2 Greenhouse gas emissions

In recent years, our planet has been witnessing an unprecedented global crisis: climate change. Climate change finds one of its main causes in the greenhouse gas emissions, a phenomenon intrinsically linked to human activities and intense industrialisation. Emissions of carbon dioxide (CO_2), methane (CH_4), and other greenhouse gases are destabilising the delicate balance of our climate, leading to disastrous consequences in form of extreme weather events, rising sea levels, and changes of ecosystems. Given the ongoing changes, some global institutions and governments are implementing policies aimed at reduce, as much as possible, humanity's impact on the balance of the Earth's ecosystem.
In this context, the project aim is to examine the time series of total greenhouse gas emissions in France. Greenhouse gases are substances present in the Earth's atmosphere that have the ability to absorb and keep heat from the sun, contributing to the so-called greenhouse effect. This natural phenomenon is vital for maintaining temperatures on Earth that support life as we know it. However, the increase in greenhouse gas emissions, mainly due to human activities, raised concerns about excessive global warming and climate change. As mentioned earlier, our focus will be on the total greenhouse gas emissions: aggregating carbon dioxide, methane, and nitrous oxide from all sources, including agriculture. The standard used to valuate the aggregation of emissions from different greenhouse gases is carbon dioxide equivalent (CO_2e).
The CO_2e is a measure that expresses the impact on global warming of a certain quantity of greenhouse gas compared to the same quantity of carbon dioxide (CO_2). Specifically, one can refer to "grams of CO_2 equivalent," "kilograms of CO_2 equivalent," "tons of CO_2 equivalent," and so forth, respectively denoting a gram, a kilogram, or a metric ton of the substance.

2.1 EDA

The dataset used includes data from different parts of the world. However, we have chosen to focus our attention on an area limited to specific European country, in particular to consider France, one of major player in the economic world. The time series has annual frequency and composite of 70 observations, whose information refers to the years 1950 to 2020. The values reported are measured in scale of tonnes of CO_2 equivalent per capita.
The original time series can be found at https://ourworldindata.org/grapher/co-emissions-per-capita?tab=chart&country=~FRA.

The assumption that our time series is a realization of a stationary process is clearly fundamental in time series analysis, so we must first determine whether the series can be considered a realization of a stationary process. In this stage a very useful tool is the plot of the series, some patterns in the data can be seen visually.

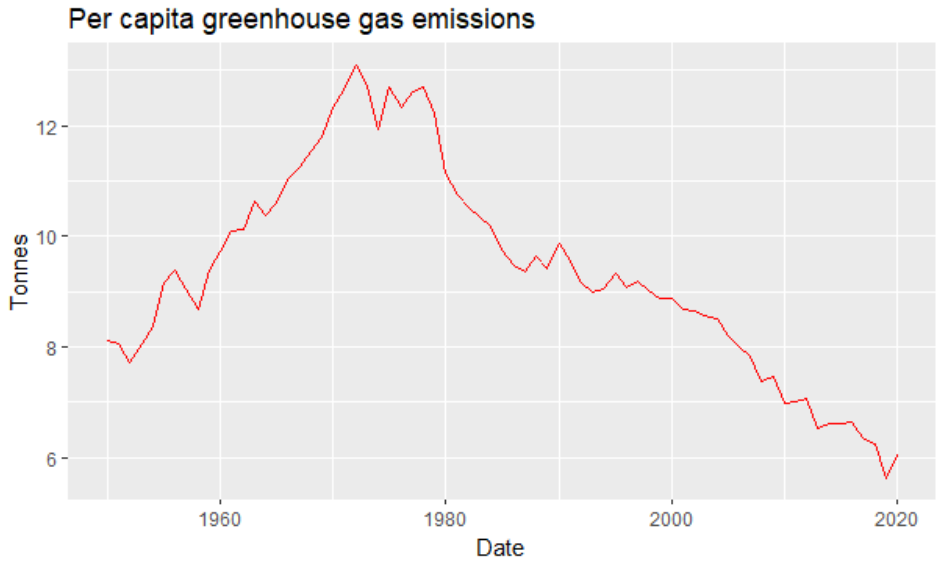
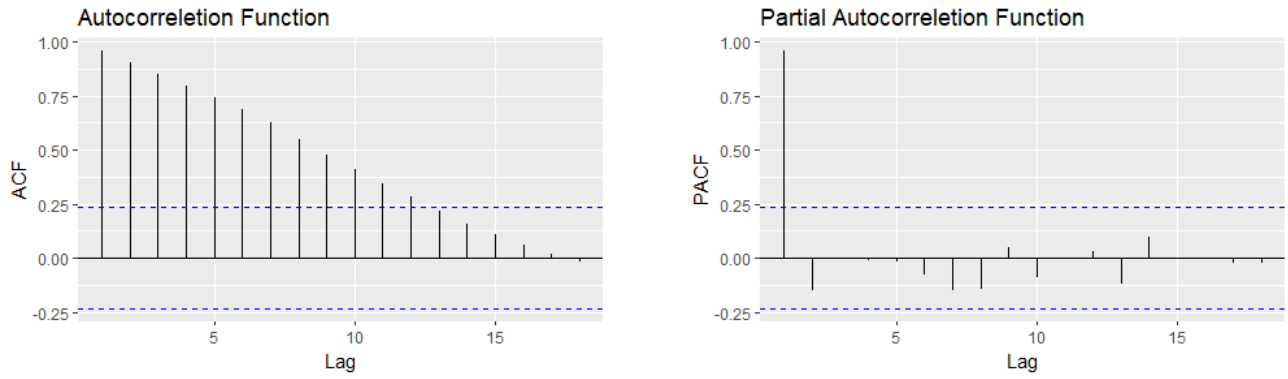


Figure 23: Time plot

In the time series plot (Figure26) we observe the values related to per capita emission of greenhouse gas measured in tonnes of CO_2e, in particular we observe the values recorded during the period 1950-2020.

