

Concentration de CO₂ dans l'atmosphère depuis 1958

Hélène Raynal

21 avril 2020

Sujet

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" (voir (https://en.wikipedia.org/wiki/Keeling_Curve) (https://en.wikipedia.org/wiki/Keeling_Curve)) pour l'histoire et l'importance de ces données).

Les données sont disponibles sur le site Web de l'Institut Scripps (<https://scrippsco2.ucsd.edu/>). J'ai utilisé par erreur les observations mensuelles (je n'ai pas le temps de refaire tout le travail en utilisant les données hebdomadaires, ce sont d'ailleurs les données mensuelles qui sont utilisées en général dans les études d'impact climatiques) observations mensuelles (https://scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/monthly/monthly_in_situ_co2_mlo.csv). Attention, ce fichier est mis à jour régulièrement avec de nouvelles observations. Notez donc bien la date du téléchargement, et gardez une copie locale de la version précise que vous analysez. Faites aussi attention aux données manquantes.

Pré-requis: Quelques références sur le traitement de suites chronologiques:

- Introduction à l'Étude des Séries Temporelles (https://perso.math.univ-toulouse.fr/jydauxoi/files/2017/04/poly_eleves.pdf) J-Y. Dauxois 2016-2017
- ARIMA modelling in R (<https://otexts.com/fpp2/arima-r.html>)
- Séries temporelles – Modèles ARIMA. (<https://didierdelignieresblog.files.wordpress.com/2016/03/arimacomplet.pdf>) Didier Delignières Séminaire EA "Sport – Performance – Santé" Mars 2000
- Introduction à la manipulation de série temporelle avec R MAP-STA2 : Séries chronologiques (https://www.imo.universite-paris-saclay.fr/~goude/Materials/time_series/cours1_R_serie_temp.pdf) Yannig Goude yannig.goude@edf.fr (<mailto:yannig.goude@edf.fr>) 2019-2020
- Séries temporelles, avec R (<http://avram.perso.univ-pau.fr/proc/geo3.pdf>) Florin Avram

Récupération des données et pré-traitements (mise en forme et description)

Extrait des métadonnées d'intérêt et présentes dans le header (57 lignes) du fichier.

The data file below contains 10 columns.

- Columns 1-4 give the dates in several redundant formats.
- Column 5 below gives monthly Mauna Loa CO₂ concentrations in micro-mol CO₂ per mole (ppm), reported on the 2008A SIO manometric mole fraction scale. This is the standard version of the data most often sought. The monthly values have been adjusted to 24:00 hours on the 15th of each month.
- Column 6 gives the same data after a seasonal adjustment to remove the quasi-regular seasonal cycle. The adjustment involves subtracting from the data a 4-harmonic fit with a linear gain factor.
- Column 7 is a smoothed version of the data generated from a stiff cubic spline function plus 4-harmonic functions with linear gain.
- Column 8 is the same smoothed version with the seasonal cycle removed.
- Column 9 is identical to Column 5 except that the missing values from Column 5 have been filled with values from Column 7.
- Column 10 is identical to Column 6 except missing values have been filled with values from Column 8.
- Missing values are denoted by -99.99
- CO₂ concentrations are measured on the '08A' calibration scale

