

PROJET RÉALISÉ PAR L'ÉQUIPE 2
RAPPORT DE GROUPE EN SCIENCES DES
DONNÉES 2 + BASES DE DONNÉES

Randriamisanta Fehizoro, Abdraman Mahamat, LE Quentin, LE Maxime, Jad El
Hage



Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Décembre 2022

SOUmis COMME CONTRIBUTION PARTIELLE
POUR LE COURS SCIENCE DES DONNÉES 2 ET BASES DE DONNÉES

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature: _____ Date: _____

Remerciements

Nos plus sincères remerciements vont à notre encadrant pédagogique pour les conseils avisés sur notre travail.

12/09/2022.

Résumé

Nous avons décidé de ne pas faire de résumé pour ce rapport.

Table des matières

Chapitre 1 Introduction	1
Chapitre 2 Base de données	2
2.1 Descriptif des tables	2
2.2 Modèles MCD et MOD	3
2.3 Import des données	3
2.4 Requêtes réalisées	4
Chapitre 3 Matériel et Méthodes	7
3.1 Logiciels	7
3.2 Modélisation statistique	7
Chapitre 4 Analyse et résultats	8
4.1 Echelle mondiale	8
4.2 Echelle locale (communes France)	11
Chapitre 5 Conclusion et perspectives	14
Bibliographie	15
Annexes	16
Codes	16

CHAPITRE 1

Introduction

Le monde subit une croissance démographique depuis le dernier siècle, et la contamination de l'environnement a souillé notre planète.

Cela signifie-t-il qu'il y a une corrélation entre l'accroissement de la population et la pollution ?

A travers nos bases de données, nous suivrons une analyse statistique de la pollution en partant d'une échelle globale, en étudiant l'émission de gaz à effets de serre totale par pays sur plusieurs années. On triera nos données pour montrer ceux qui jouent le plus grand rôle dans l'émission de ces gaz . Ensuite on se lancera vers une analyse locale en posant dans notre spectre la France afin d'analyser la répartition des gaz en fonction des secteurs (agriculture, tertiaire,...) et en fonction du nombre d'habitants

CHAPITRE 2

Base de données

2.1 Descriptif des tables

- **Emissions de gaz à effet de serre par commune :** Sur les 12 colonnes de la table, nous en avons gardé 10. On a supprimé les colonnes « Autres transports » et « Autres transports internationaux » car il y a trop de lignes vides pour ces colonnes. Nombre de colonnes : 10 Nombre de lignes : 35779 Description des colonnes (type) :
 - Id_commune : c'est la clé (code officiel géographique par l'INSEE, de type entier)
 - Nom_commune : caractères (nom de chaque commune)
 - Agriculture, biomasse_hors_total,dechets,industrie_energie,industrie_hors_energie, residence, routier, tertiaire : numériques (Les valeurs sont exprimées en tonne d'équivalent CO2)

Lien: [GES par commune](#)

- **Nombre d'habitants par commune :** Sur les 10 colonnes de la table, nous en avons gardé 3 (nom_commune, id_commune et population_totale). Toutes les lignes ont été gardées pour garantir la cohérence des différentes tables. Nombre de colonnes : 10 Nombre de lignes : 34996 Description des colonnes (type) :
 - Id_commune : c'est la clé (code officiel géographique par l'INSEE, de type entier)
 - Nom_commune : caractères (nom de chaque commune)
 - Population_totale : entier (nombre d'habitants pour chaque commune)

Lien: [Nombre d'habitants par commune](#)

- **Evolution de l'émission de GES par pays :** Sur les 15 colonnes de la table, nous en avons gardé 5. Nous avons trié la colonne « Unit » par « tonnes d'équivalent CO2 » puis nous avons supprimé cette dernière. Dans la colonne « pays », il y a d'autres enregistrements comme « Afrique » ou encore « Union européenne ». Nous allons les supprimer sur pandas en utilisant la fonction « merge ». Nombre de colonnes : 4 Nombre de lignes : 31855 Lien: [Emission de GES par pays](#)

Description des colonnes (type) : - Code_pays : caractères composés de 3 lettres, unique pour chaque pays (clé) - Nom_pays : caractères - Annee : entier - Value : numériques (Les valeurs sont exprimées en tonne d'équivalent CO2). C'est la quantité de GES émise par un pays

- **Population par pays et par date :** Pour cette table, nous avons gardés toutes les colonnes et toutes les lignes. Nous n'allons supprimer des lignes que lors de la création du MOD, en supprimant les enregistrements qui ne sont pas des « pays ». Nombre de colonnes : 3 Nombre de lignes : 267 Description des colonnes (type) :
 - Code_pays : caractères composés de 3 lettres, unique pour chaque pays (clé)
 - Nom_pays : caractères
 - Année : entier Lien: [Population par pays](#)