We observe distinct trends in the series over time. In the first period of series, from year '50 to '73, there is a evident upward trend in greenhouse gas emissions per capita. However, from year '74 onward, we observe a downward trend in greenhouse gas emissions per capita. Regarding the seasonal component, given the annaual frequency of the series this cannot be observed. To confirm the presence of the trend component we are going to use suitable graphical tool, such as plot of ACF.

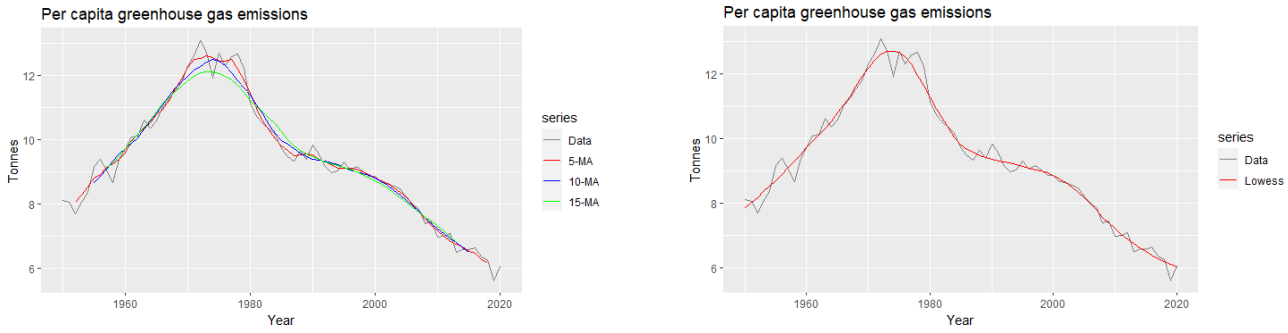


(a) (b)

Figure 24: ACF and PACF

From the ACF (refer to Figure 25) we observe that the autocorrelations exhibit a slow decay to 0, this phenomenon results a further confirmation of the presence of the trend component in the series. After confirming the presence of the trend component, the next step is to estimate it. Trend component estimation problem was approached through two different ways, in particular:

- Moving Avarage ($m - MA$)
- Lowess



(a) (b)

Figure 25: Trend Estimate

As expected, the Moving Average method, with a window size of 5, produces an unsmooth trend estimate. By increasing the window size, we notice an improvement in the smoothness of the trend estimate, which is particularly evident when using a window size of 15. However, although the trend is smoother with a window size of 15, this has a significant drawback: the loss of the first and last m/2 values.

The locally weighted scatter smoothing method, with a smoother span equal to 50% of the sample, obtain a considerable approximation of trend estimate. This approximation shows a high degree of smoothing and does not result in the loss of any value, unlike moving average methods. In general,the considerations regarding the nature of the trend in the first phase of analysis and exploration of the time series are confirmed in the trend estimate by both the Lowess and Moving Average methods.

2.2 ARIMA

The assumption that our time series is a realization of a stationary process is clearly fundamental in time series analysis. The Box-Jenkins methodology requires that the $ARIMA(p, d, q)$ process to be used in describing the DGP to be both stationary and invertible. Thus, in order to construct an ARIMA model, we must fall under the hypothesis that our time series can be considered a realization of a stationary process.

Detecting Stationary in variance
Looking the plot of the time series is difficult to determine the nature of the variance, about it being constant or not over time, in fact there are no fluctuations that increase with time.
To establish the nature of the variance, the Plot Std-Mean was cosidered, observing how the value of the variance change among different subgroups of the initial sample. As a function of this it was decreed that no large disparities are present between the variances of the sub-samples, however those present are not negligible.
In this case we have therefore considered a parametric transformation of the data that allows us to stabilize the variance of the series, the Box-Cox transform, defined as follows:

$$w_t = \begin{cases} log(x_t) & \text{if } \lambda = 0 \\ \frac{(x_t^{\lambda}-1)}{\lambda} & \text{if } \lambda \neq 0 \end{cases} \tag{2}$$

The problem is to find λ^* such that the series is as homoschedastic as possible. The choice of the optimal λ value can be made according to different criteria. Among them, we selected the one that makes the standard deviation as invariant as possible with respect to the mean, relative to several subgroups of observations.