Chargement des librairies R, du fichier de données, renommage des colonnes, premières statistiques

```
setwd("C:/Users/hraynal/EspaceTravailBadet/FORMATION_PERSO/2020-Reproductibilite/mooc-rr/module3/exo3")

library(tidyverse)
library(forecast)
library(lubridate)
library(car)
library(scales)
library(patchwork)
library(kableExtra)
library(tinytex)

data_url <- c("https://scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/monthly/monthly_in_situ_co2_m1
o.csv")
dataC02 <- read.csv(data_url, sep="," ,skip = 57)

colnames(dataC02) <- c("Year", "Month","Date1", "Date2", "ObsC02", "SeasAdjC02","SplineAdjC02", "SplineAdjC02Trend", "O
bsC02Comp", "SeasAdjC02Comp")

summary(dataC02)
```

```
##      Year      Month      Date1      Date2
## Min.   :1958   Min.    : 1.000   Min.    :21231   Min.    :1958
## 1st Qu.:1973   1st Qu.: 4.000   1st Qu.:26968   1st Qu.:1974
## Median :1989   Median : 7.000   Median :32704   Median :1990
## Mean   :1989   Mean    : 6.507   Mean    :32705   Mean    :1990
## 3rd Qu.:2005   3rd Qu.: 9.500   3rd Qu.:38442   3rd Qu.:2005
## Max.   :2020   Max.    :12.000   Max.    :44180   Max.    :2021
##      ObsC02      SeasAdjC02      SplineAdjC02      SplineAdjC02Trend
## Min.   :-99.99   Min.    :-99.99   Min.    :-99.99   Min.    :-99.99
## 1st Qu.:328.50   1st Qu.:328.94   1st Qu.:328.46   1st Qu.:328.82
## Median :351.67   Median :352.18   Median :351.33   Median :352.03
## Mean   :346.87   Mean    :346.86   Mean    :348.95   Mean    :348.95
## 3rd Qu.:377.70   3rd Qu.:377.38   3rd Qu.:377.69   3rd Qu.:377.37
## Max.   :416.18   Max.    :413.35   Max.    :414.88   Max.    :413.29
##      ObsC02Comp      SeasAdjC02Comp
## Min.   :-99.99   Min.    :-99.99
## 1st Qu.:328.50   1st Qu.:328.94
## Median :351.67   Median :352.18
## Mean   :349.64   Mean    :349.63
## 3rd Qu.:377.70   3rd Qu.:377.38
## Max.   :416.18   Max.    :413.35
```

Traitement des valeurs manquantes notées -99.99

Il s'avère que dans la série des valeurs observées complétée par les valeurs interpolées, il reste des observations avec des valeurs manquantes (-99.99). On supprime ces observations.

```
dataC02 <- dataC02[dataC02$ObsC02Comp != "-99.99", ]
```

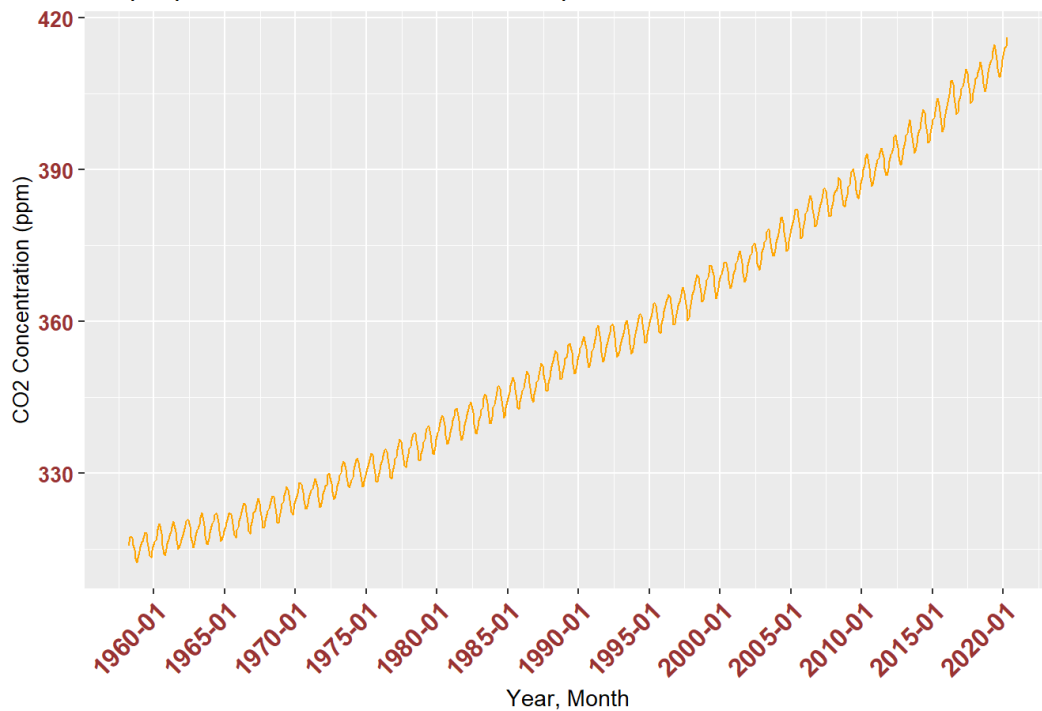
Création d'une variable date au format YYYY MM DD

```
dataC02$Date <- ymd(paste0(dataC02$Year, " ", dataC02$Month, " ", "15")) #Les données étant mensuelles, je fixe le 15 d
u mois comme jour de référence du mois
dataC02$Date <- ymd(paste0(dataC02$Year, "-", dataC02$Month, "-", "15"))
```