2.2 Modèles MCD et MOD

Voici le modèle conceptuel et organisationnel de nos données

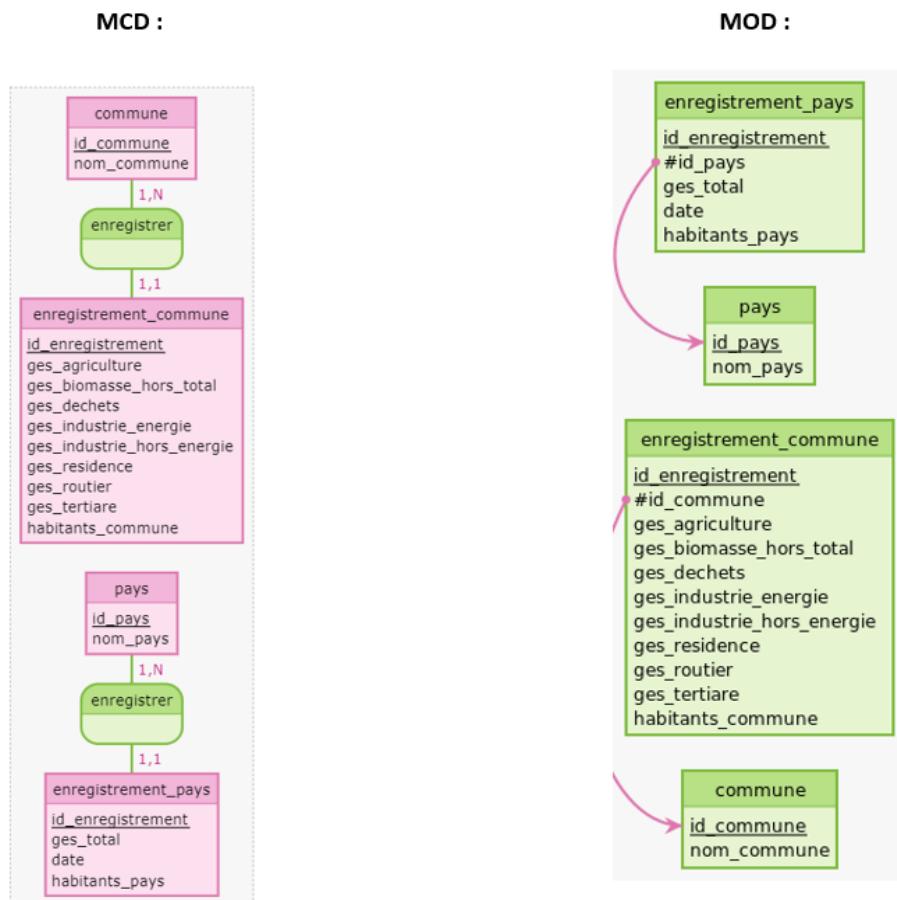


Figure 2.1: Relations.

2.3 Import des données

Avant l'import des données sur sql, on a trié les colonnes selon nos besoins (on n'a gardé que “GES” par exemple et on a éliminé les lignes avec “méthane”). On avait des données par secteur pour les pays mais on a décidé, pour chaque année allant de 1990 à 2018, de faire la totale de GES par pays en regroupant les données par date. Notons que tous les pays n'ont pas de données pour chaque date mais on a décidé

de les garder et faire la moyenne des GES par pays pour une date donnée si on veut faire des analyses (on ne fait donc pas la somme de GES de 1990 à 2018 mais la moyenne). On a vérifié le type de variable de chaque attribut de chaque table et on a utilisé l'encodage UTF-8 pour chacun des fichiers csv qu'on a créés. On uniformisé les valeurs de GES total par pays en ne gardant pas de chiffres après la virgule. Pour les communes, on a gardé trois chiffres après la virgule pour les GES par secteur. On a remplacé les id_commune de *l'Insee* par des entiers ordonnés pour faciliter les futures jointures. On a aussi supprimé les doublons en utilisant pandas sur python.

2.4 Requêtes réalisées

On commence par appeler la base de données sur php myadmin

```
require(RMySQL)
con <- dbConnect(RMySQL::MySQL(), host = "localhost", dbname="pollution", user = "root", password = "")
```

1- Quelle est la commune qui émet le plus de GES ?

```
select commune.nom_commune,enregistrement_commune.ges_total
from commune,enregistrement_commune
where commune.id_commune=enregistrement_commune.id_enregistrement and enregistrement_commune.ges_total=(select max(ges_total)
from enregistrement_commune)
```

Table 2.1: 1 records

nom_commune	ges_total
Fos-sur-Mer	9514690

2- Quelle est la commune qui émet le moins de GES ?

```
select commune.nom_commune,enregistrement_commune.ges_total
from commune,enregistrement_commune
where commune.id_commune=enregistrement_commune.id_enregistrement and enregistrement_commune.ges_total=(select min(ges_total)
from enregistrement_commune)
```

Table 2.2: 1 records

nom_commune	ges_total
Charmes-en-l'Angle	95.337

3- Quelle est l'année où il y a le moins de GES en moyenne ?

```
SELECT enregistrement_pays.habitants_par_pays as annee, AVG(enregistrement_pays.ges_total) as sm
FROM enregistrement_pays
GROUP BY enregistrement_pays.habitants_par_pays
ORDER BY sm
LIMIT 1
```

Table 2.3: 1 records

annee	sm
38392	83.607

4- Quels sont les 5 pays les plus pollueurs ?

```
SELECT pays.nom_pays, AVG(enregistrement_pays.ges_total) as ges_moyenne
FROM pays,enregistrement_pays
WHERE pays.id_pays=enregistrement_pays.id_pays
GROUP BY enregistrement_pays.id_pays
ORDER BY ges_moyenne DESC
LIMIT 5
```

