Projet d'étude de Statistiques

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

2024-01-29

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.

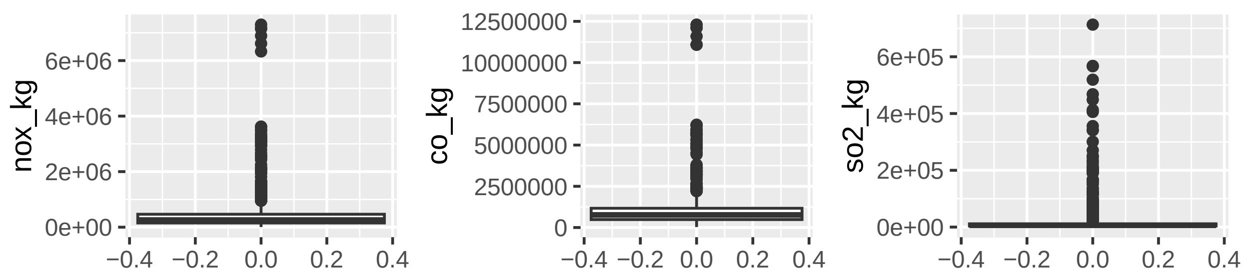

Figure 1: Boxplot des variables nox_kg,co_kg,so2_kg

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.

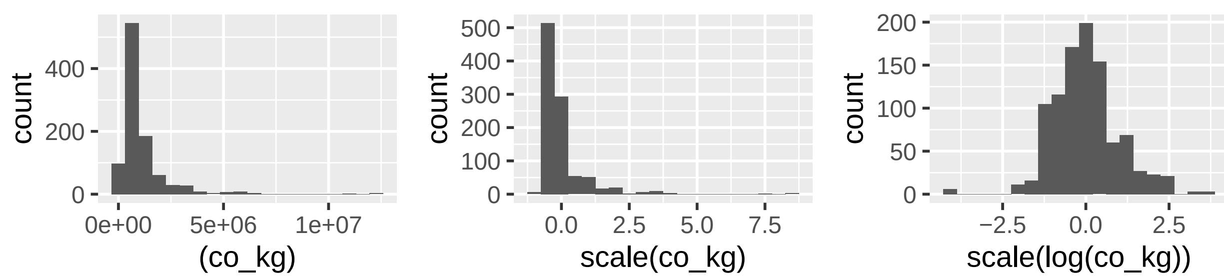

Figure 2: Histogramme de la variable co_kg en brute, scale et scale(log())

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.

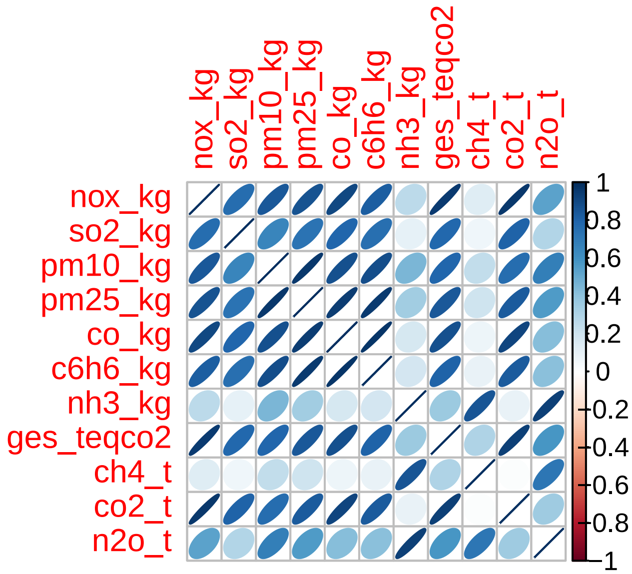

Figure 3: Corrélation entre les variables

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...