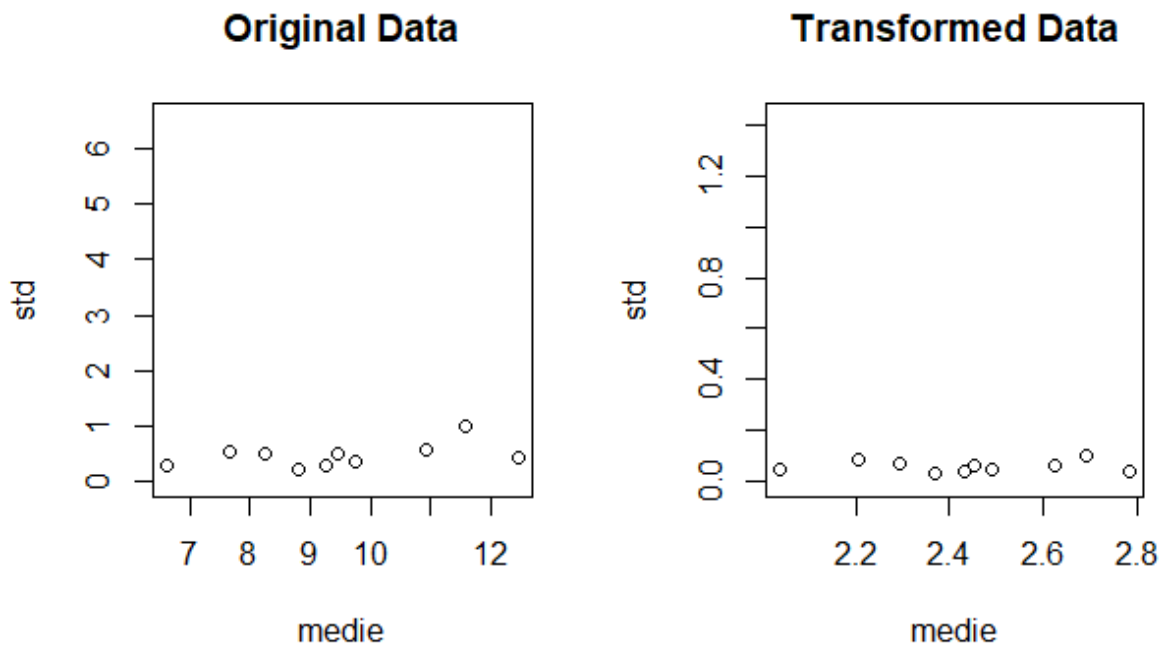


Figure 26: Std-Mean Plot

The λ that optimizes this relationship is the one identified by the method of Guerrero (1993). In this case we have $\lambda = 0.07803634$. Therefore, we can establish that the time series as a result of the transformation performed is stationary in variance.

Detecting Stationary in mean

Based on what was asserted in the previous sections, given the presence of the trend component, the time series considered w_t(following the Box-Cox transformation) is non-stationary on mean, differentiation can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore removing (or reducing) trend.

For this reason, we considered 2 types of differentiations of the series:

- *First-difference*: $\Delta w_t = (1 - L)w_t$
- *Second-difference*: $\Delta^2 w_t = (1 - L)^2 w_t$

The differentiations on the time series w_t are shown below:

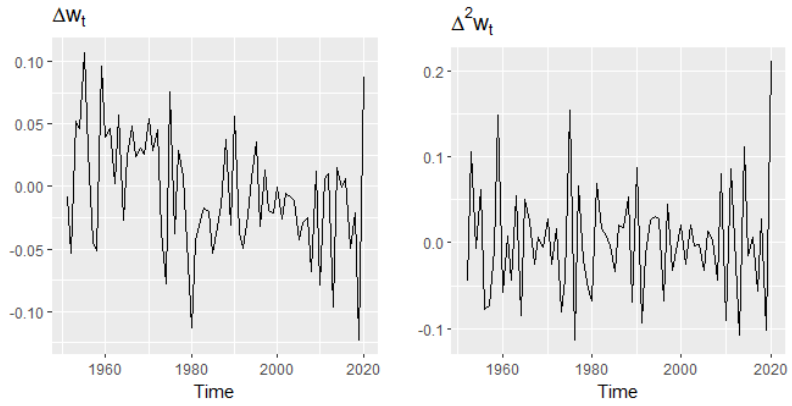


Figure 27: Differentiated Time series

From a first look at the plot of the produced series we note how Second-difference series seem to be stable around an average. For the selection of differentiation we exploited two criteria,one more objective, looking at the differentiation that minimizes variance, another more subjective one, which consists in selecting the differentiation that produces an ACF as close as possible to a stationary process.

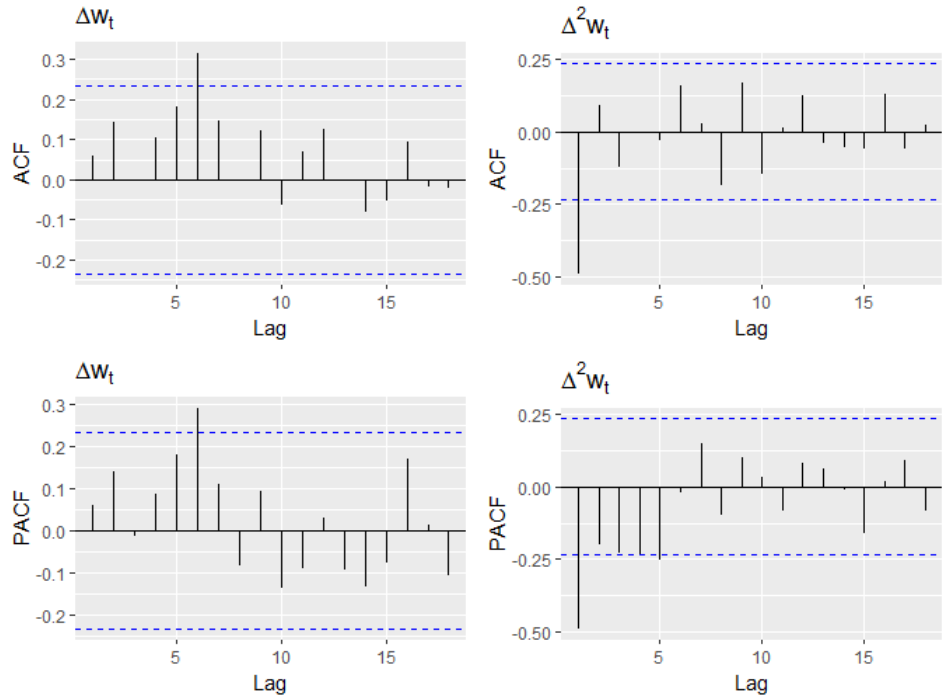


Figure 28: Acf and Pacf of differentiated Time series

Focus on the Figure 28, we can observe that the autocorrelations of the both series differences appear to be similar those of a stationary process.

Series	Variance	Fraction of Variance
w_t	0.05984404 5	-
Δw_t	0.002207931	0.03689476
$\Delta^2 w_t$	0.004090662	0.06835538

Variance table of series

Looking the objective method, based on the selection of the differentiation which minimizes the variance of the series, the choice falls on Δw_t.

