

Projet d'étude de Statistiques

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

2024-01-11

Contents

1	Introduction	1
2	Analyse descriptive des données	3
2.1	Statistiques descriptive des données qualitatives	5
2.2	Analyse des données	7
3	Classification des EPCI	7
3.1	Clustering	7
3.2	Analyse discriminante linéaire	7
4	EMS	7
4.1	Modèle linéaire	7
4.2	Modèle linéaire généralisé	10
5	Conclusion	10

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 <- read.csv("Data-projetmodIA-2324.csv")
```

```
summary(data)
```

```
##      code_epci      lib_epci      annee_inv      nox_kg
## Min.   :200006930 Length:984   Min.   :2014   Min.   :  4085
## 1st Qu.:200066342 Class :character 1st Qu.:2015   1st Qu.: 151025
## Median :200071178 Mode  :character Median :2016   Median : 257531
## Mean   :219957739          Mean   :2016   Mean   : 466703
## 3rd Qu.:243200433          3rd Qu.:2018   3rd Qu.: 465286
## Max.   :248400251          Max.   :2019   Max.   :7296028
```

##	so2_kg	pm10_kg	pm25_kg	co_kg
##	Min. : 91.3	Min. : 2366	Min. : 1971	Min. : 28231
##	1st Qu.: 3694.6	1st Qu.: 56986	1st Qu.: 40502	1st Qu.: 482779
##	Median : 6302.6	Median : 104200	Median : 65338	Median : 764707
##	Mean : 21041.6	Mean : 142992	Mean : 93646	Mean : 1117926
##	3rd Qu.: 11327.3	3rd Qu.: 172210	3rd Qu.:106391	3rd Qu.: 1169215
##	Max. :713262.5	Max. :1186604	Max. :927205	Max. :12302750
##	c6h6_kg	nh3_kg	ges_teqco2	ch4_t
##	Min. : 437.2	Min. : 7408	Min. : 2481	Min. : 30.57
##	1st Qu.: 8497.0	1st Qu.: 83409	1st Qu.: 76282	1st Qu.: 276.97
##	Median : 12599.3	Median : 216921	Median : 118249	Median : 533.78
##	Mean : 17863.6	Mean : 263843	Mean : 209643	Mean : 905.64
##	3rd Qu.: 19764.0	3rd Qu.: 364883	3rd Qu.: 212410	3rd Qu.:1074.44
##	Max. :178741.5	Max. :1202364	Max. :3171595	Max. :5313.16
##	co2_t	n2o_t	TypeEPCI	nomdepart
##	Min. : 865.1	Min. : 0.952	Length:984	Length:984
##	1st Qu.: 31906.3	1st Qu.: 18.017	Class :character	Class :character
##	Median : 59008.8	Median : 41.718	Mode :character	Mode :character
##	Mean : 138412.7	Mean : 48.855		
##	3rd Qu.: 131054.4	3rd Qu.: 65.378		
##	Max. :2678898.2	Max. :726.911		
##	Ardèche	Ariège	Aude	Aveyron
##	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000
##	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000
##	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000
##	Mean :0.006098	Mean :0.04878	Mean :0.06098	Mean :0.122
##	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000
##	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000
##	Gard	Haute.Garonne	Gers	Hérault
##	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000
##	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
##	Median :0.0000	Median :0.0000	Median :0.0000	Median :0.0000
##	Mean :0.1037	Mean :0.1159	Mean :0.1037	Mean :0.1037
##	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.0000
##	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000
##	Landes	Lot	Lot.et.Garonne	Lozère
##	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000
##	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000
##	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000
##	Mean :0.006098	Mean :0.06098	Mean :0.006098	Mean :0.06707
##	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000
##	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000
##	Pyrénées.Atlantiques	Hautes.Pyrénées	Pyrénées.Orientales	Tarn
##	Min. :0.0000	Min. :0.000000	Min. :0.000000	Min. :0.000000
##	1st Qu.:0.0000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000
##	Median :0.0000	Median :0.000000	Median :0.000000	Median :0.000000
##	Mean :0.0122	Mean :0.06098	Mean :0.07317	Mean :0.09756
##	3rd Qu.:0.0000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000
##	Max. :1.0000	Max. :1.000000	Max. :1.000000	Max. :1.000000
##	Tarn.et.Garonne	Vaucluse	latit	longit
##	Min. :0.000000	Min. :0.000000	Min. :42.44	Min. : -0.295
##	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:43.33	1st Qu.: 1.257
##	Median :0.000000	Median :0.000000	Median :43.70	Median : 2.157
##	Mean :0.06098	Mean :0.006098	Mean :43.70	Mean : 2.159