Table 2.4: 5 records

nom_pays	ges_moyenne
Chine	9362383
États-Unis	6932452
Fédération de Russie	2104303
Inde	1928566
Japon	1339262

5- Quelle est le pourcentage de chaque secteur par rapport au total de tous les secteurs ?

```

SELECT ((SUM(enregistrement_commune.ges_agriculture)/(SELECT SUM(enregistrement_commune.ges_total) FROM enregistrement_commune)*100)
as industrie_hors_energie,(SUM(enregistrement_commune.ges_residence)/(SELECT SUM(enregistrement_commune.ges_total)

```

NB: On a remarqué que latex considère le symbole “/” suivi de certains caractères comme des caractères spéciaux. C'est pour cela qu'on n'a pas évalué ce code.

moyenne ?

```
SELECT pays.nom_pays, AVG(enregistrement_pays.ges_total) as ges_moyenne
FROM pays,enregistrement_pays
WHERE pays.id_pays=enregistrement_pays.id_pays
GROUP BY enregistrement_pays.id_pays
ORDER BY ges_moyenne DESC
LIMIT 5
```

Table 2.5: 5 records

nom_pays	ges_moyenne
Chine	9362383
États-Unis	6932452
Fédération de Russie	2104303
Inde	1928566
Japon	1339262

7- Quels sont les 5 pays les plus peuplés en moyenne ?

```
SELECT pays.nom_pays, AVG(enregistrement_pays.habitants_par_pays) as nb_population
FROM pays,enregistrement_pays
WHERE pays.id_pays=enregistrement_pays.id_pays
GROUP BY enregistrement_pays.id_pays
ORDER BY nb_population DESC
LIMIT 5
```

Table 2.6: 5 records

nom_pays	nb_population
Chine	1311862000
Inde	1140243947
États-Unis	291519126
Indonésie	224988870
Brésil	180065604

On peut voir que deux des pays les plus peuplés (Etats-Unis et Chine) font aussi partie du top 5 des pays les plus pollueurs mais on ne peut rien conclure pour l'instant.

8- En quelle année l'émission du ges est le plus haut ?

```
SELECT enregistrement_pays.habitants_par_pays as annee, AVG(enregistrement_pays.ges_total) as sm
FROM enregistrement_pays
GROUP BY enregistrement_pays.habitants_par_pays
ORDER BY sm DESC
LIMIT 1
```

Table 2.7: 1 records

annee	sm
1371860000	12300200

9- En quelle année l'émission du ges et le plus bas ?

```
SELECT enregistrement_pays.habitants_par_pays as annee, AVG(enregistrement_pays.ges_total) as sm
FROM enregistrement_pays
GROUP BY enregistrement_pays.habitants_par_pays
ORDER BY sm
LIMIT 1
```

Table 2.8: 1 records

annee	sm
38392	83.607

10- quelle était la valeur de l'émission du pays plus gros polluer pendant l'année où le ges était le plus haut ?

```
SELECT pays.nom_pays, AVG(enregistrement_pays.ges_total) as ges_moyenne
FROM pays,enregistrement_pays
WHERE pays.id_pays=enregistrement_pays.id_pays AND enregistrement_pays.habitants_par_pays=(SELECT enregistrement_pays.habitants_par_pays
FROM enregistrement_pays
GROUP BY enregistrement_pays.habitants_par_pays
ORDER BY AVG(enregistrement_pays.ges_total) DESC
LIMIT 1)
GROUP BY enregistrement_pays.id_pays
ORDER BY ges_moyenne DESC
LIMIT 1
```

Table 2.9: 1 records

nom_pays	ges_moyenne
Chine	12300200

CHAPITRE 3

Matériel et Méthodes

3.1 Logiciels

On a utilisé Excel pour l'uniformisation des données et le filtrage des colonnes. La bibliothèque Pandas de python a été d'une grande aide pour le nettoyage complet des données et les jointures des tables car on a croisé plusieurs jeux de données . R a été utilisé pour mettre en relation tous les logiciels et pour produire le rapport final. Voici les informations sur les versions des logiciels et sur l'ordinateur qui a servi pour les analyses.

ordinateur: - système d'exploitation: Windows

- modèle: VivoBook_ASUSLaptop X421DA_D413DA
- version: 10.0.19044 Build 19044
- processeur: AMD Ryzen 5 3500U with Radeon Vega Mobile Gfx, 2100MHz, 4 cœur(s), 8 processeur(s) logique(s)
- RAM: 8 Go

Python:

- version: 3.9.12

R: - version: 4.2.1 (2022-06-23 ucrt)

