Sujet 1 - Concentration de CO2 dans l'atmosphère depuis 1958

Mélanie Debelgarric

05/04/2022

Thématique

En 1958, Charles David Keeling a initié une mesure de la concentration de CO2 dans l'atmosphère à l'observatoire de Mauna Loa, Hawaii, États-Unis qui continue jusqu'à aujourd'hui. L'objectif initial était d'**étudier la variation saisonnière**, mais l'intérêt s'est déplacé plus tard vers l'**étude de la tendance croissante dans le contexte du changement climatique**. En honneur à Keeling, ce jeu de données est souvent appelé "Keeling Curve".

Ouverture et Préparation des données

Les données sont récupérées sur le site Web de l'institut Scripps. Dans la consigne, il est demandé de télécharger les données hebdomadaires. Les données qui seront traitées ici décrivent donc la **concentration en CO2 atmosphérique pour chaque semaines de l'année 1958 à l'année actuelle (2022)**.

Téléchargement des données

Tout d'abord nous devons charger le fichier de données (à partir d'un lien url) :

```
data_url <- "https://scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/weekly/weekly_in_s
data <- read.csv(data_url, skip = 44, header=FALSE)
head(data)
```

```
##           V1     V2
## 1 1958-03-29 316.19
## 2 1958-04-05 317.31
## 3 1958-04-12 317.69
## 4 1958-04-19 317.58
## 5 1958-04-26 316.48
## 6 1958-05-03 316.95
```

Les deux colonnes du jeu de données correspondent respectivement à la date, et à la concentration en CO2. Je vais donc nommer ces deux colonnes pour faciliter la lecture et l'analyse:

```
colnames(data) <- c("date","CO2")
colnames(data) # Vérification : OK
```

```
## [1] "date" "CO2"
```

Nous allons ensuite vérifier s'il existe des données manquantes dans le jeu de données :

```
na_records = apply(data, 1, function (x) any(is.na(x)))
data[na_records,]
```

```
## [1] date CO2
## <0 lignes> (ou 'row.names' de longueur nulle)
```

Il ne semble pas qu'il y ait de valeur manquante. Maintenant, regardons la classe des deux variables :

```
class(data$date)
```

```
## [1] "character"
```

```
class(data$CO2)
```

```
## [1] "numeric"
```

La variable “CO2” est bien une variable numérique (logique pour une contration en CO2). Par contre, la variable “date” est sous la forme de “caractère”. Nous devons donc convertir cette variable pour que ce soit une date par la suite.

Préparation des données : gérer les données manquantes

Tout d’abord, nous devons convertir la colonne date en classe “Date” :

```
data$Date <- as.Date(data$date)
head(data)
```

```
##         date    CO2       Date
## 1 1958-03-29 316.19 1958-03-29
## 2 1958-04-05 317.31 1958-04-05
## 3 1958-04-12 317.69 1958-04-12
## 4 1958-04-19 317.58 1958-04-19
## 5 1958-04-26 316.48 1958-04-26
## 6 1958-05-03 316.95 1958-05-03
```

```
class(data$Date)
```

```
## [1] "Date"
```

Selon les commentaires de l’institut Scripps à propos de ce jeu de données, les valeurs hebdomadaires sont ajustées à midi au milieu de chaque semaines. Nous devons donc vérifier si il y a bien 7 jours de différences entre chaque semaines dans le jeu de données :

```
ligne_FALSE <- which((diff(data$Date) == 7)==FALSE) # Vecteur contenant les différentes lignes pour les
ligne_FALSE
```

```
##  [1]    6    8   15   17   30   34   44   54  211  226  232  242  270  278  280
## [16]  286  386  399  409  899 1303 1369 2300 2316 2386 2432 2480 2776 3154
```

Ce n’est pas le cas. D’après le vecteur “ligne_FALSE” il existe 29 lignes pour lesquelles il n’y a pas 7 jours de différences entre elles. Si l’on regarde la première valeur pour laquelle ce n’est pas le cas, c’est à dire à la 6ème ligne :

```
diff(c(data$Date[ligne_FALSE[1]],data$Date[ligne_FALSE[1]+1])) # différence entre ces deux dates
```

```
## Time difference of 14 days
```

Il existe une différence de 14 jours entre la date de la ligne 6 (03/05/1958) et de la date de la ligne 7 (17/05/1958). L’idéal serait d’avoir une valeur manquante “NA” entre ces deux dates, donc une ligne supplémentaire pour la date du 10/05/1958, pour avoir un jeu de données qui est hebdomadaire.