```
## 3rd Qu.:0.00000 3rd Qu.:0.000000 3rd Qu.:44.10 3rd Qu.: 3.034
## Max. :1.00000 Max. :1.000000 Max. :44.88 Max. : 4.825
```

```
data[1,]
```

```
## code_epci lib_epci annee_inv nox_kg so2_kg pm10_kg pm25_kg
## 1 200006930 CC du Haut Allier 2019 65633.66 3866.599 15728.87 10975.55
## co_kg c6h6_kg nh3_kg ges_teqco2 ch4_t co2_t n2o_t TypeEPCI
## 1 173194.3 2319.199 133686.2 43995.12 617.104 17831.59 17.114 CC
## nomdepart Ardèche Ariège Aude Aveyron Gard Haute.Garonne Gers Hérault Landes
## 1 Lozère 0 0 0 0 0 0 0 0 0 0
## Lot Lot.et.Garonne Lozère Pyrénées.Atlantiques Hautes.Pyrénées
## 1 0 0 1 0 0
## Pyrénées.Orientales Tarn Tarn.et.Garonne Vaucluse latit longit
## 1 0 0 0 0 44.7324 3.769267
```

- Notations :

2 Analyse descriptive des données

Dans un premier temps on extrait les données du fichier Data-projetmodIA-2324.csv. Puis on extrait les données quantitatives de ce jeu de données (les quantités de gaz). Puis on visualise globalement les données quantitatives brutes.

```
data_quant=data[,c("nox_kg","so2_kg","pm10_kg","pm25_kg","co_kg","c6h6_kg","nh3_kg","ges_teqco2","ch4_t")
data_quant=as.data.frame(data_quant)
head(data_quant)
```

```
## nox_kg so2_kg pm10_kg pm25_kg co_kg c6h6_kg nh3_kg
## 1 65633.66 3866.599 15728.87 10975.55 173194.3 2319.199 133686.18
## 2 310288.20 8083.028 50929.20 38591.71 593036.6 8349.081 114533.40
## 3 337655.55 9373.106 143623.67 82143.61 1275976.8 18806.497 71177.45
## 4 298100.30 4091.852 126735.60 63331.88 780230.6 12250.430 244266.82
## 5 447186.53 13650.148 143525.58 111854.31 1386798.2 21346.289 130426.82
## 6 2110865.51 57993.192 506888.15 353513.88 5270166.1 78510.229 111010.72
## ges_teqco2 ch4_t co2_t n2o_t
## 1 43995.12 617.104 17831.59 17.114
## 2 127777.47 436.445 93016.70 19.755
## 3 161136.84 251.623 125004.72 17.606
## 4 116802.18 275.749 79458.03 50.976
## 5 216301.79 447.677 161747.97 24.640
## 6 1057760.61 398.561 806673.97 59.760
```