To validate the selection of the chosen differentiation, the Augmented Dickey-Fuller (ADF) test is used, in our case, the null hypothesis is not rejected for significance level of test α equal to 5% (refers Figure 29), so Δw_t is not a stationary series.

```
        Augmented Dickey-Fuller Test

data:  dxt
Dickey-Fuller = -3.3181, Lag order = 4, p-value =
0.07607
alternative hypothesis: stationary
```

Figure 29: Augmented Dickey-Fuller test on Δw_t

So after discarding the first difference series we repeat the Augumented-Dickey-Fuller test(refers Figure 30) to verify hypothesis stationarity of the series $\Delta^2 w_t$, in this case we reject null Hypothesis for significance level of test α equal to 5%, so we can assume that $\Delta^2 w_t$ is a stationary series.

```
        Augmented Dickey-Fuller Test

data:  dxt2
Dickey-Fuller = -8.5734, Lag order = 4, p-value =
0.01
alternative hypothesis: stationary
```

Figure 30: Augmented Dickey-Fuller test on $\Delta^2 w_t$

After employing the graphical method by analyzing the ACF and PACF and taking into account the minimization of variance in relation to differentiation, the initial choice moved toward the application of the first difference(Δw_t). During the choice validation phase, the hypothesis of non-stationarity of the series was not rejected; therefore, the decision was revised in favor of the second difference(($\Delta^2 w_t$), which satisfies the conditions of stationarity for α equal to 5%. In the images below there are the time plot for it.

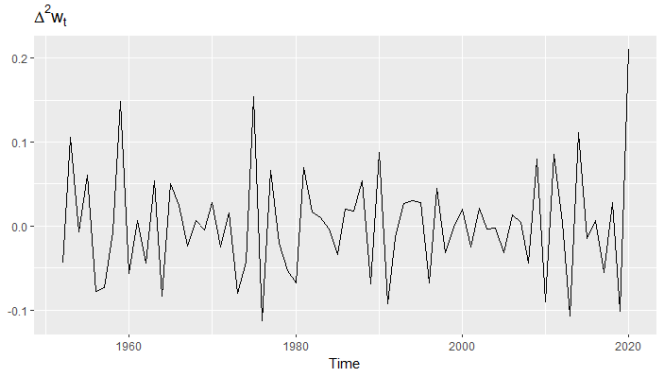


Figure 31: Plot of $\Delta^2 w_t$ series

Take into account the results shown, after a Box-Cox transformation and applied a Second Difference to transformed series, the time series $\Delta^2 w_t$ turns out to be stationary in mean and variance.

2.3 Model identification

Given the assumption that our time series $\Delta^2 w_t$ is the realization of a stationary process, the next step is model identification. Model Identification involves determining the order of the model required in way to capture the salient dynamic features of the data. Due to presence only of trend component, we research to identify an ARIMA, that could be write like $ARIMA(p, d, q)$, on the basis of what was stated in the previous section the parameter of d(difference) is set to 2. Therefore our problem of model identification is focus on the $ARIMA(p, 2, q)$ with zero mean(from now on, due to only models with zero mean were taken into account in the project, it will no longer be specified next to each model 'with zero mean') and it can be written as: $\phi_p(L)\Delta^2 w_t = \theta_q(L)\epsilon_t$, where:

- $\phi_p(L) = 1\text{-}\ \phi_1 L\text{-}...\text{-}\phi_p L^p$
- $\Delta^2 = (1\text{-}L)^2$
- $\theta_q(L) = 1\text{-}\ \theta_1 L\text{-}...\text{-}\theta_q L^q$

Therefore set parameter of difference, the problem of model identification focuses on parameter of AR and MA, to achieve aim of identification of parameter we use two different way:

- A procedure considered more subjective, based on the ACF and the PACF of the differential series($\Delta^2 wt$), trying to identify any exponential decays or cut-offs in the correleogram and partial correlogram.
- A procedure considered more objective, based on the estimation of different ARIMA models, whose parameters are different possible combinations of p and q.

Graphical method
As mentioned above this procedure is based on identification of parameter's model by looking the autocorreletions function plot and the partial autocorreletions function plot of time series. Acf and Pacf of the series Δw_t are shown below.

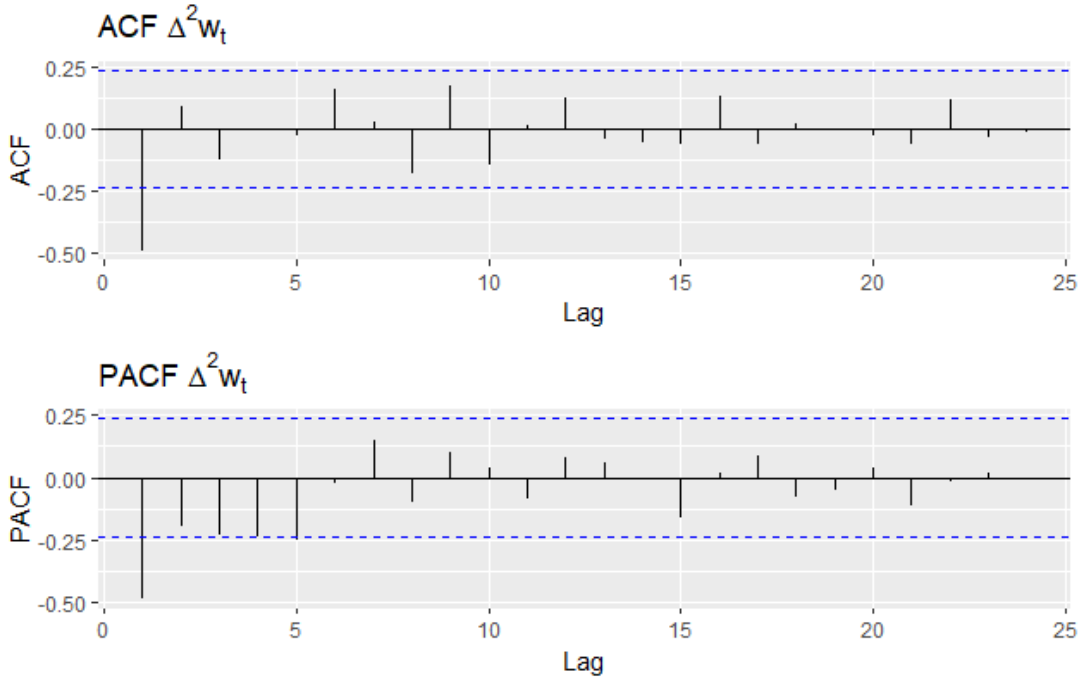


Figure 32: ACF and PACF of $\Delta^2 w_t$.