Tout d’abord, vérifions la différence de jours entre chacune des lignes pour laquelle cette différence n’est pas égale à 7 :

```
ecart <- rep(NA, length = length(ligne_FALSE)) # vecteur vide qui contiendra les ecarts de jours entre
for (i in 1:length(ligne_FALSE)){
  date_avant <- data$Date[ligne_FALSE[i]]
  date_apres <- data$Date[ligne_FALSE[i]+1]
  ecart[i] <- diff(c(date_avant, date_apres))
```

```
}
ecart
```

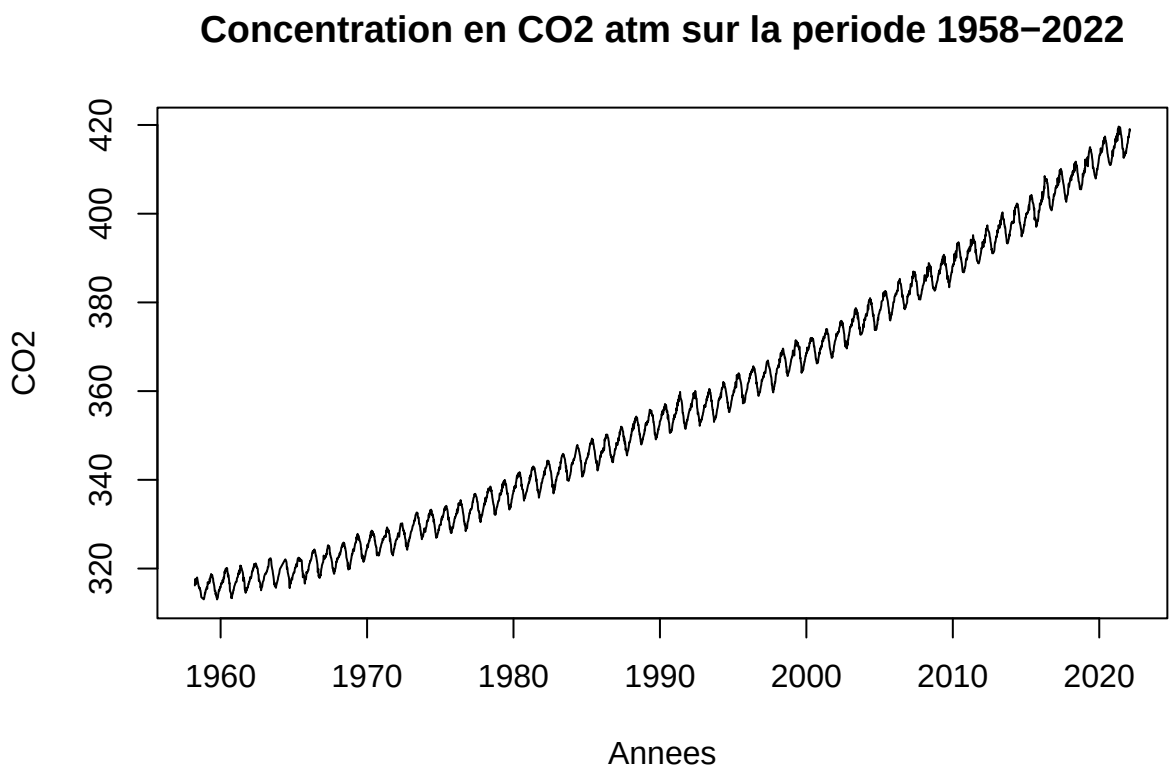
```
##  [1]  14  42  14  63  14  14  14  14  28  14  14  14  14 133  21  14  28  14  21
## [20]  14  35  14  14  21  35  21  14  21  14
```

Il existe une grande variabilité dans les écarts de jours entre deux dates. Notamment, nous avons un écart de 133 jours entre le 18/01/1964 et la date suivante.