```
g1=ggplot(data_quant)+geom_boxplot(aes(y = nox_kg))
g2=ggplot(data_quant)+geom_boxplot(aes(y = co_kg))
g3=ggplot(data_quant)+geom_boxplot(aes(y =so2_kg ))
grid.arrange(g1,g2,g3,ncol=3)
```

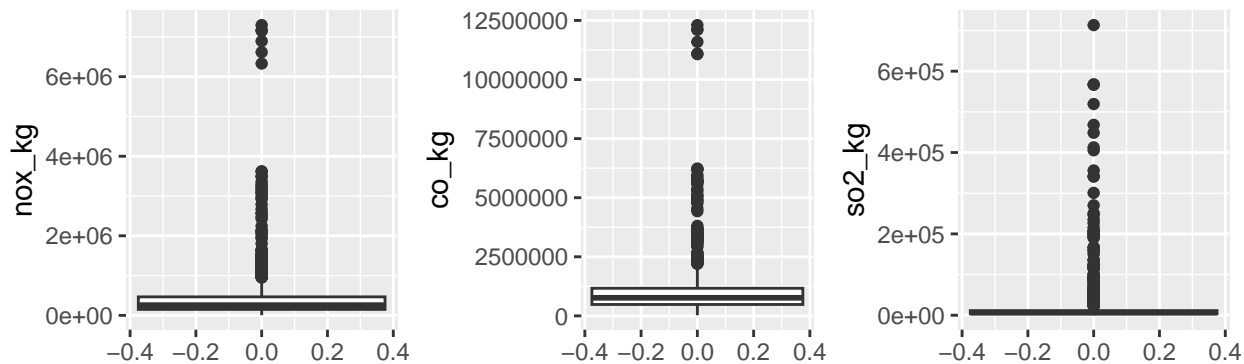


Figure 1: Boxplot des variables nox_kg,co_kg,so2_kg