Représentations graphiques des résultats

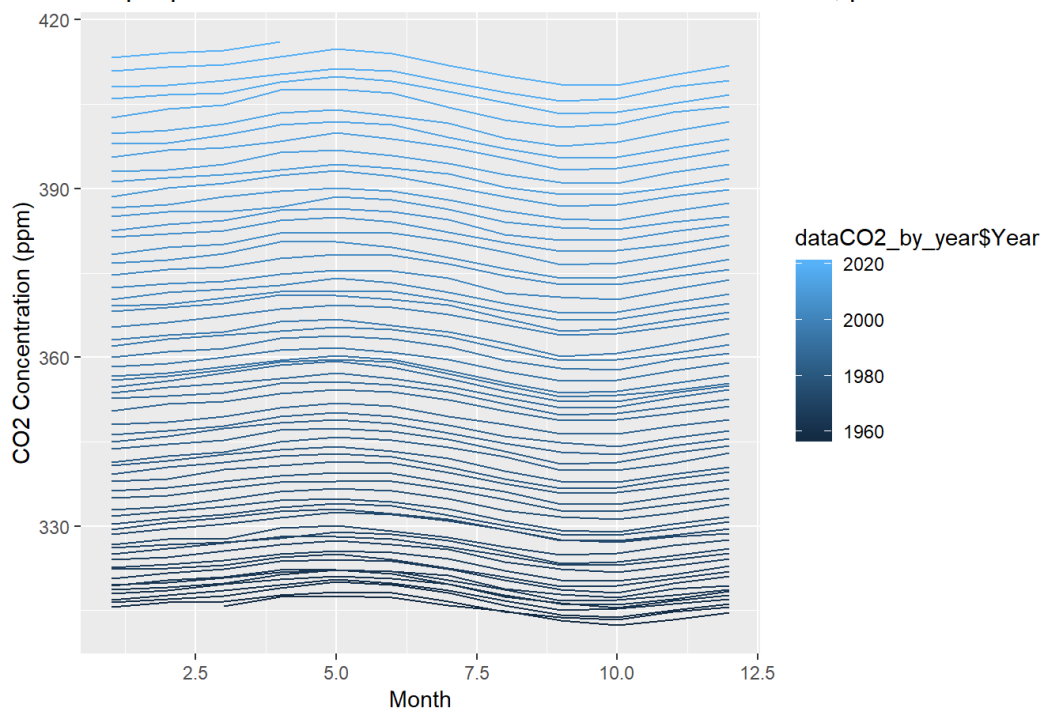
```
ggplot(dataC02,aes(Date, dataC02$ObsC02Comp)) +
  geom_line(color='orange') +
  xlab("Year, Month") +
  scale_x_date(date_labels = "%Y-%m", date_breaks = "5 year") +
  theme(axis.text.x = element_text(face = "bold", color = "#993333",
    size = 12, angle = 45, hjust = 1)) +
  ylab("CO2 Concentration (ppm)") +
  scale_y_continuous() +
  theme(axis.text.y = element_text(face = "bold", color = "#993333",
    size = 10, hjust = 1),axis.title.y = element_text(size = 10)) +
  ggtitle("Graphique 1: Evolution au cours du temps de la teneur en CO2")
```