Analyses graphiques

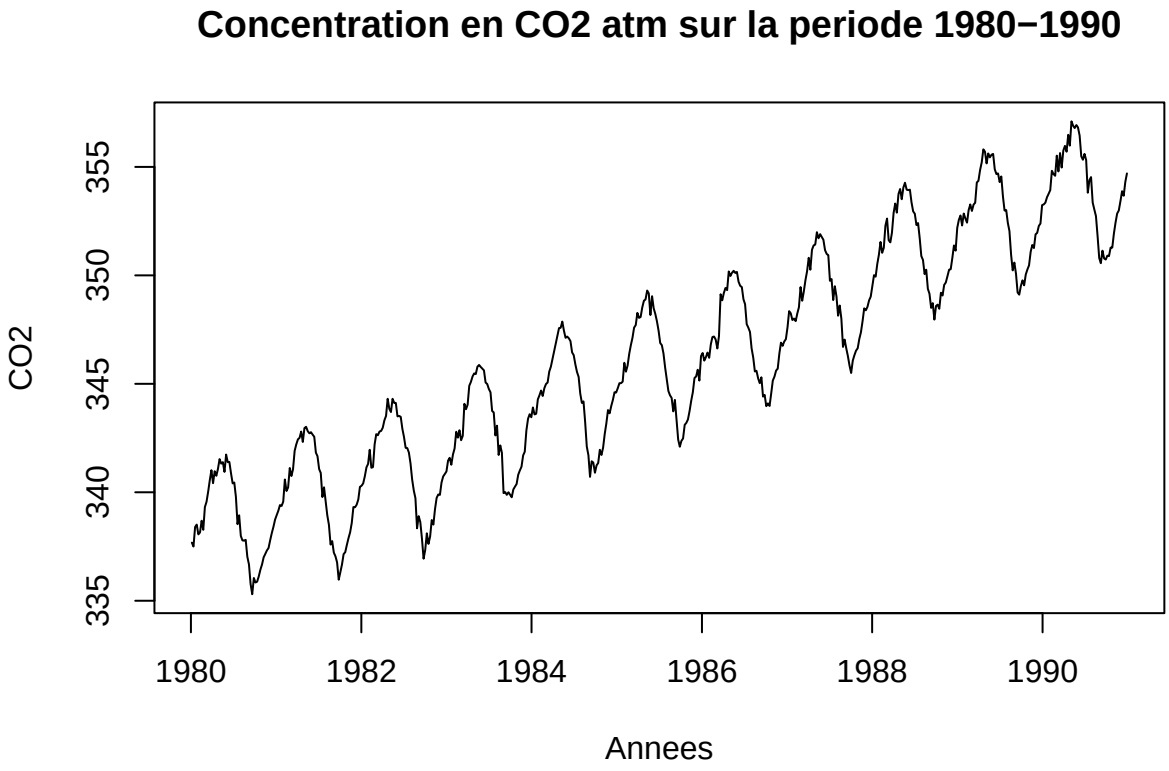
Nous pouvons en premier lieu faire un graphique montrant la concentration en CO2 au cours du temps.

```
plot(data$Date, data$CO2,
     type="l",
     xlab = "Annees", ylab="CO2",
     main = "Concentration en CO2 atm sur la periode 1958-2022")
```