Focusing on Acf where we identify q, the orders of the moving average component, it is easy to see that there is a cut-off at the first autocorrelation, so we identify q=1. Regarding the autoregressive component,then looking at the PACF, this appears to exhibit a cut-off at the first partial autocorrelation, so we identify p=1.
Given the above considerations, by the graphical method, we identified $ARIMA(1,2,1)$.

Combination of parameter
By using this procedure we select parameters of ARIMA model taking different combination of them. In particular for the choice of the model it has been chosen an interval of values for p,q:

$$0 \leq p \leq 3$$
$$0 \leq q \leq 3$$

Thereby 16 models were estimated and selection of model has done by looking two information criteria:

- *The Schwarz bayesian information criterion* $(SBIC)$
- $The\ Hannan - Quinn\ information\ criterion\ (HQIC)$

Next figure represent different combination of model's parameter $ARIMA(p,2,q)$, in particular each row represent a particular combination of parameters, additionaly last two columns show the value of SBIC and HQIC for each model.

p	d	q	SBIC	HQIC
0	2	0	-170.6907	-168.8285
1	2	0	-185.3872	-181.6628
2	2	0	-186.6080	-181.0214
3	2	0	-191.4743	-184.0255
0	2	1	-200.2509	-196.5266
1	2	1	-199.3006	-193.7140
2	2	1	-198.3398	-190.8910
3	2	1	-200.8214	-191.5105
0	2	2	-199.3007	-193.7141
1	2	2	-198.4139	-190.9651
2	2	2	-197.7209	-188.4099
3	2	2	-204.7906	-193.6174
0	2	3	-198.3357	-190.8869
1	2	3	-197.9154	-188.6045
2	2	3	-199.1580	-187.9849
3	2	3	-204.0109	-190.9755

Figure 33: Table with model's parameter and IC.

Comparing the two information criteria for the 16 *models* examined, there turn out to be two optimal models each for a particular information criterion, by minimizing the information criterion $HQIC$ the optimal model turns out to be $ARIMA(0,2,1)$.
By minimizing the $SBIC$ information criterion, the optimal model among those estimated result $ARIMA(3,2,2)$.
Since discordance is present between the two criteria, it was decided to continue the estimation phase by considering both models that minimizing the information criteria.
So at the identification step, the models that have been identified are:

- $ARIMA(1,2,1)$ (identified by using ACF/PACF)
- $ARIMA(0,2,1)$ & $ARIMA(3,2,2)$ (identified by using IC)

2.4 Model Estimation & Checking

Once the models are established, the parameters and the corresponding standard errors are estimated using least square estimation method.
The next two figures represent information about the estimates of the three models.

```
ARIMA(3,2,2)

Coefficients:
         ar1      ar2      ar3      ma1     ma2
      0.4623  -0.2040  -0.3768  -1.3917  0.6914
s.e.  0.2095   0.1374   0.1535   0.1987  0.1696

sigma^2 = 0.001738:  log likelihood = 105.29
AIC=-198.58   AICc=-196.97   BIC=-186.12
```

(a)

```
ARIMA(1,2,1)

Coefficients:
         ar1      ma1
      0.0183  -0.8679
s.e.  0.1511   0.0799

sigma^2 = 0.001924:  log likelihood = 101.1
AIC=-196.2   AICc=-195.76   BIC=-189.96
```

(b)

```
ARIMA(0,2,1)

Coefficients:
          ma1
      -0.8632
s.e.   0.0720

sigma^2 = 0.001891:  log likelihood = 101.09
AIC=-198.18   AICc=-197.97   BIC=-194.03
```

(c)

Figure 34: Estimate models

In the figure 34 we observe the information regarding the estimation of models:

- $ARIMA(3,2,2)$ refer to Figure 34a
- $ARIMA(1,2,1)$ refer to Figure 34b
- $ARIMA(0,2,1)$ refer to Figure 34c

We observe the coefficients for each estimated parameter of the model components and their standard deviations. In addition, other information such as the estimated residual variance, value of the log-likelihood and information criteria are shown.
After estimation, the next step is to validate the estimated model, in particular, we focus on the model residuals: analyzing the residuals ACF & PACF plot and verifying that they have a white noise structure.