Graphique 1: Evolution au cours du temps de la teneur en CO2



```
dataCO2_by_year <- dataCO2 %>% group_by("Year")
ggplot(dataCO2_by_year, aes(dataCO2_by_year$Month, dataCO2_by_year$obsCO2Comp)) +
  geom_line(aes(group = dataCO2_by_year$Year, colour = dataCO2_by_year$Year)) +
  xlab("Month") +
  ylab("CO2 Concentration (ppm)") +
  ggtitle("Graphique 2: Evolution saisonnière de la concentration en CO2, par année")
```

Graphique 2: Evolution saisonnière de la concentration en CO2, par année



- On déduit du graphique 1 que la série n'est pas stationnaire, elle présente une **tendance** (augmentation du taux de CO2 au cours du temps).
- La série montre une **composante saisonnière**, qui semble suivre le cycle annuel, comme en témoigne le graphique 2. En effet les courbes annuelles ont la même forme, tout en étant décalée dans le temps.
- On ne voit pas sur ces graphiques de ruptures dans le comportement de la série.
- On ne détecte pas d'observations aberrantes.

Question 1: Réalisez un graphique qui vous montrera une oscillation périodique superposée à une évolution systématique plus lente.

Le graphique 3 montre une oscillation périodique saisonnière, c'est la courbe orange. Cette oscillation périodique a été mise en évidence dans le graphique 2 et on l'a rapproché d'une composante saisonnière. La courbe bleue montre l'évolution systématique: la tendance. On voit que cette tendance n'est pas linéaire puisqu'elle ne peut pas être ajustée à une régression linéaire (courbe grise).

```
# Définition en tant que série temporelle
Co2_train <- ts(dataCO2$ObsCO2Comp, start = c(1958,3), frequency = 12)
T <- time(Co2_train)
base <- data.frame(Y=dataCO2$ObsCO2Comp,X1=as.numeric(T))

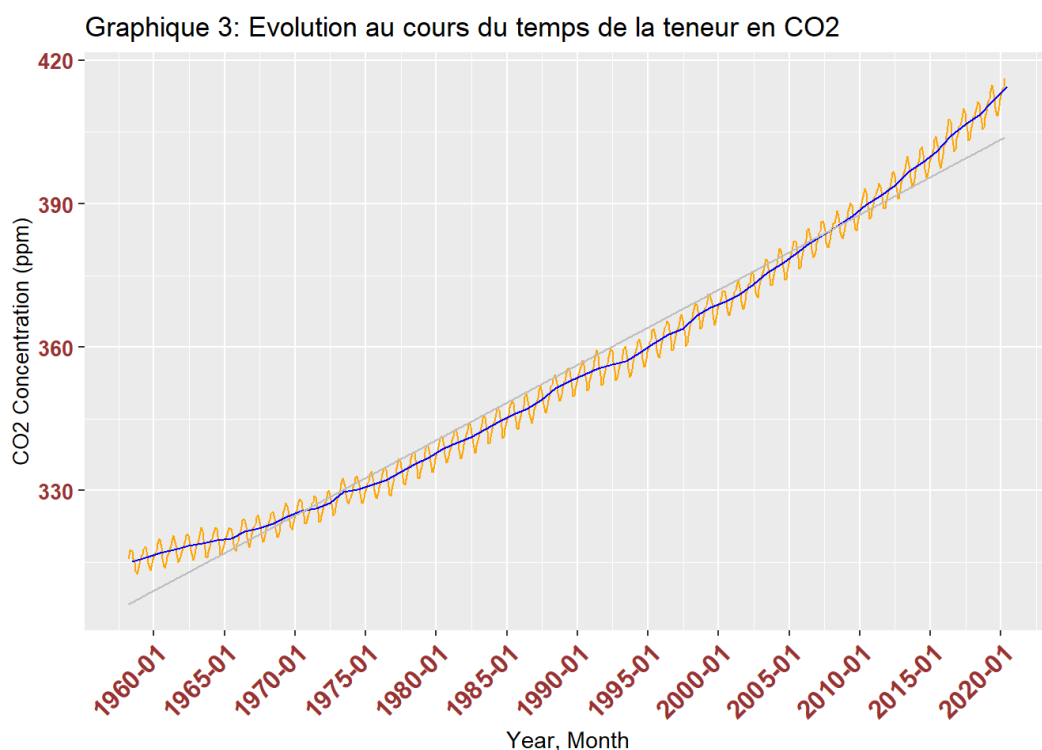
#essai d'ajustement à un modèle de régression linéaire
reg1 <- lm(Y~X1,data=base)
tendance1 <-predict(reg1)
T <-as.numeric(T)

#Calcul des moyennes annuelles de teneur en CO2
# <- cbind( Levels(factor(dataCO2$Year)), tapply(dataCO2$ObsCO2Comp, dataCO2$Year, mean))

MeanCO2 <- tapply(dataCO2$ObsCO2Comp, dataCO2$Year, mean)

tmp <- ymd(paste0(unique(dataCO2$Year), "-06-15"))
MeanCO2Year <- data.frame(tmp , as.numeric(MeanCO2) )
colnames(MeanCO2Year) <- c("Year", "MeanCO2")
dataCO2M <- merge( dataCO2, MeanCO2Year, by="Year")

ggplot() +
  geom_line(aes( x=dataCO2$Date, y=dataCO2$ObsCO2Comp), color='orange') +
  geom_line(aes(x=MeanCO2Year$Year, y=MeanCO2Year$MeanCO2 ), color="blue") +
  geom_line(aes(x=dataCO2$Date, y=tendance1 ), color="grey") +
  #geom_line(aes(x=T, y=tendance1 ), color="grey") +
  xlab("Year, Month") +
  scale_x_date(date_labels = "%Y-%m", date_breaks = "5 year") +
  theme(axis.text.x = element_text(face = "bold", color = "#993333",
    size = 12, angle = 45, hjust = 1)) +
  ylab("CO2 Concentration (ppm)") +
  scale_y_continuous() +
  theme(axis.text.y = element_text(face = "bold", color = "#993333",
    size = 10, hjust = 1),axis.title.y = element_text(size = 10)) +
  ggtitle("Graphique 3: Evolution au cours du temps de la teneur en CO2")
```