```
g1=ggplot(data_quant)+ geom_histogram(aes(x = (co_kg)),bins =20 )
g2=ggplot(data_quant)+ geom_histogram(aes(x = scale(co_kg)),bins =20)
g3=ggplot(data_quant)+ geom_histogram(aes(x = scale(log(co_kg))),bins =20)
grid.arrange(g1,g2,g3,ncol=3)
```

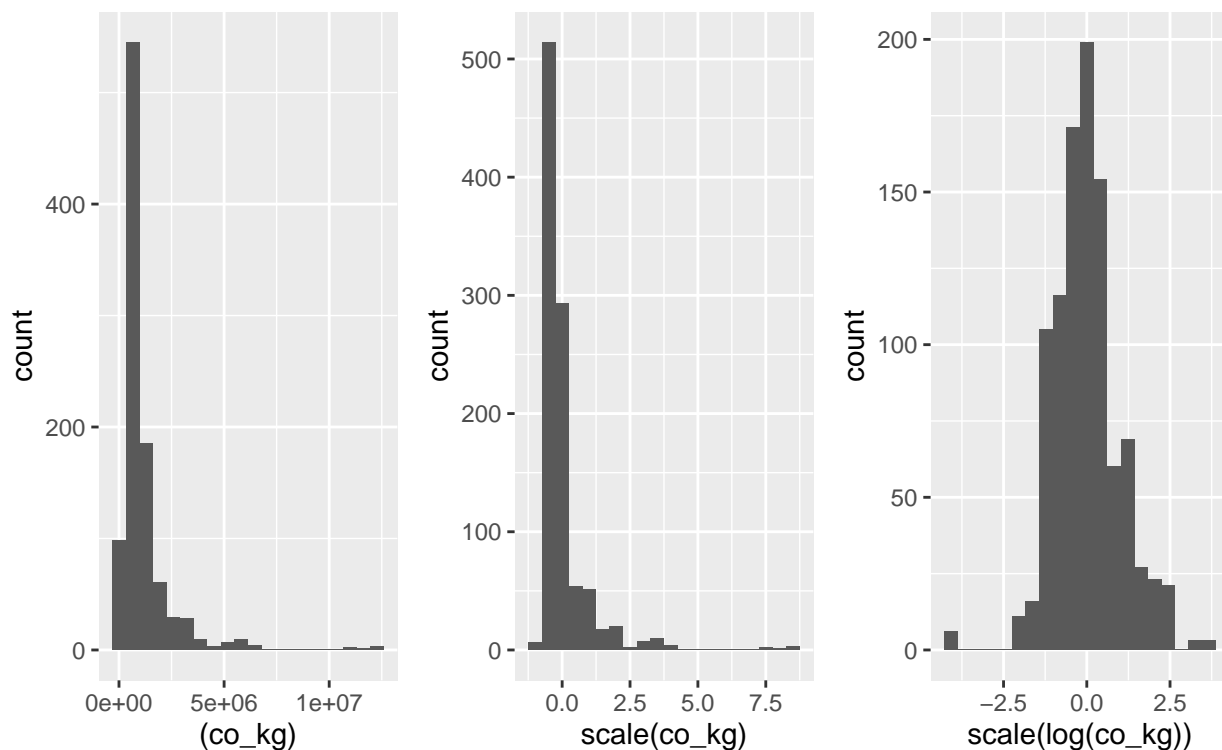


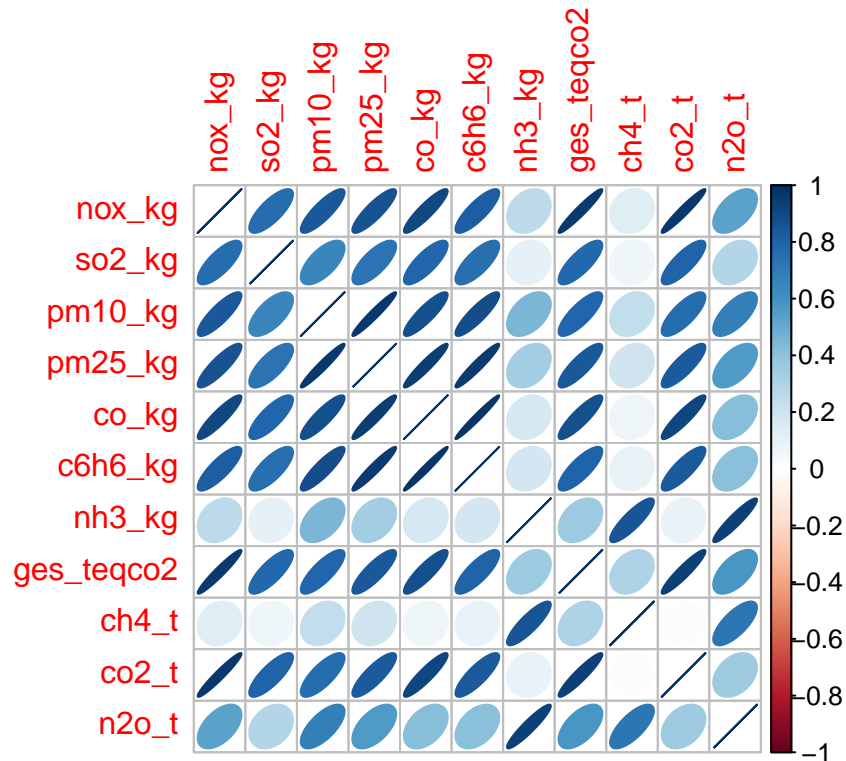
Figure 2: Histogramme de la variable co_kg en brute, scale et scale(log())

On effectue une transformation des données car d'après les boxplots de la figure ?? on remarque une variance énorme de certaines données comme co_kg. En examinant l'historgramme des données quantitatives, on observe une distribution fortement asymétrique. On peut donc appliquer une log-transformation pour normaliser la distribution des données. Certaines variables ont pour unité la tonne et d'autre le kg on peut donc scale les données. On peut visualiser l'interet de ces transformations grâce à la figure ?? avec la variable

co_kg. Par la suite on scale log les données de la manière suivante.