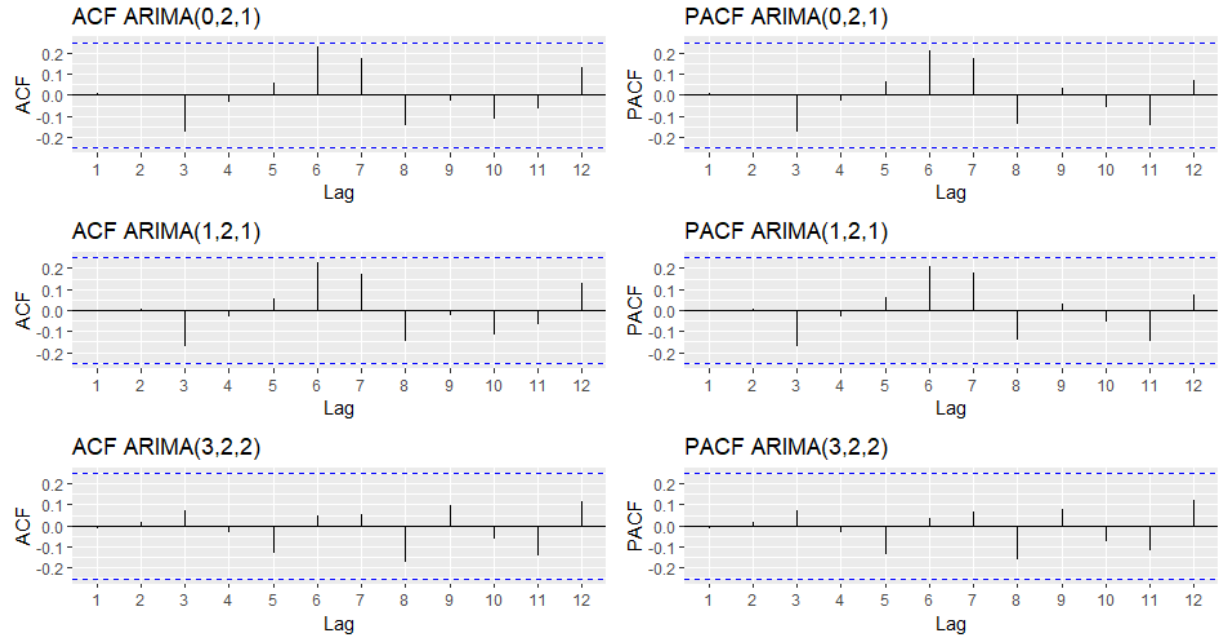
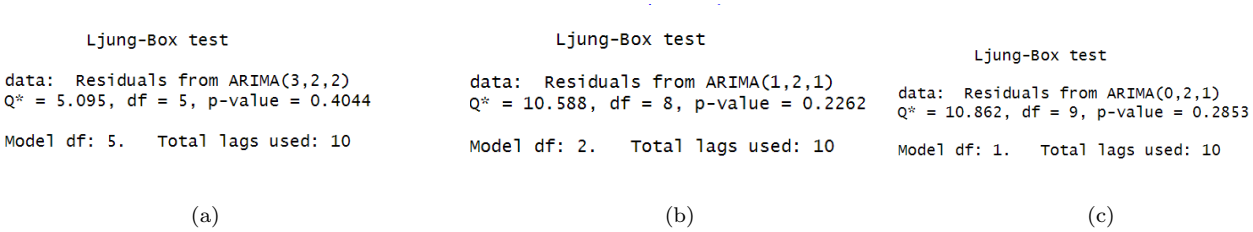


Figure 35: Acf and Pacf of model's residuals

Examining correlogram and partial correlogram (refer to Figure 35) of the models's residuals, all plot reveals that autocorrelations and partial autocorrelation are not significantly different from zero. This property in concordance to the notion that the residuals follow a white noise distribution.

To further validate the proposition of white noise structure in the residuals,for each model's residuals, the Ljung-Box test is conducted. The results state, for all models, that the first 10 lags are not jointly different from 0 for α at the 5% level(refer to Figure 36), providing additional evidence for the presence of white noise characteristics.

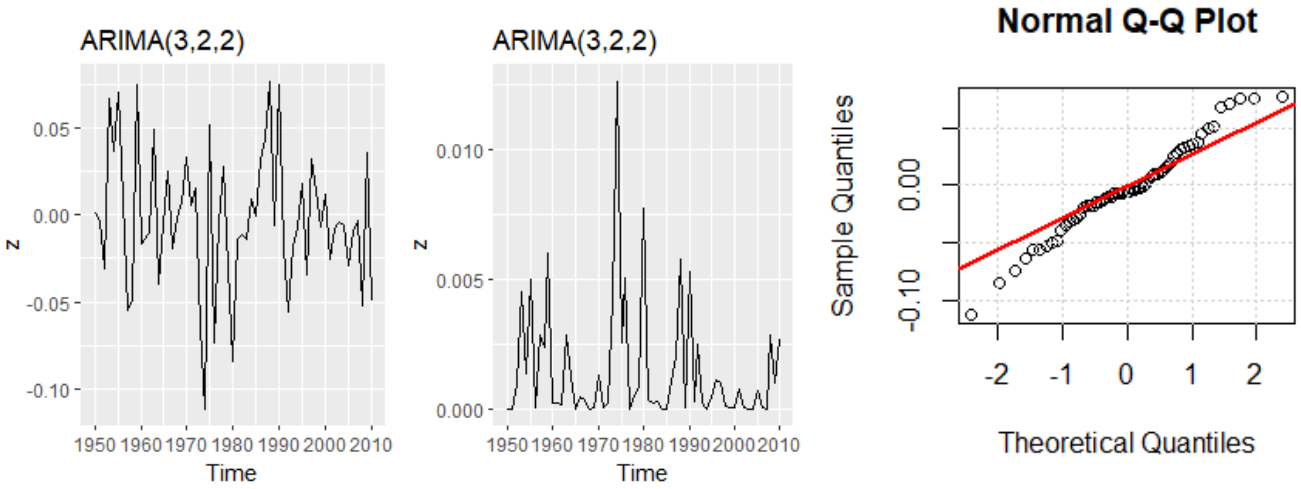


(a) (b) (c)

Figure 36: the Ljung-Box test

Thus, after analysing the autocorrelations of the residuals and conducting the Lijung-Box test, we can state that the residuals of the three models considered appear to be the realization of a white noise process.