Question 2: Séparez ces deux phénomènes. Caractérisez l'oscillation périodique. Proposez un modèle simple de la contribution lente, estimez ses paramètres et tentez une extrapolation jusqu'à 2025 (dans le but de pouvoir valider le modèle par des observations futures).

Modélisation de la série temporelle

On va chercher à modéliser la série en utilisant une catégorie de modèles qui cherche à déterminer chaque valeur de la série en fonction des valeurs qui la précède ($y_t = f(y_{t-1}, y_{t-2}, \dots)$). C'est le cas des modèles ARIMA ("AutoRegressive – Integrated – Moving Average"). Cette catégorie de modèles a été popularisée et formalisée par Box et Jenkins (1976).

Les processus autorégressifs supposent que chaque point peut être prédit par la somme pondérée d'un ensemble de points précédents, plus un terme aléatoire d'erreur.

Le processus d'intégration suppose que chaque point présente une différence constante avec le point précédent.

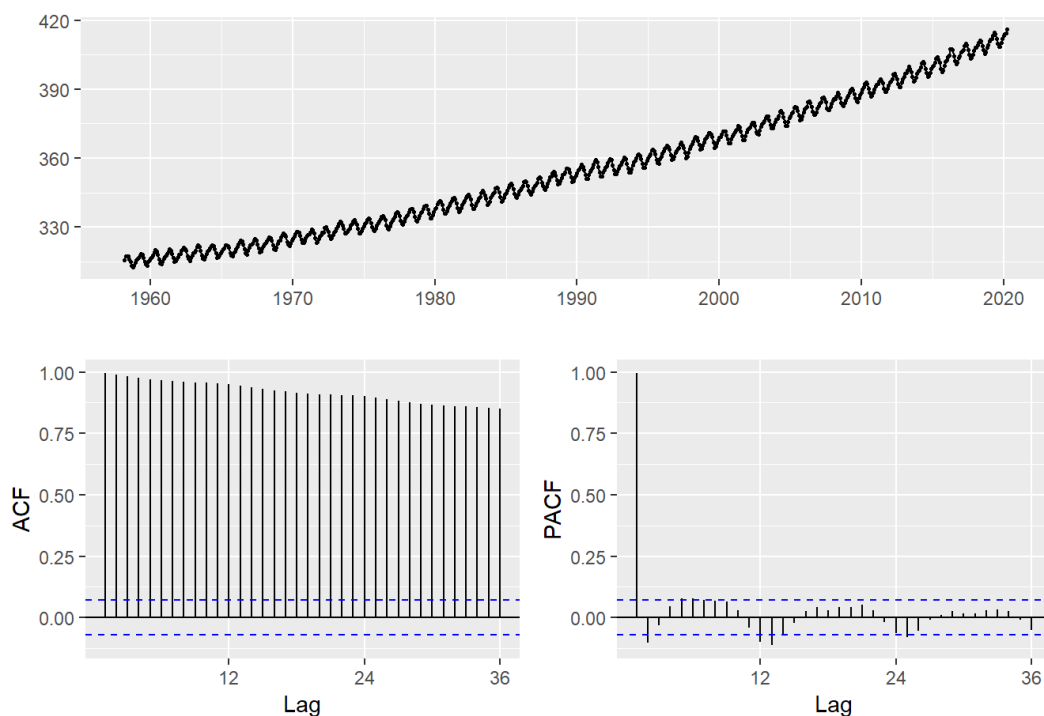