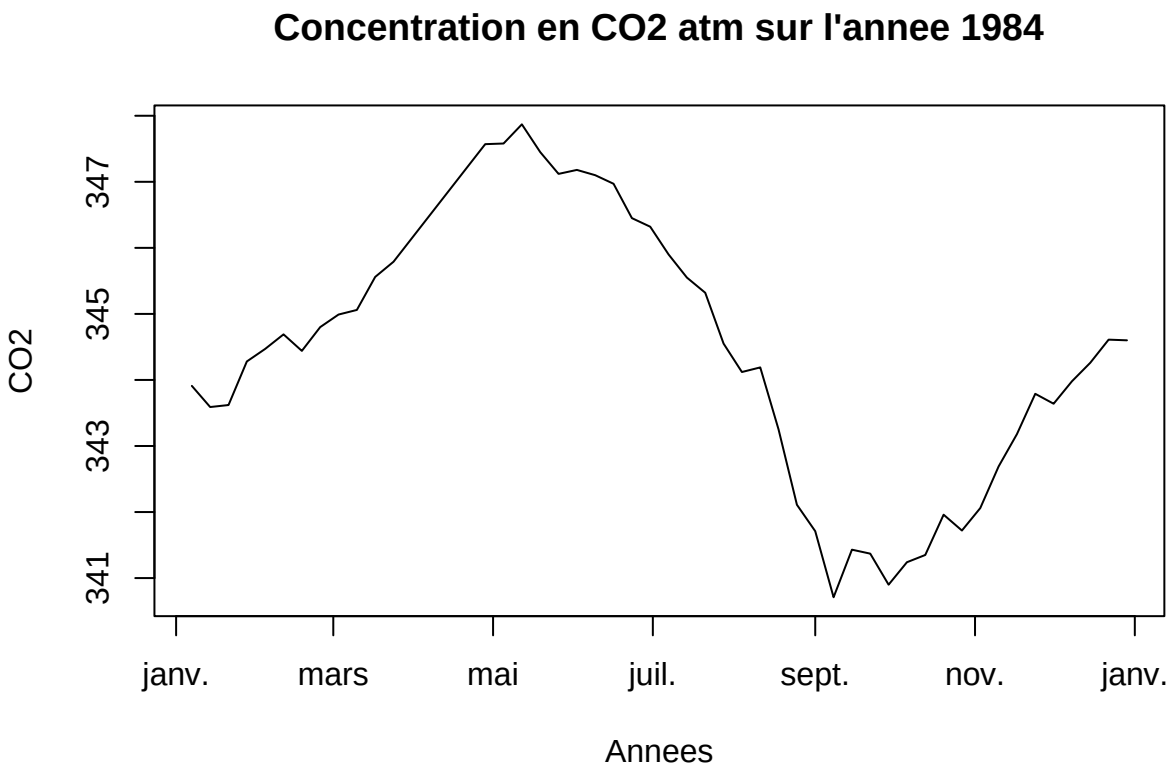
Sur ce plot, nous pouvons voir qu'il y a une augmentation de la concentration en CO2 atmosphérique depuis les années 1958. Sur ce plot, nous pouvons également voir qu'il existe des oscillations, correspondant aux variations annuelles de CO2 et dépendantes de la saison. Nous pouvons essayer d'identifier les "pics" de concentration en CO2 sur une période de 10 ans, comme le décrit le graphique suivant :

```
ann8090 <- subset(data, Date > "1979-12-31" & Date < "1991-01-01" )
plot(ann8090$Date, ann8090$CO2,
     type="l",
     xlab = "Annees", ylab="CO2",
     main = "Concentration en CO2 atm sur la periode 1980-1990")
```



La variation semble effectivement annuelle, vérifion cela sur une année choisie arbitrairement :

```
ann84 <- subset(data, Date > "1983-12-31" & Date < "1985-01-01" )
plot(ann84$Date, ann84$CO2,
     type="l",
     xlab = "Annees", ylab="CO2",
     main = "Concentration en CO2 atm sur l'annee 1984")
```