- mode: desktop

3.2 Modélisation statistique

Le principal outil statistique que nous avons utilisé est le coefficient de corrélation de Pearson. Cet outil nous a permis de déterminer l'intensité de l'association linéaire entre la variable "nombre d'habitants" et "quantité de gaz à effet de serre émis. La condition pour pouvoir l'utiliser, c'est que les deux variables doivent avoir une variance non nulle (pour éviter la division par zéro). Rappelons que plus ce coefficient est proche de -1 ou +1, plus l'association entre les deux variables est forte. La corrélation est généralement exprimé avec un autre chiffre que l'on nomme p. En fait, avant de d'interpréter les résultats, on suppose qu'il n'existe pas de relation linéaire entre les deux variables (hypothèse nulle). La valeur p est la probabilité d'observer un coefficient de corrélation différent de zéro lorsqu'on fait l'hypothèse nulle est vraie. Si p est faible, on pourrait rejeter l'hypothèse nulle. En général, le seuil de rejet d'une hypothèse nulle est une valeur p de 0,05. Pour que le coefficient de corrélation puisse être interprété, la valeur de p doit être plus petite que 0,05. La limite du coefficient de corrélation est qu'il ne peut résumer la qualité d'une régression multiple. Il n'est pas robuste car il est très sensible aux valeurs aberrantes.

CHAPITRE 4

Analyse et résultats

4.1 Echelle mondiale

Premièrement, on va voir un choropleth pour avoir une vision globale de l'émission de GES dans le monde en 2014 car c'est la date à laquelle on a le plus de données. Voici une partie du code:

```
choropletch  
  ggplot(carte_ges,aes(long,lat,group=group,fill=ges_total))+geom_polygon(color="black")+coord_fixed()+ coord_map
```

```
choropletch
```

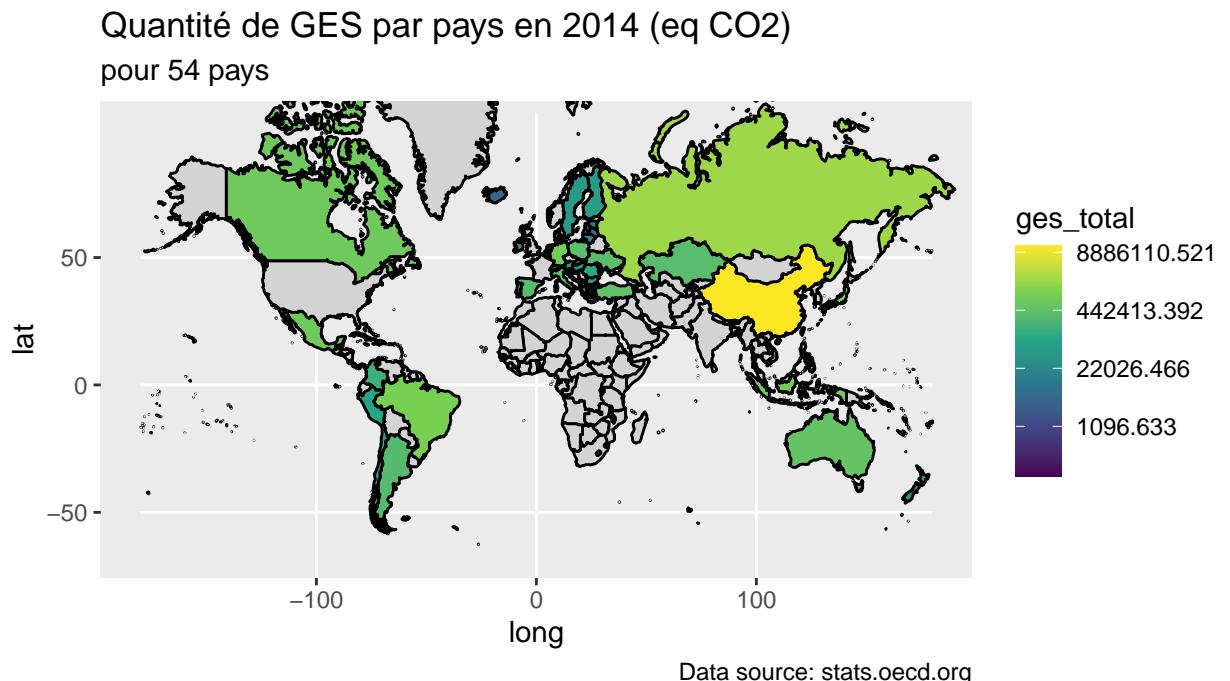
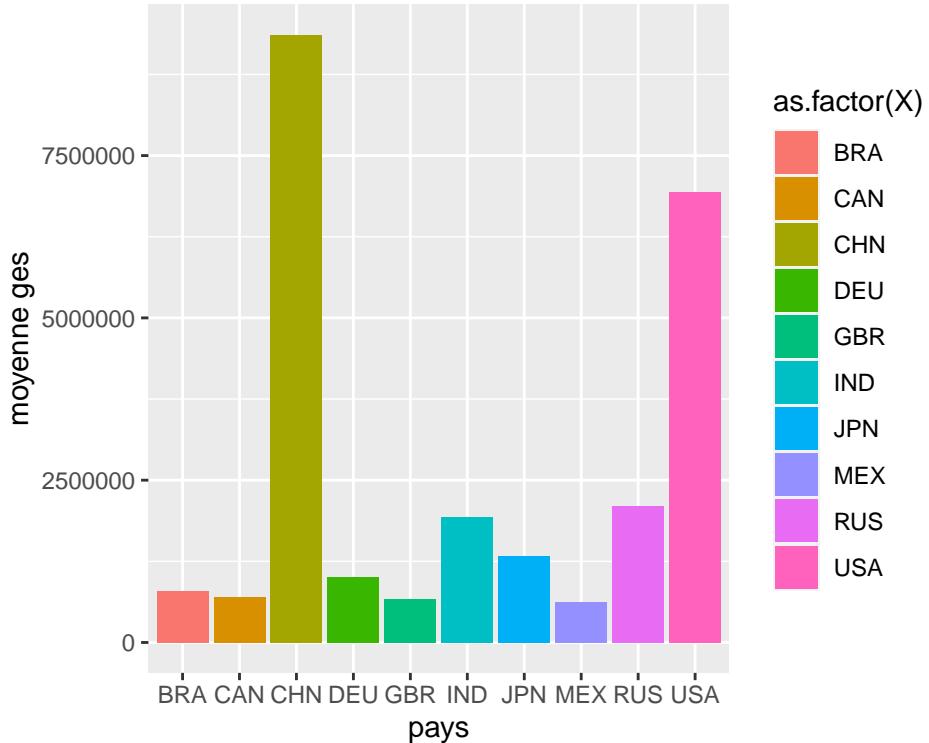