Next, further graphical analysis regarding the residuals is shown, specifically the plot of standardized residuals, standardized squared residuals, and a plot comparing the cumulative distribution of the model residuals with the cumulative distribution of the normal (QQ-plot)

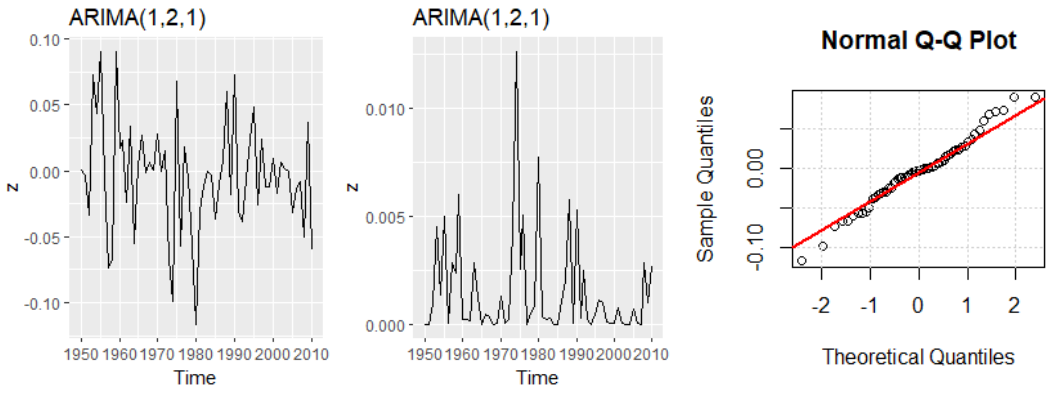


(a) (b)

Figure 37: Residual Analisys of model $ARIMA(3,2,2)$

The residuals of the model $ARIMA(3,2,2)$ appear to be approximately normal; indeed, the standardized residuals mostly fall within the range of [-2, 2] (refer to Figure 37a). By comparing the cumulative distribution of residuals to the normal distribution (refer to Figure 37b), clear similarities emerge, affirming the Gaussian nature of the residuals. To provide a comprehensive assessment, the Jarque-Bera test is conducted. The null hypothesis is not rejected at level α equal 5%, further confirming the normality of the residuals.

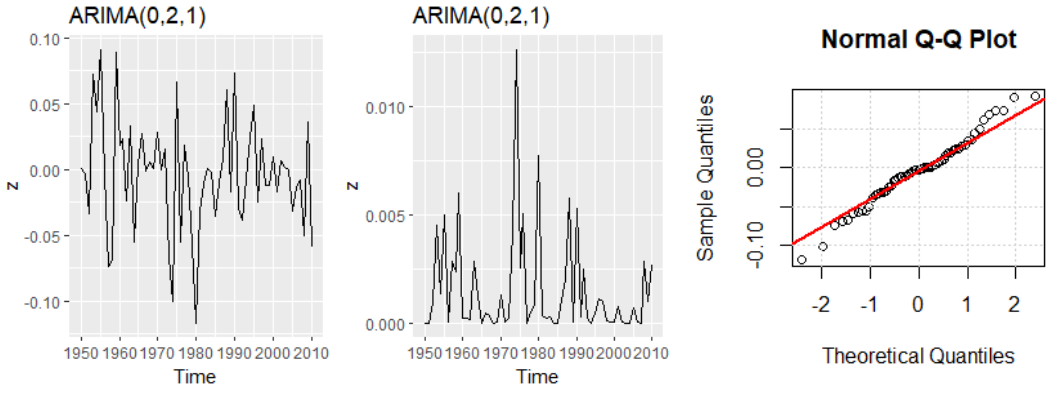
Additionally, the standardized residuals squared (refer to Figure 37a) are examined to detect any potential patterns. As illustrated in the figure, no distinct patterns or clusters are observed, indicating that the model is correctly specified.



(a) (b)

Figure 38: Residual Analisys of model $ARIMA(1, 2, 1)$

The residuals of the model $ARIMA(1, 2, 1)$ appear to be approximately normal; indeed, the standardized residuals mostly fall within the range of [-2, 2] (refer to Figure 38a). By comparing the cumulative distribution of residuals to the normal distribution (refer to Figure 38b), clear similarities emerge, affirming the Gaussian nature of the residuals. To provide a comprehensive assessment, the Jarque-Bera test is conducted. The null hypothesis is not rejected at level α equal 5%, further confirming the normality of the residuals.
Additionally, the standardized residuals squared (refer to Figure 38a) are examined to detect any potential patterns. As illustrated in the figure, no distinct patterns or clusters are observed, indicating that the model is correctly specified.



(a) (b)

Figure 39: Residual Analisys of model $ARIMA(0, 2, 1)$

The residuals of the model $ARIMA(0, 2, 1)$ appear to be approximately normal; indeed, the standardized residuals mostly fall within the range of [-2, 2] (refer to Figure 39a). By comparing the cumulative distribution of residuals to the normal distribution (refer to Figure 39b), clear similarities emerge, affirming the Gaussian nature of the residuals. To provide a comprehensive assessment, the Jarque-Bera test is conducted. The null hypothesis is not rejected at level α equal 5%, further confirming the normality of the residuals.
Additionally, the standardized residuals squared (refer to Figure 39a) are examined to detect any potential patterns. As illustrated in the figure, no distinct patterns or clusters are observed, indicating that the model is correctly specified.

Being under the assumption that residuals of model $ARIMA(3,2,2)$,$ARIMA(1,2,1)$ and $ARIMA(0,2,1)$ are normal, we can check the significance of the model parameters.