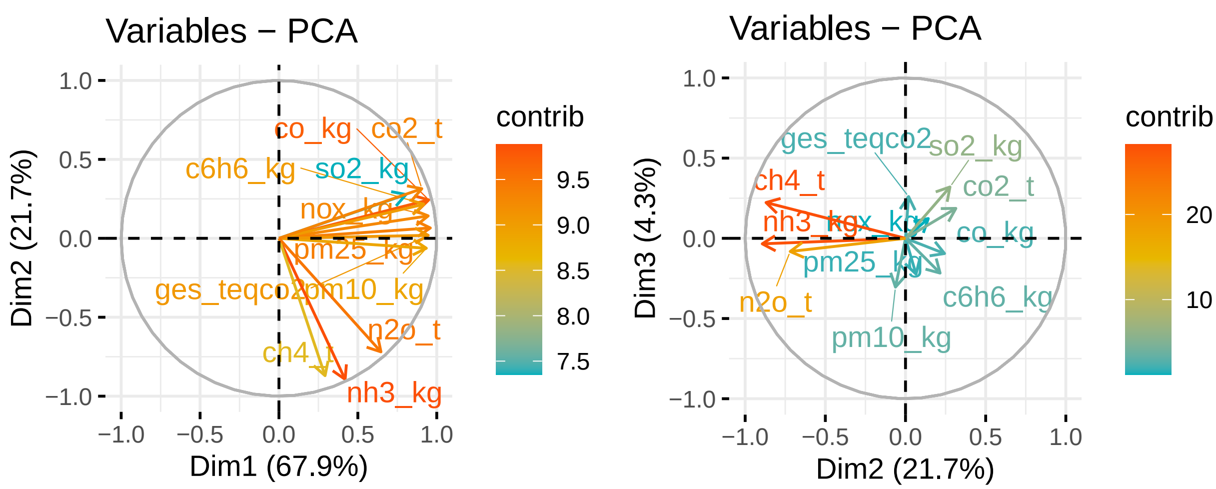

Figure 4: Cercle des corrélations

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

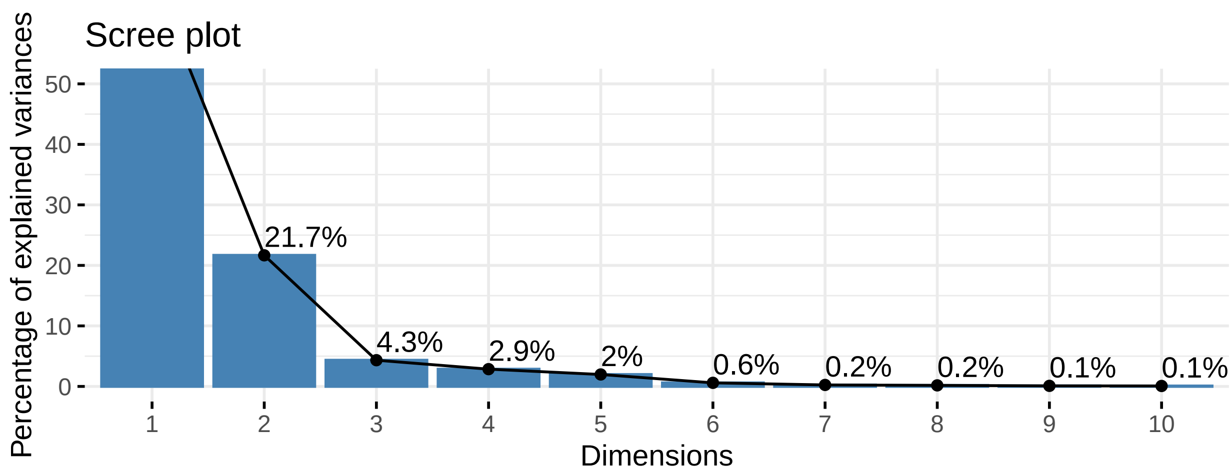

Figure 5: Pourcentage de variance expliquée par chaque axe

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 …

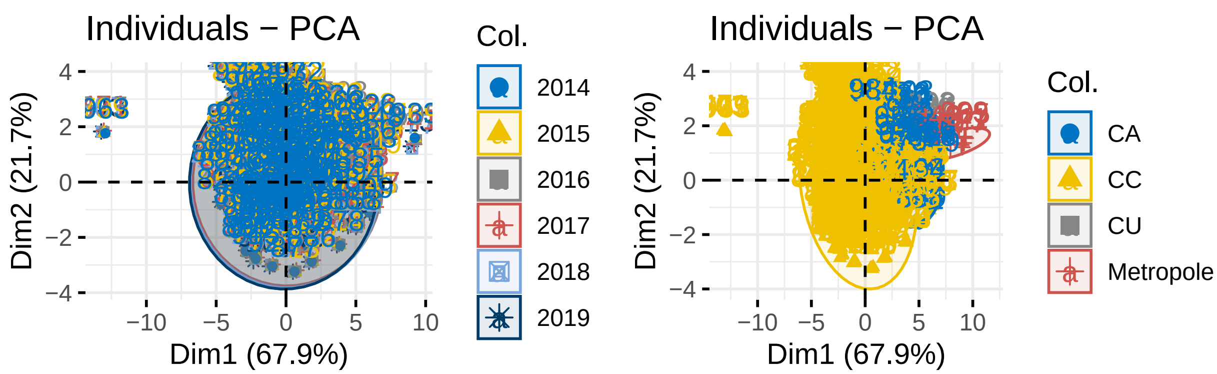

Figure 6: ACP des variables quantitatives

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.