Figure 4.1: GES sur un choropleth.

Dans carte et avec les données qu'on a, on peut voir que c'est l'Islande qui émet le moins de GES et que la chine émet le plus. Une conclusion trop hative serait de

dire les grands pays émettent plus de GES que les petits pays. Mais vu l'absence de données, on s'appuiera plutôt sur un diagramme pour mieux interpréter les résultats.



La Chine et les États Unis se démarquent et occupent la première et deuxième position, ces deux pays sont les plus gros polluant.

Pour avoir une meilleure idée de la répartition, on va aussi analyser les données de 2018 avec des calculs et des boxplots.

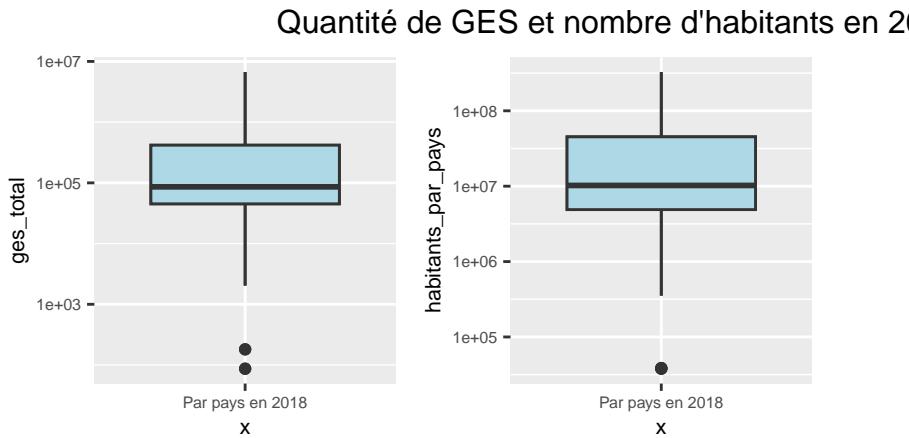
voici un résumé statistique de GES émis par tous les pays en 2018, c'est à dire l'année la plus récente pour laquelle on a des données.

```
enregistrement_pays<-dbReadTable(conn =con, "enregistrement_pays")

ges_2018<-subset.data.frame(enregistrement_pays, enregistrement_pays$année == 2018)
summary(ges_2018["ges_total"])

##      ges_total
##  Min.   :    87
##  1st Qu.:  45213
##  Median :  86198
##  Mean   : 395629
##  3rd Qu.: 417252
##  Max.   :6687510
```

Voici le boxplot pour visualiser ces données.



On va faire un graphe de corrélation pour représenter l'émission moyenne de GES par pays et la population moyenne par pays, groupé par date

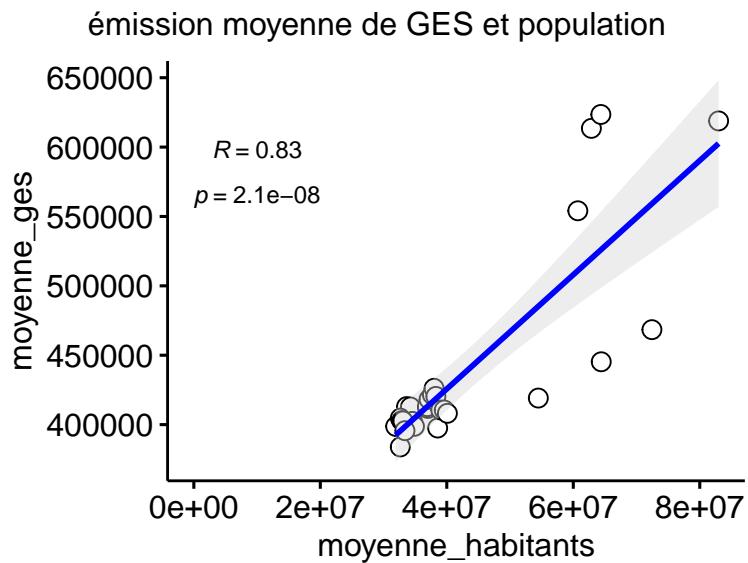


Figure 4.2: Nuage de points: émission moyenne de GES et population moyenne.