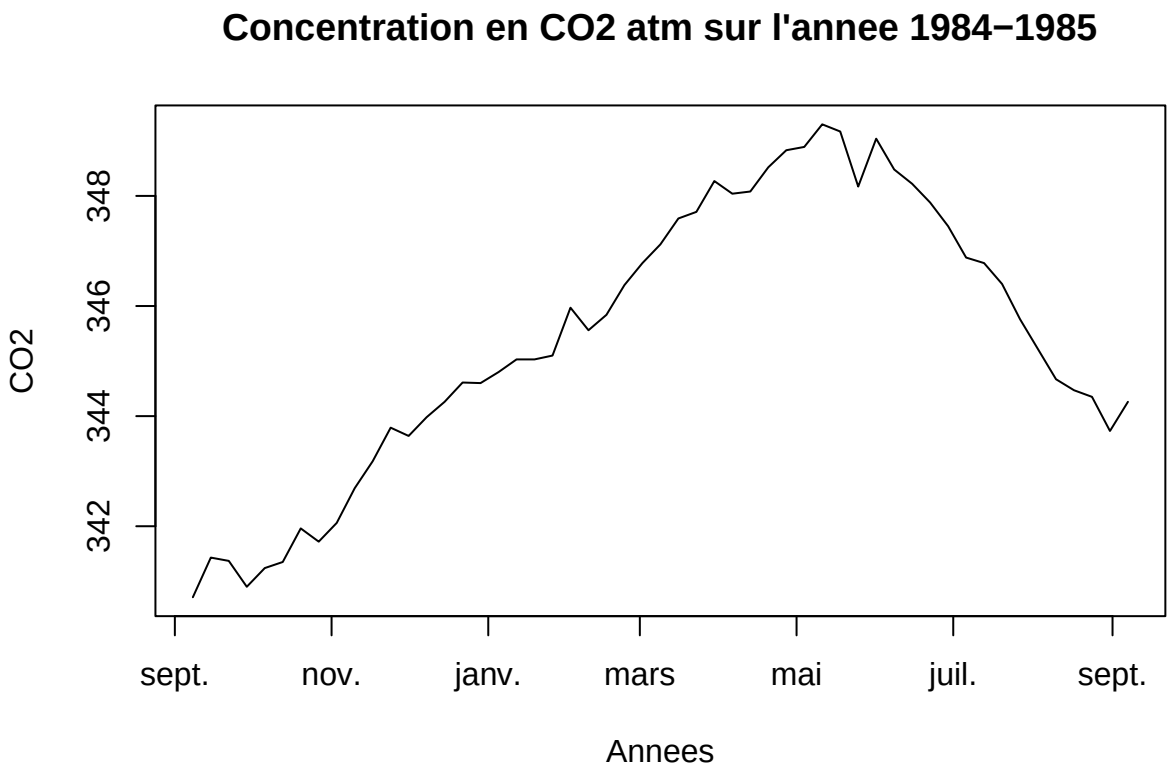


L'oscillation semble bien être annuelle, et semble débuter aux alentours de Septembre. Nous pouvons vérifier ceci graphiquement et le valider par le calcul :

```
mini <- ann84$Date[which.min(ann84$CO2)]
maxi <- ann84$Date[which.max(ann84$CO2)]
print(c(mini, maxi))
```

```
## [1] "1984-09-08" "1984-05-12"
```

```
ann8485 <- subset(data, Date > "1984-09-01" & Date < "1985-09-10" )
plot(ann8485$Date, ann8485$CO2,
     type="l",
     xlab = "Annees", ylab="CO2",
     main = "Concentration en CO2 atm sur l'annee 1984-1985")
```



En conclusion de cette analyse graphique, nous pouvons voir que :

1. La concentration en CO2 atmosphérique augmente au cours du temps depuis les années 1958 jusqu'à aujourd'hui
2. Cette concentration suit des oscillations périodiques annuelles au sein même de la période 1958-2022

Etude des oscillations annuelles sur la période donnée

Calcul de l'augmentation de CO2 au sein d'une oscillations

Si on reprend l'exemple de l'année 1984-1985, débutant et terminant un mois de Septembre, nous pouvons calculer la différence en concentration de CO2 existante.

```
diff8485 <- max(ann8485$CO2) - min(ann8485$CO2)
diff8485
```

```
## [1] 8.59
```

Il existe donc une différence de 8.59 entre le maximum (le "pic") et le minimum de l'oscillation annuelle 1984-1985. Maintenant nous pouvons créer une fonction sortant **la différence entre le maximum et le minimum de chaques oscillations** sur la totalité de la période (1958 - aujourd'hui).

Tout d'abord, j'utilise une fonction me permettant de connaître le minimum (pic_min) et le maximum (pic_max) de chaques oscillations annuelles. Ces oscillations débutent et se finissent en Septembre, comme on a pu le visualiser sur les graphiques précédents. La période étudiée sera alors de Septembre 1958 à Septembre 2021 (dernières données en date pour Septembre).

```
annee <- c(1958:2021) # Définition de la période étudiée en année

pic_min = function(annee) {
```

```
      debut = paste0(annee,"-09-01")
      fin = paste0(annee+1,"-09-01")
      semaines = data$Date > debut & data$Date <= fin
      min(data$CO2[semaines], na.rm=TRUE)
}

pic_max = function(annee) {
      debut = paste0(annee,"-09-01")
      fin = paste0(annee+1,"-09-01")
      semaines = data$Date > debut & data$Date <= fin
      max(data$CO2[semaines], na.rm=TRUE)
}
```

Je peux alors utiliser ces fonctions pour calculer la différence entre le maximum et le minimum pour chaque années, comme ceci :