```
data_quant_scaled <- scale(log(data_quant))
data_scaled_df <- as.data.frame(data_quant_scaled)
```

```
mat_cor <- cor(data_scaled_df)
corrplot(mat_cor,method="ellipse")
```



La visualisation de la figure ?? 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.

2.1 Statistiques descriptive des données qualitatives

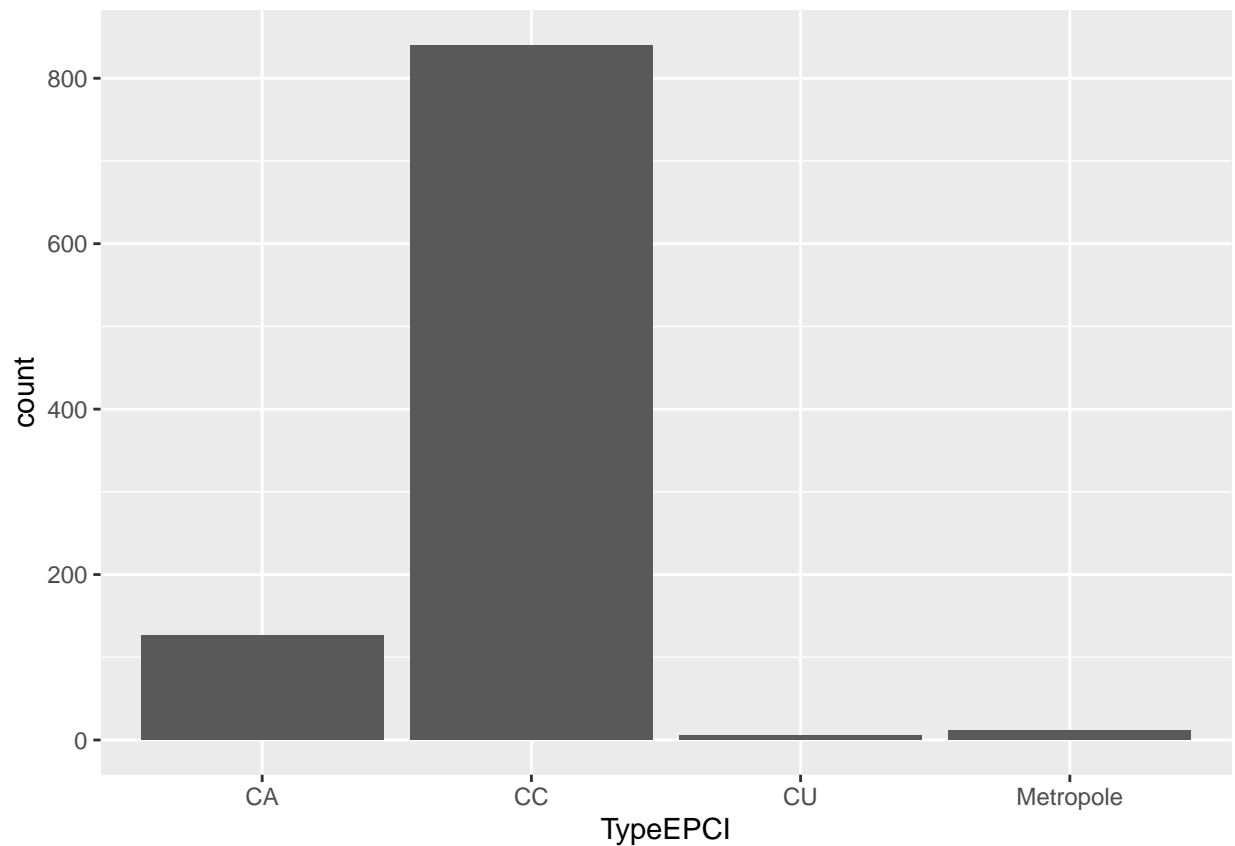
Dans le jeu de données nous avons aussi des variables qualitatives comme le code epci, le lib_epci ou des infos sur les départements.