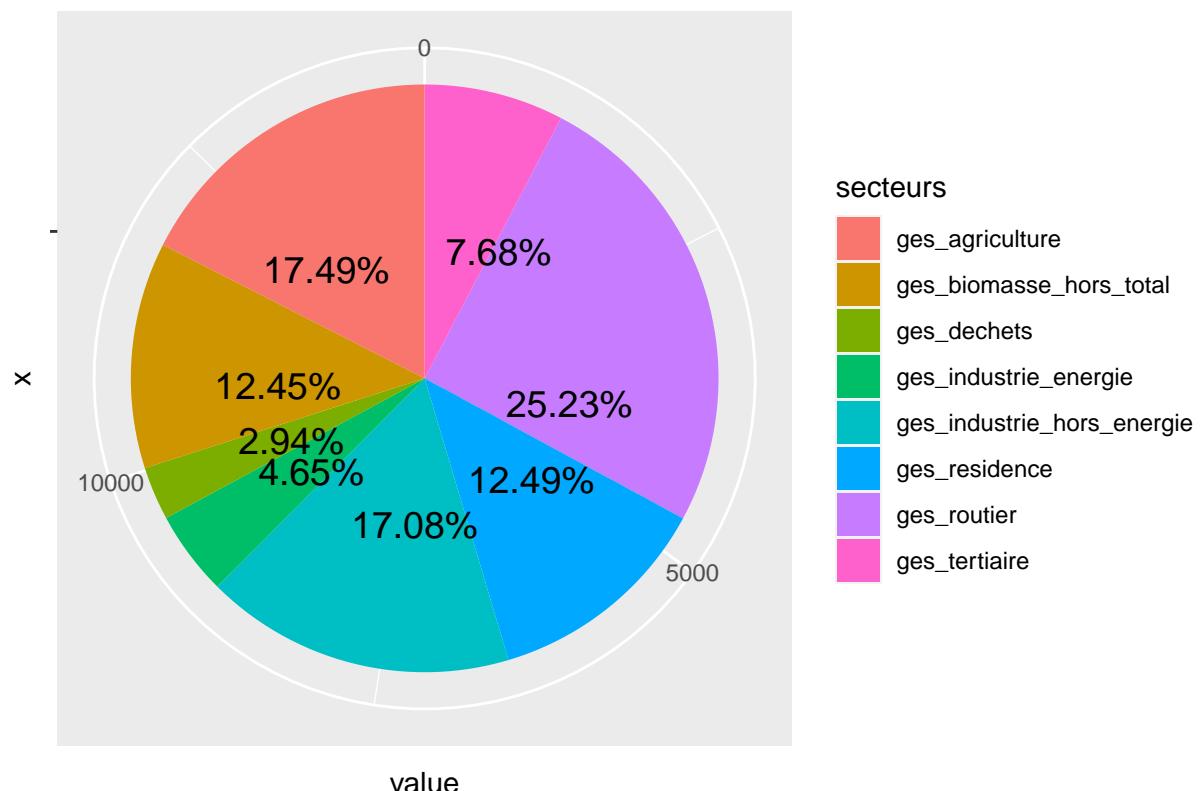
Au niveau mondial, on a une forte corrélation positive entre le nombre d'habitants et l'émission de GES. De plus la probabilité qu'il n'y ait pas de corrélation est très

faible ($2.1 \cdot 10^{-8}$). Donc on peut conclure que plus le nombre d'habitants augmente, plus l'émission de GES augmente (donc plus on pollue).

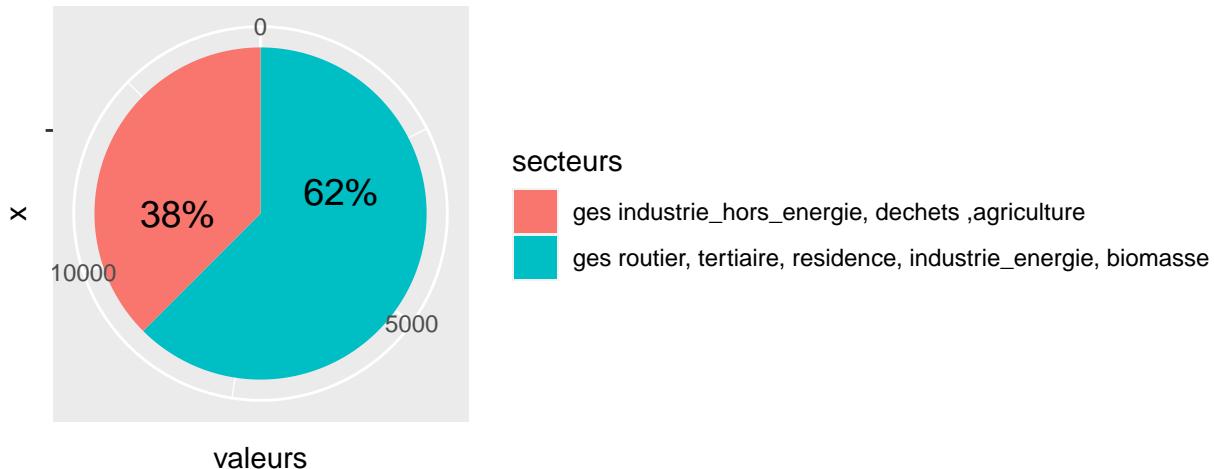
4.2 Echelle locale (communes France)

Passons aux données sur les communes. Nous avons réalisé plusieurs graphiques. Notons que les codes sont fournis dans les annexes. Pour commencer on représentera la répartition des différentes émissions de gaz en fonction de leurs secteurs respectifs à travers un graphe en camembert qui optimisera la visualisation.