Les processus de moyenne mobile supposent que chaque point est fonction des erreurs entachant les points précédant, plus sa propre erreur.

Un modèle ARIMA est étiqueté comme modèle ARIMA (p,d,q), dans lequel:

- p est le nombre de termes auto-régressifs
- d est le nombre de différences
- q est le nombre de moyennes mobiles.

On va utiliser la fonction **auto.arima()** qui est un algorithme automatique qui permet de trouver le modèle ARIMA qui estime le mieux la série temporelle.

```
Co2_train <- ts(dataCO2$ObsCO2Comp, start = c(1958,3), frequency = 12)
Co2_train %>% ggtsdisplay()
```



```
set.seed(12356) # spécification de la graine pour assurer la replicabilité de nos résultats, ("fixer l'aléatoire")
auto.arima(Co2_train)
```

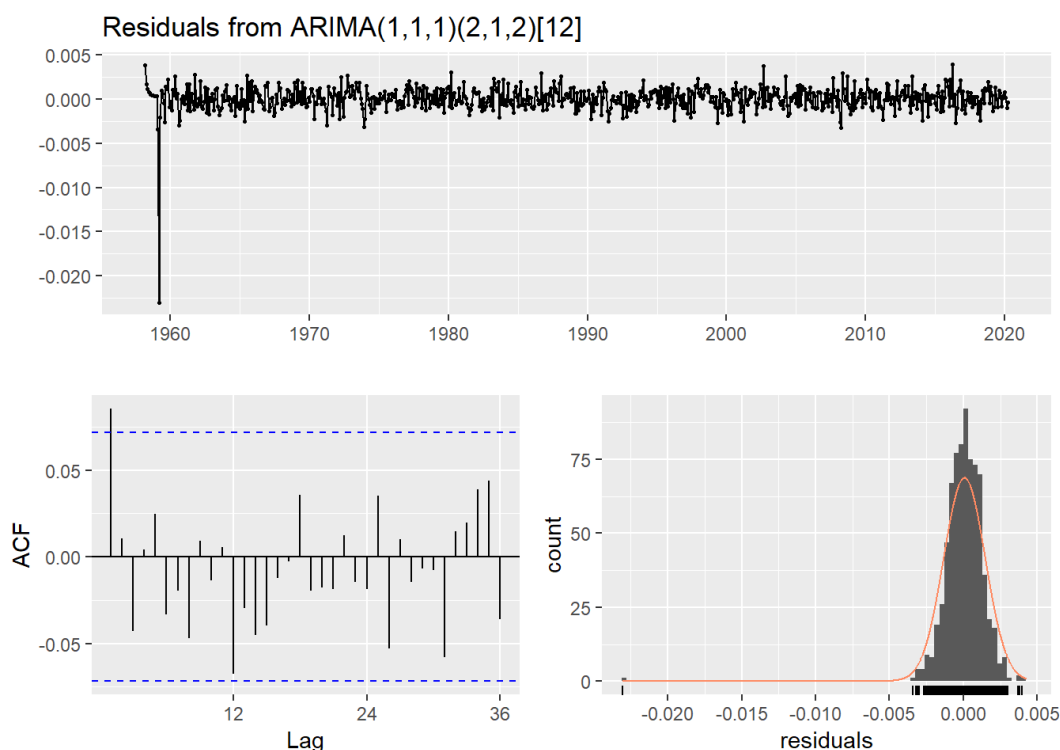
```
## Series: Co2_train
## ARIMA(1,1,1)(2,1,2)[12]
##
## Coefficients:
##      ar1      ma1      sar1      sar2      sma1      sma2
##    0.2193 -0.5802 -0.3041 -0.0090 -0.5446 -0.2647
## s.e. 0.0888  0.0744    NaN  0.0389    NaN    NaN
##
## sigma^2 estimated as 0.09685: log likelihood=-182.66
## AIC=379.31  AICc=379.47  BIC=411.49
```

