate : $T^{obs} = 13,643$, puis pvaleur $= \mathbb{P}(T > T^{obs}) = 0.55$.

On a pvaleur > 0.05, on conserve donc le modèle réduit au risque 5%.

```
## [1] 0.6490327
```

Le pseudoR^2 du modèle réduit vaut 0.6490327, c'est une valeur satisfaisante pour ajuster notre modèle.

Conclusion

En conclusion, ce projet d'analyse de données vise à fournir une compréhension globale des émissions de polluants dans la région Occitanie, de 2014 à 2019. Pour y parvenir, nous avons mis en œuvre plusieurs techniques d'analyse de données afin de trouver des informations dans ce jeu de données. Tout d'abord, nous avons effectué des visualisations des données pour mieux comprendre le jeu de données, trouver les points qui doivent être analysés et effectuer les transformations appropriées. Ensuite, nous nous appuyons sur les méthodes de Clustering et de LDA pour classifier les données, avec label et sans label, en fonction des niveaux de pollution et d'émission. Enfin, nous utilisons des modèles linéaires et linéaires généralisés pour trouver des relations entre les émissions et ainsi prédire les niveaux futurs d'émissions de gaz à effet de serre. En résumé, les enseignements tirés de cette analyse peuvent contribuer à une prise de décision éclairée et à des politiques environnementales ciblées dans la région.