Camembert représentant la part d'émissions de chaque secteur en France



Camembert représentant la pourcentage d'émissions de groupe de secteurs



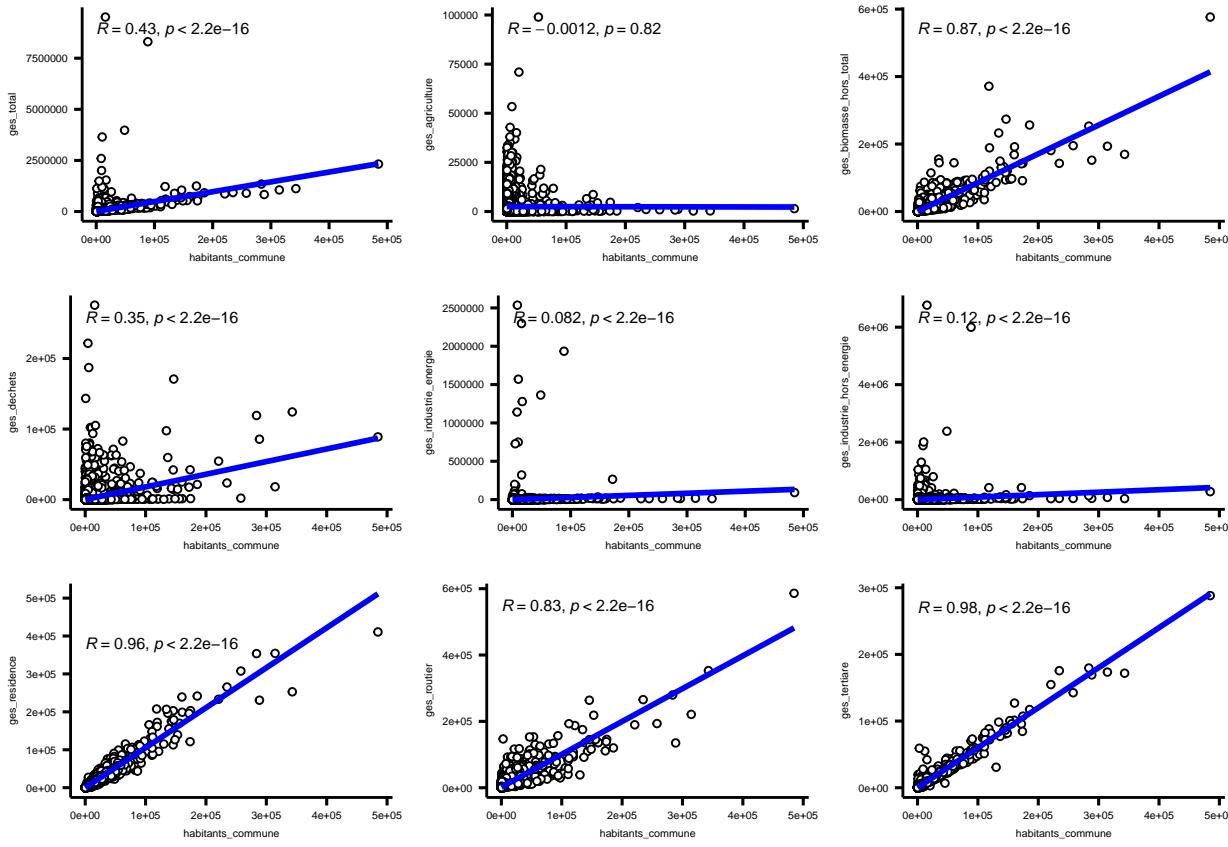
On aperçoit que le secteur routier est facteur principal dans l'émission de GES, l'industrie et l'agriculture suivent juste après. Ces données sont nécessaire pour nous rediriger par la suite : Nous analyserons postérieurement la corrélation entre l'émission de ces secteurs et le nombre d'habitants présents par commune.

Voici un résumé statistique du GES total par commune:

```
summary(enregistrement_commune["ges_total"])
```

```
##      ges_total
##  Min.   :    95
##  1st Qu.:  2625
##  Median :  5030
##  Mean   : 14301
##  3rd Qu.: 10700
##  Max.   :9514690
```

Analysons alors la corrélation entre le nombre d'habitants par commune et la quantité de GES émise par secteur. Nous analyserons aussi la corrélation pour l'émission de GES total de tous les secteurs de chaque commune et le nombre d'habitants de chaque commune.