En se basant sur le résultat de l'analyse précédente, on retient le modèle $ARIMA(1, 1, 1)(2, 1, 2)[12]$. On peut ensuite vérifier si les résidus ressemblent à un bruit blanc ou non.

```
(fit_minaicc <- Arima(Co2_train, order=c(1,1,1),seasonal=list(order=c(2,1,2),period=12),
  lambda = "auto"
))
```

```
## Series: Co2_train
## ARIMA(1,1,1)(2,1,2)[12]
## Box Cox transformation: lambda= 0.04611171
##
## Coefficients:
##      ar1      ma1      sar1      sar2      sma1      sma2
##      0.2169 -0.5781 -0.3383 -0.0105 -0.5332 -0.312
## s.e.  0.0910  0.0767      NaN      NaN      NaN      NaN
##
## sigma^2 estimated as 2.036e-06: log likelihood=3932.16
## AIC=-7850.33  AICc=-7850.17  BIC=-7818.15
```

```
checkresiduals(fit_minaicc, lag=36)
```



```
##
## Ljung-Box test
##
## data: Residuals from ARIMA(1,1,1)(2,1,2)[12]
## Q* = 30.041, df = 30, p-value = 0.4636
##
## Model df: 6. Total lags used: 36
```

```
fit_minaicc$aicc
```

```
## [1] -7850.174
```

Les graphiques précédents montre que le résidu ressemble suffisamment au bruit blanc, que la valeur p est élevée et que le modèle a passé le test de la Ljung-Box.

Extrapolation jusqu'à 2025

Les prévisions ci-dessous (aussi représentées sur le graphique 4) permet de visualiser les prévisions de teneur en CO2 jusqu'en 2025, en utilisant le modèle ARIMA estimé précédemment. Ainsi, la valeur prédite par le modèle est 428.3680 pour le mois d'avril 2025, avec un intervalle de confiance [423.5816, 433.2060].

```
forecast(fit_minaicc, h=60)
```