```
data_quali=data[,c("code_epci","lib_epci","annee_inv","TypeEPCI","nomdepart")]
table(data_quali[,c("nomdepart")])
```

```
##
##                      Ardèche,Gard                      Ariège
##                      6                      48
##                      Aude                      Aude,Haute-Garonne,Tarn
##                      48                      6
##                      Aude,Pyrénées-Orientales                      Aveyron
##                      6                      102
##                      Aveyron,Lot                      Aveyron,Lozère
```

```
##          12          6
##          Gard          Gard,Hérault
##          78          6
##          Gard,Lozère          Gard,Vaucluse
##          6          6
##          Gers          Gers,Haute-Garonne
##          84          6
##          Gers,Landes Gers,Lot-et-Garonne,Tarn-et-Garonne
##          6          6
##          Haute-Garonne          Haute-Garonne,Tarn
##          96          6
##          Hautes-Pyrénées Hautes-Pyrénées,Pyrénées-Atlantiques
##          48          12
##          Hérault          Hérault,Tarn
##          90          6
##          Lot          Lozère
##          48          54
##          Pyrénées-Orientales          Tarn
##          66          72
##          Tarn-et-Garonne          Tarn,Tarn-et-Garonne
##          48          6
```

```
ggplot(data=data_quali)+geom_bar(aes(x = TypeEPCI))
```



2.2 Analyse des données

2.2.1 PCA

On visualise les individus à partir des émissions de polluants.

2.2.2 Réduction de dimension (MCA)

3 Classification des EPCI

3.1 Clustering

3.2 Analyse discriminante linéaire

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 :

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

```
dlog=data[4:15]
data_quant=scale(log(data[4:14]))
dlog[1:11]=data_quant
dlog=data.frame(dlog,annee_inv=data$annee_inv)

anov2= lm(ges_teqco2 ~TypeEPCI * annee_inv, data=dlog)
summary(anov2)
```

```
##
## 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.

```
anov_sans_int=lm(ges_teqco2 ~TypeEPCI + annee_inv, data=dlog)
anova(anov_sans_int,anov2)
```

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ TypeEPCI + annee_inv
## Model 2: ges_teqco2 ~ TypeEPCI * annee_inv
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      979 570.31
## 2      976 570.30   3  0.0085333 0.0049 0.9995
```

On obtient une p-value de 1 > 0.05.

On peut donc enlever les interactions :

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

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

```
anov_annee=lm(ges_teqco2 ~annee_inv, data=dlog)
anov_type=lm(ges_teqco2 ~TypeEPCI, data=dlog)
anova(anov_annee,anov_sans_int)
```

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ annee_inv
## Model 2: ges_teqco2 ~ TypeEPCI + annee_inv
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      982 982.84
## 2      979 570.31   3    412.53 236.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(anov_type,anov_sans_int)
```

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ TypeEPCI
## Model 2: ges_teqco2 ~ TypeEPCI + annee_inv
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      980 570.47
## 2      979 570.31   1    0.16143 0.2771 0.5987
```


Pour le modèle dépendant uniquement du type d'EPCI, on obtient une p-value de $0.6 > 0.05$.
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.

```
anova(lm(ges_teqco2 ~ 1, data=dlog), anov_type)
```

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ 1
## Model 2: ges_teqco2 ~ TypeEPCI
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     983 983.00
## 2     980 570.47  3    412.53 236.23 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 :

```
anova(anov_type, anov2)
```

```
## Analysis of Variance Table
##
## Model 1: ges_teqco2 ~ TypeEPCI
## Model 2: ges_teqco2 ~ TypeEPCI * annee_inv
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     980 570.47
## 2     976 570.30  4    0.16996 0.0727 0.9904
```

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

```
summary(anov_type)
```

```
##
## Call:
## lm(formula = ges_teqco2 ~ TypeEPCI, data = dlog)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.2233 -0.4145 -0.0369  0.3839  2.8504
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.28521    0.06797  18.909 < 2e-16 ***
## TypeEPCICC     -1.53937    0.07289 -21.119 < 2e-16 ***
## TypeEPCICU      1.05473    0.31881   3.308 0.000973 ***
## TypeEPCIMetropole 1.84123    0.23050   7.988 3.83e-15 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.763 on 980 degrees of freedom
## Multiple R-squared:  0.4197, Adjusted R-squared:  0.4179
## F-statistic: 236.2 on 3 and 980 DF,  p-value: < 2.2e-16
```

4.1.2 Régression linéaire

4.1.3 ANCOVA

4.2 Modèle linéaire généralisé

5 Conclusion