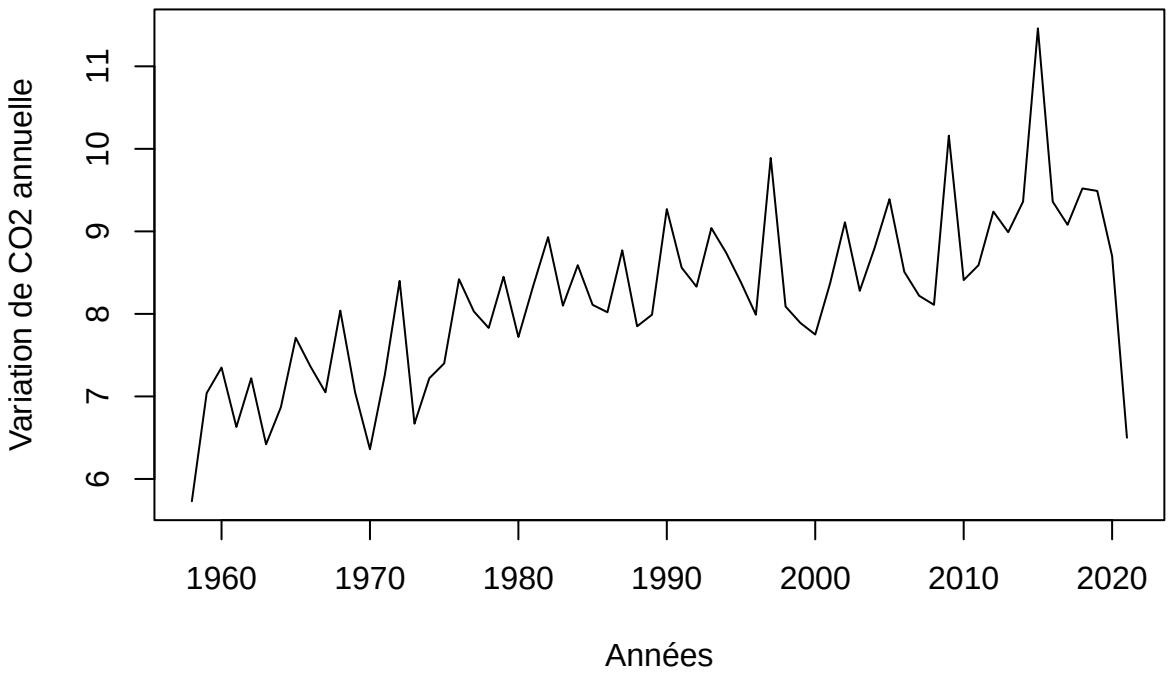
```
diff <- rep(NA, length = length(annee))
for (i in 1:length(annee)){
  min_an <- pic_min(annee[i])
  max_an <- pic_max(annee[i])

  diff[i] <- max_an - min_an
}
diff
```

```
##  [1]  5.73  7.04  7.35  6.63  7.22  6.42  6.87  7.71  7.36  7.05  8.04  7.05
## [13]  6.36  7.26  8.40  6.67  7.22  7.40  8.42  8.03  7.83  8.45  7.72  8.34
## [25]  8.93  8.10  8.59  8.11  8.02  8.77  7.85  7.99  9.27  8.56  8.33  9.04
## [37]  8.74  8.38  7.99  9.89  8.09  7.89  7.75  8.37  9.11  8.28  8.80  9.39
## [49]  8.51  8.22  8.11 10.16  8.41  8.59  9.24  8.99  9.36 11.46  9.36  9.08
## [61]  9.52  9.49  8.70  6.50
```

Maintenant que j'ai obtenu un vecteur avec la différence entre la concentration maximale et minimale de CO2 pour chaque année de la période étudiée, je peux regarder graphiquement s'il existe une variation dans l'oscillation périodique :

```
plot(annee, diff,
     type="l",
     xlab = "Années", ylab = "Variation de CO2 annuelle")
```



Nous pouvons remarquer sur ce graphique qu'il semble y avoir une augmentation de la différence en CO2 pour les oscillations sur la période étudiée (1958-2021). En effet, dans les années 1960, la variation de concentration de CO2 est de l'ordre de 6 à 8. Dans les années 2000, cette variations est de l'ordre de 7 à 10. Il semble qu'en plus d'une augmentation de concentration de CO2 atmosphérique depuis les années 1958, il existe également une augmentation de la variation de ce taux à l'échelle annuelle au cours du temps.