##	Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
##	May 2020	417.0723	416.4952	417.6501	416.1900	417.9563
##	Jun 2020	416.3214	415.6379	417.0060	415.2765	417.3689
##	Jul 2020	414.4725	413.7199	415.2265	413.3220	415.6262
##	Aug 2020	412.3290	411.5184	413.1411	411.0899	413.5716
##	Sep 2020	410.6638	409.8000	411.5292	409.3435	411.9881
##	Oct 2020	410.8612	409.9443	411.7799	409.4598	412.2671
##	Nov 2020	412.6284	411.6578	413.6012	411.1449	414.1170
##	Dec 2020	414.1127	413.0911	415.1368	412.5513	415.6798
##	Jan 2021	415.4457	414.3752	416.5187	413.8096	417.0878
##	Feb 2021	416.2138	415.0976	417.3329	414.5079	417.9264
##	Mar 2021	416.9610	415.8006	418.1245	415.1875	418.7417
##	Apr 2021	418.6402	417.4343	419.8493	416.7974	420.4907
##	May 2021	419.4669	418.1985	420.7390	417.5285	421.4139
##	Jun 2021	418.7110	417.3948	420.0312	416.6997	420.7317
##	Jul 2021	416.8524	415.4954	418.2136	414.7787	418.9359
##	Aug 2021	414.6915	413.2969	416.0906	412.5605	416.8330
##	Sep 2021	413.0196	411.5876	414.4562	410.8315	415.2187
##	Oct 2021	413.2307	411.7566	414.7098	410.9783	415.4948
##	Nov 2021	415.0151	413.4945	416.5411	412.6916	417.3510
##	Dec 2021	416.4950	414.9296	418.0660	414.1032	418.8999
##	Jan 2022	417.8384	416.2296	419.4533	415.3803	420.3105
##	Feb 2022	418.6140	416.9644	420.2698	416.0937	421.1488
##	Mar 2022	419.3644	417.6749	421.0604	416.7832	421.9608
##	Apr 2022	421.0540	419.3217	422.7932	418.4074	423.7166
##	May 2022	421.8933	420.1093	423.6846	419.1678	424.6358
##	Jun 2022	421.1279	419.3046	422.9587	418.3424	423.9310
##	Jul 2022	419.2550	417.3992	421.1187	416.4200	422.1085
##	Aug 2022	417.0878	415.2021	418.9816	414.2072	419.9875
##	Sep 2022	415.4071	413.4907	417.3320	412.4796	418.3544
##	Oct 2022	415.6117	413.6571	417.5751	412.6259	418.6181
##	Nov 2022	417.4045	415.4051	419.4131	414.3504	420.4801
##	Dec 2022	418.8977	416.8553	420.9496	415.7780	422.0398
##	Jan 2023	420.2495	418.1652	422.3438	417.0658	423.4564
##	Feb 2023	421.0278	418.9046	423.1612	417.7848	424.2947
##	Mar 2023	421.7791	419.6178	423.9511	418.4779	425.1052
##	Apr 2023	423.4777	421.2738	425.6925	420.1116	426.8694
##	May 2023	424.3196	422.0647	426.5859	420.8757	427.7903
##	Jun 2023	423.5518	421.2599	425.8556	420.0515	427.0799
##	Jul 2023	421.6700	419.3488	424.0035	418.1250	425.2437
##	Aug 2023	419.4890	417.1413	421.8494	415.9036	423.1040
##	Sep 2023	417.7989	415.4231	420.1878	414.1706	421.4576
##	Oct 2023	418.0071	415.5935	420.4341	414.3212	421.7243
##	Nov 2023	419.8102	417.3504	422.2838	416.0539	423.5988
##	Dec 2023	421.3100	418.8060	423.8282	417.4863	425.1671
##	Jan 2024	422.6689	420.1218	425.2306	418.7794	426.5927
##	Feb 2024	423.4519	420.8654	426.0535	419.5023	427.4369
##	Mar 2024	424.2085	421.5830	426.8495	420.1994	428.2540
##	Apr 2024	425.9164	423.2466	428.6023	421.8398	430.0306
##	May 2024	426.7636	424.0436	429.5002	422.6104	430.9557
##	Jun 2024	425.9910	423.2355	428.7636	421.7837	430.2382
##	Jul 2024	424.0984	421.3161	426.8983	419.8503	428.3876
##	Aug 2024	421.9061	419.0998	424.7302	417.6215	426.2325
##	Sep 2024	420.2067	417.3743	423.0575	415.8823	424.5740
##	Oct 2024	420.4153	417.5451	423.3044	416.0333	424.8414
##	Nov 2024	422.2281	419.3100	425.1657	417.7730	426.7286
##	Dec 2024	423.7367	420.7727	426.7206	419.2117	428.3082
##	Jan 2025	425.1031	422.0945	428.1322	420.5101	429.7440
##	Feb 2025	425.8903	422.8412	428.9603	421.2356	430.5940
##	Mar 2025	426.6507	423.5616	429.7613	421.9350	431.4166
##	Apr 2025	428.3680	425.2325	431.5255	423.5816	433.2060

```
autoplot(forecast(fit_minaicc, h=60), title="Graphique 4: Extrapolation jusqu'en 2025 ")
```

Forecasts from ARIMA(1,1,1)(2,1,2)[12]