À l'exception du secteur agriculture et industrie_energie, tous les graphes ci-dessus possèdent un coefficient de corrélation positif et assez élevé et un p minime, ce qui montre une corrélation positive entre le nombre d'habitants et le GES émis. À l'exception du secteur “dechets” et “agriculture”, tous les graphiques montrent un coefficient de Pearson supérieur à 0,8 et un p inférieur à $2,2 \cdot 10^{-16}$. La conclusion est la même qu'au niveau mondial, plus le nombre d'habitants par commune est élevé, plus le GES est lui aussi élevé.

CHAPITRE 5

Conclusion et perspectives

Que ce soit au niveau mondial ou à une échelle réduite, on a vu que la quantité de GES émise est correlée positivement avec le nombre d'habitants. En réalité, nous n'avons utilisé que le coefficient de Pearson pour calculer la corrélation mais il nous a permis d'avoir une conclusion fiable car c'est un outil fait pour 2 variables. Une autre approche serait d'étudier, avec le nombre d'habitants, la taille du pays ou encore le PIB. Par ailleurs, l'étude globale nous a permis de pointer du doigts les pays jouant le rôle principal dans l'émission de gaz, or d'après nos données on s'est rapidement rendu compte de l'impact dont la Chine et les Etats Unis infligent sur le monde, la pollution de leur habitations est largement supérieures aux autres pays, est-ce réellement leur habitation qui est responsable ? On n'oubliera pas de prendre en compte que la Chine est un pays industriel et que les Etats Unis sont la première puissance industrielle dans le monde. Pourra t on pointer du doigts la population et l'accuser comme responsable de cette pollution ?

Bibliographie

Annexes

Codes

code choropleth et jointure

```
# Jointure des dataframes pour avoir les noms en anglais des pays

# importation des libraires
library(dplyr)
library(maps) #Package maps provides lots of different map outlines and point
library(ggplot2) # ggplot2 provides the map_data() function
library(mapproj) # pour la projection

require(RMySQL)

con <- dbConnect(RMySQL::MySQL(), host = "localhost", dbname="pollution", user =
"root", password="")

# Creation variables
enregistrement_pays<-dbReadTable(conn =con, "enregistrement_pays")
enregistrement_commune<-dbReadTable(conn=con,"enregistrement_commune")
commune<-dbReadTable(conn=con,"commune")

# On rajoute les noms en anglais dans enregistrement_pays
data=read.csv2(file = "C:/PAUL VALERY 3/informatique/mini projet/tables_csv/nom_pays.csv")
#data

df=merge(x=data,y=enregistrement_pays,by="id_pays")

# On garde seulement les données pour 2014
df <- subset.data.frame(df, df$année == 2014)
#df

# On ne garde pas antarctique
world<-map_data("world")
world <- map_data("world") %>%
  filter(region != "Antarctica")

#On va joindre ce dataframe avec la carte du monde du package map
carte_ges=merge(x =world,y=df,by="region",all.x =TRUE)
carte_ges=arrange(carte_ges,group,order)
```

code diagramme en barre

```
tab <- dbReadTable(con, "enregistrement_pays")
attach(tab)
df_groupe_annee<-tab%>% group_by(id_pays) %>% summarise(moyenne_ges=mean(ges_t))
df_tri<-df_groupe_annee%>% arrange(desc(moyenne_ges))
df_tri<-head(df_tri,10)

X=df_tri$id_pays
Y=df_tri$moyenne_ges
library(ggplot2)
ggplot(df_tri,aes(x=X, y=Y, fill=as.factor(X)))+
    geom_bar(position=position_dodge(), stat="identity")+
    ylab("moyenne mpg")+xlab("pays")
```

code camembert

```
#Import des packages
library(ggplot2)
library(scales)

#Regroupement des moyennes dans chaque variables
routier<- mean(ges_routier)
tertiaire <- mean(ges_tertiaire)
residence <- mean(ges_residence)
ihe <- mean(ges_industrie_hors_energie)
ie <- mean(ges_industrie_energie)
dechets <- mean(ges_dechets)
biomasse <- mean(ges_biomasse_hors_total)
agriculture <- mean(ges_agriculture)

#Cr?ation d'un dataframe pour faire le graphique
df <- data.frame(
  secteurs = c("ges_routier","ges_tertiaire","ges_residence","ges_industrie_ho",
  value = c(routier, tertiaire, residence, ihe, ie, dechets, biomasse, agricultu
  )

#Barplot
bp<- ggplot(df, aes(x="", y=value, fill=secteurs)) + geom_bar(width = 1, stat = "identity")

#Camembert + ajout des pourcentages
#Calcul pour avoir les pourcentages
total <- sum(c(routier, tertiaire, residence, ihe, ie, dechets, biomasse, agricultu
proutier <- tertiaire/total*100
ptertiaire <- routier/total*100
```

```
presidence <- residence/total*100
pihe <- ihe/total*100
pourcent_ie <- ie/total*100
pdechets <- dechets/total*100
pbiomasse <- biomasse/total*100
pagriculture <- agriculture/total*100

val <- round(c(proutier, ptertiaire, presidence, pihe, pourcent_ie, pdechets, pbiomasse, pagriculture))

##Camembert
pie <- bp +
  coord_polar("y", start=0) + geom_text(aes(y = value/3 + c(0, cumsum(value)[-1]),
```