Modélisation et Prédictions futures

Modélisation de l'augmentation de CO2 jusqu'en 2025

Supposons qu'une simple modèle linéaire suffit pour expliquer l'augmentation du CO2 atmosphérique au cours du temps. Je construis ici un GLM pour faire une prédiction de la concentration en CO2 pour les prochaines semaines jusqu'à fin 2025. Après la construction de ce modèle, je peux le regarder graphiquement.

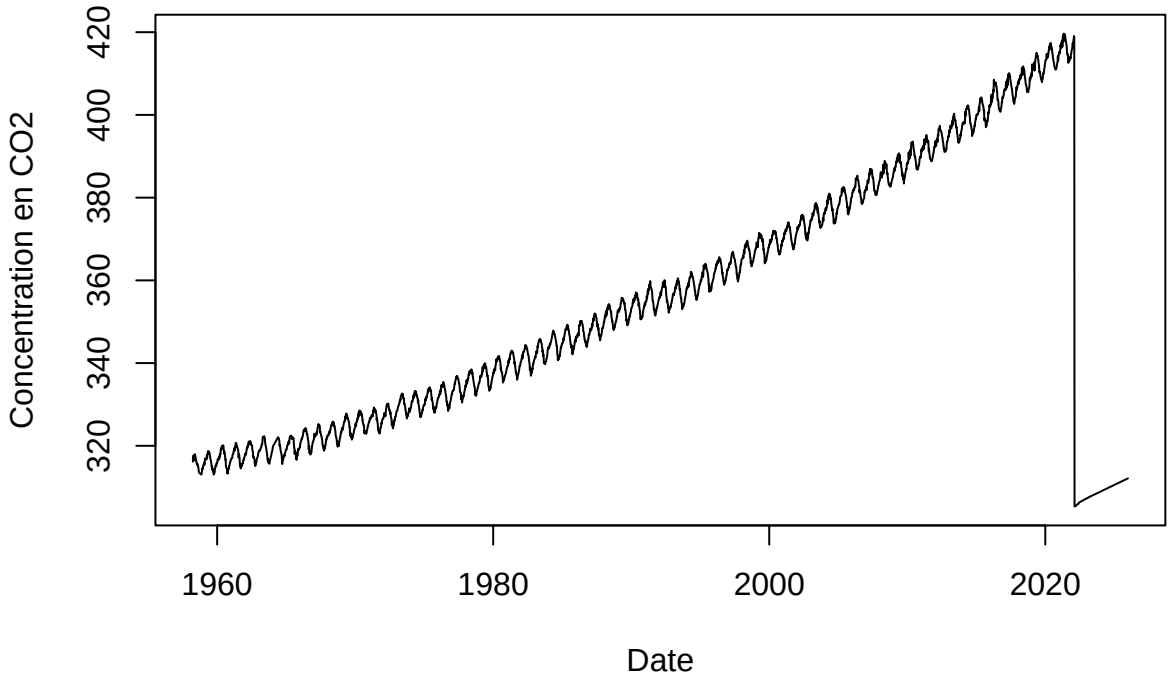
```
# Construction du modèle
model1 <- glm(CO2~Date, data = data)
new_date <- seq(from = as.Date("2022-02-12"), to = as.Date("2025-12-31"), by = "weeks")
pred <- predict.glm(model1, type="response")[1:length(new_date)]

# Data des données prédites par le GLM
data.glm <- data.frame(pred, new_date)
colnames(data.glm) <- c("CO2","Date")

# Fusion de mon dataset de base et du data issu du GLM
newdata <- data[,-1]
newdata <- rbind(newdata, data.glm)

# Graphique
```

```r
plot(newdata$CO2~newdata$Date,
     type="l",
     xlab = "Date", ylab = "Concentration en CO2")
```



Comme nous pouvons le voir sur le dernier plot, mon modèle ne prédit pas correctement l'augmentation de CO2 jusqu'en 2025. Ce modèle n'est donc pas approprié.

Commentaire pour le relecteur : Je ne suis pas sûre de répondre correctement à la consigne. Mais je pense que ce travail est suffisant pour m'évaluer sur la manière d'écrire un document computationnel.