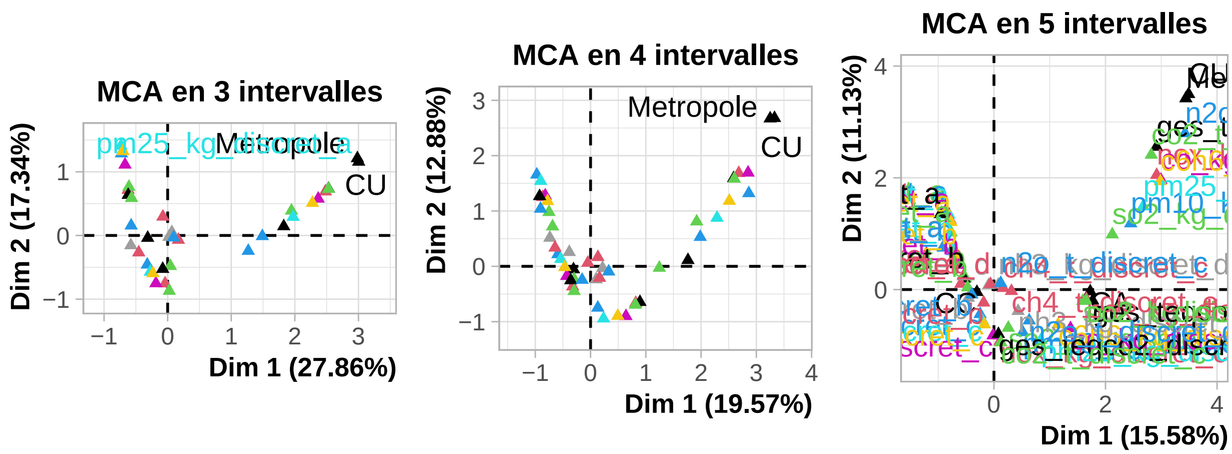

Figure 7: MCA avec découpage des données en 3, 4 et 5 intervalles

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 :

Blabla sur les méthodes de clustering

3.2 Analyse discriminante linéaire

Explication sur la méthode de l'analyse discriminante linéaire

3.2.1 LDA sur le dépassement d'émissions de méthane de 1000 tonnes par an

```
# On sépare les données en train et test (70% train, 30% test)
taille_train=round(0.7*nrow(data_mel2))
d_train=data_mel2[1:taille_train,]
```

```
d_test=data_mel2[taille_train:nrow(data_mel2),]

# On applique la LDA
lda_model <- lda(ch4_t ~ .,data=d_train)
```

```
## Warning in lda.default(x, grouping, ...): les variables sont colinéaires
```

```
# On colorie les points en fonction du dépassement ou non de $1000$ tonnes par an
color2 <- data_lda2$ch4_t ;
color2[color2=="TRUE"] <- "black";
color2[color2=="FALSE"] <- "red"

# Afficher les résultats de la LDA
#print(lda_model)
```

Commentaire sur les résultats obtenus.

3.2.2 LDA sur le type d'EPCI

```
## Call:
## lda(TypeEPCI ~ ., data = d_train)
##
## Prior probabilities of groups:
##          CA          CC          CU   Metropole
## 0.114658926 0.867924528 0.004354136 0.013062409
##
## Group means:
##                nox_kg     so2_kg    pm10_kg   pm25_kg      co_kg    c6h6_kg
## CA          1.2490874  1.0375280  0.9595484  1.077756  1.2616229  1.1498144
## CC         -0.2750418 -0.2231439 -0.2308489 -0.249439 -0.2665294 -0.2447147
## CU          2.2112065  1.7943235  1.9763324  2.179173  2.4785591  2.4388058
## Metropole   3.1511132  2.4307324  2.6158259  2.909730  3.1626560  2.9428386
##                 nh3_kg ges_teqco2       ch4_t      co2_t      n2o_t
## CA          0.08040308  1.2169429 -0.06652089  1.3231960  0.3911695
## CC         -0.04706438 -0.2704604 -0.03108375 -0.2867971 -0.1198924
## CU         -0.51312187  2.3445933 -0.12937263  2.2628365  0.5752615
## Metropole   0.10567936  3.2609723  0.41304985  3.1155983  1.3446043
##
## Coefficients of linear discriminants:
##                   LD1          LD2        LD3
## nox_kg      0.7725267 -1.533221072  1.5351048
## so2_kg      0.2203251  0.008773746  0.1955207
## pm10_kg     0.2051146  0.093440408 -3.9819502
## pm25_kg     0.2757635  0.681823698  5.7191508
## co_kg      -1.3401119 -0.074442145  2.6826379
## c6h6_kg     0.2286017 -0.483877317 -4.6619866
## nh3_kg      2.7931848 -5.863041210 -0.9178086
## ges_teqco2 -0.5713383  0.373287997 -2.9348409
## ch4_t      -0.7586916  1.570970804  0.5066352
## co2_t      -0.4727215 -0.610639436  0.3022481
## n2o_t      -2.3050890  4.907610579  1.8756469
```

```
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.8815 0.1126 0.0058
```

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 :

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

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) :

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

et

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

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