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ar1  0.46234    0.20949  2.2070   0.02732 *
ar2 -0.20397    0.13736 -1.4850   0.13755
ar3 -0.37676    0.15349 -2.4546   0.01410 *
ma1 -1.39168    0.19871 -7.0035 2.497e-12 ***
ma2  0.69141    0.16965  4.0757 4.589e-05 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a)

```
z test of coefficients:

     Estimate Std. Error  z value Pr(>|z|)
ar1  0.018311   0.151073   0.1212   0.9035
ma1 -0.867927   0.079884 -10.8649   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b)

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ma1 -0.863233   0.071993 -11.991 < 2.2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(c)

Figure 40: Test of significance of the coefficients

From the significance test, in the $ARIMA(3,2,2)$ model (refers to Figure 40a), the coefficients are all significant except for the second autoregressive component for α at level of 5%.
The significance test of the coefficients performed on the moving average component of the $ARIMA(0,2,1)$ model(refers to Figure 40c) confirms its significance at the α level of 5%.
By the significance test, in the $ARIMA(1,2,1)$ model (refers to Figure 40b), the coefficient of autoregressive component is not significant for α at level of 5%. Thus, given the non-significance of the autoregressive component in the model $ARIMA(1,2,1)$, it is removed with the result that the model is reduced to an $ARIMA(0,2,1)$.

Thus performed all the steps of the Box-Jenkins methodology we can say that the identified and estimated models are adequate for the considered time series, so in the next section through these two models forecast will be developed.

2.5 Forecasting

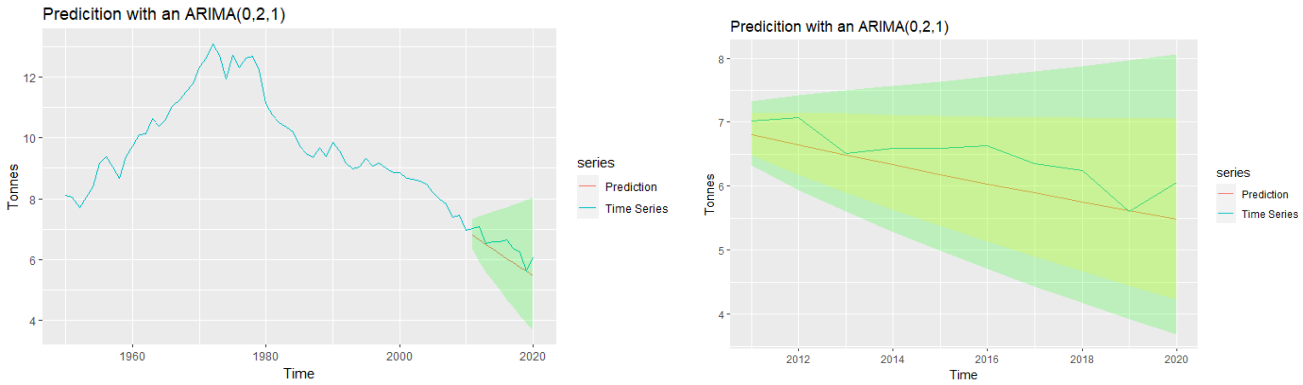
The dataset has been divided into train(from year 1950 to 2010 $\sim$ 85% of observation) and test set (from year 2011 to 2020 $\sim$ 15% of observation). On the train set the parameters of the model are estimated, on the test, after the forecast are obtained, the model being validate. Before evaluating the predictive accuracy of the model, it is necessary to back transform the forecasts to the original scale of the series and correct them to take into account the bias introduced in the forecasting process. The metrics used to evaluate models are

- RMSE
- MAPE
- MAE

Model	**RMSE**	**MAE**	**MAPE**
$ARIMA(3,2,2)$ **with zero-mean**	0.719	0.650	10.18
$ARIMA(0,2,1)$ **with zero-mean**	0.361	0.319	4.92

Table 2: Metrics

It can be seen from the table that the $ARIMA(0,2,1)$ model has higher prediction capabilities than the $ARIMA(3,2,2)$ model, in fact the model with a moving average component minimises all three metrics considered. To fully understand how well model does on the test set considered, it is sufficient to look at the MAPE metric where the $ARIMA(0,2,1)$ model on average produces forecast that deviate approximately 5% from the test observations, whereas the $ARIMA(3,2,2)$ model produces forecast that deviate from the test observations by approximately 10%. Therefore, in view of the above considerations, we can consider $ARIMA(0,2,1)$ a model with a better forecasting capabilities.



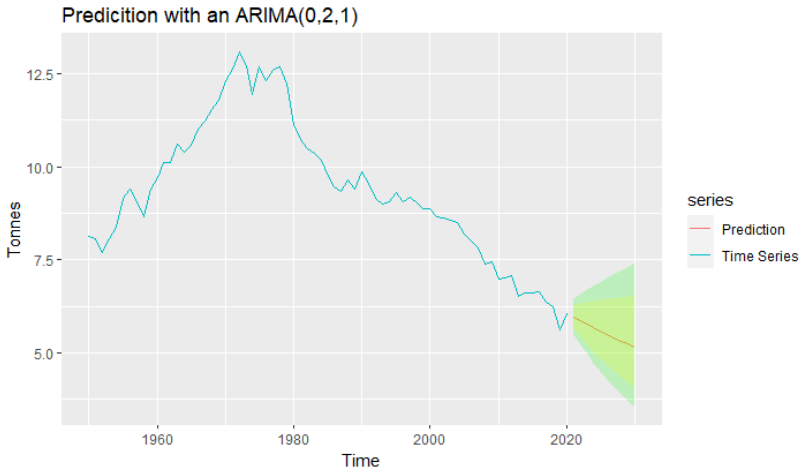
(a) (b)

Figure 41: Forecast with model $ARIMA(0, 2, 1)$

In the figure, it is evident that the model faithfully reproduces the downward trend observed in the series also in its forecasts. The plot also illustrates the forecast intervals, in particular, the yellow band denotes the 80% forecast interval, while the green band indicates the 95% forecast interval.

2.5.1 Potential future scenario

To conclude this work, we have chosen to make forecasts for the next 10 years since the last observation of the time series. Therefore, we propose to extend the forecasts until 2030 in order to outline a possible scenario in the context of per capita greenhouse gas emissions in France.



(a)

Figure 42: Forecast

The forecasts obtained from the $ARIMA(0, 2, 1)$ model for the period between 2020 and 2030 paint an encouraging picture in terms of per capita emission reductions, showing a significant decrease of 50% compared to the early 2000s. However, it is crucial to emphasise that these are only one possible evolution, and do not take into account some key factors that could influence Greenhouse gas emissions. These include demographics and policies economics at both national and international levels, including the European Union. Consideration of these elements is crucial to fully understanding the overall impact on emissions dynamics.