Universidad Politecnica de Catalunya

Time Series

Study Period 2

ARIMAX model for CO_2 emissions from the industrial sector in the USA

Author
Silje Marie Anfindsen
Carolin Drenda
Jonathan Stålberg

June 12, 2021

Contents

1 Introduction

In this report the Box-Jenkins ARIMA methodology with extensions for the treatment of calendar effects and outliers is applied to find a model for the time series of CO_2 emissions from the industrial sector in the USA (https://www.eia.gov/totalenergy/data/monthly/). A general idea of the methodology can be found in Figure 1. The data is published by the US Energy Information Administration and includes monthly data from 1990 until 2020 in Millions of Tm. The goal is to find a model to perform forecasting. The first step is to transform the model into stationary if this is not already the case. Then a model is identified for this stationary series and validated. In the second step an improved model is estimated possibly including treatment for calendar effects and oultiers. Both models are used for forecasting. In the end the models are compared with respect to different measures for accuracy of predictions and model fit.

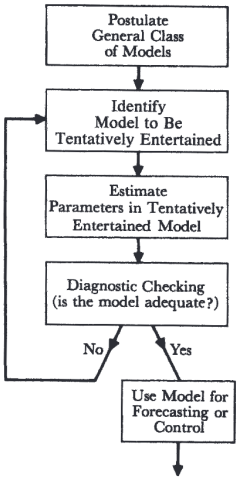


Figure 1: Scheme of Box-Jenkins Methodology (Box et al., 2008, p. 18)

2 Identification

We will start by identifying which transformations are necessary to make the time series stationary. The original time series is shown in Figure 2. The first step is to identify if a transformation is needed to make the variance stationary. Neither the mean-variance plot nor the boxplot (Figure 15a and 15b in the Appendix) show a sign of non-constant variance and therefore no transformation is needed for the variance.

To find out whether there is a seasonal pattern we take a look at the monthplot in Figure 3. We notice that there are some differences between the CO_2 emission for the different months, whith the highest values in January and the lowest in February. Thus we apply a seasonal difference to the series, that is $W_t = (1 - B^{12})X_t$, with X_t as the original time series. In Figure 4a this new time series is displayed. We apply a regular difference of the series, that is $W_t = (1-B)(1-B^{12})X_t$. This series is shown in Figure 4b. In order to know how many regular differences the series needs to be considered stationary the variance for the different transformed series is used. From Table 1 notice that the variance increases for the second applied regular difference so we choose the series $W_t = (1 - B)(1 - B^{12})X_t$ with one seasonal difference and one regular difference to work on further.

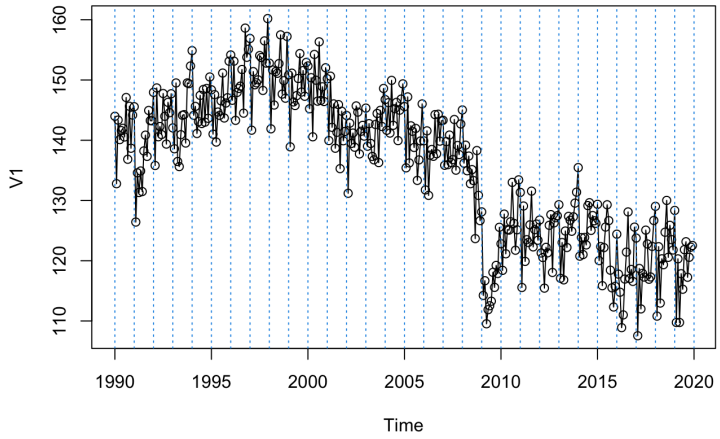


Figure 2: Plot of the original time series for CO2 emissions (in Millions of Tm) in USA before being transformed

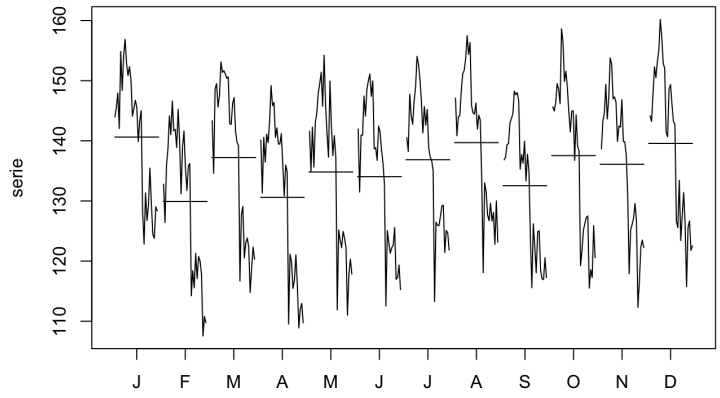
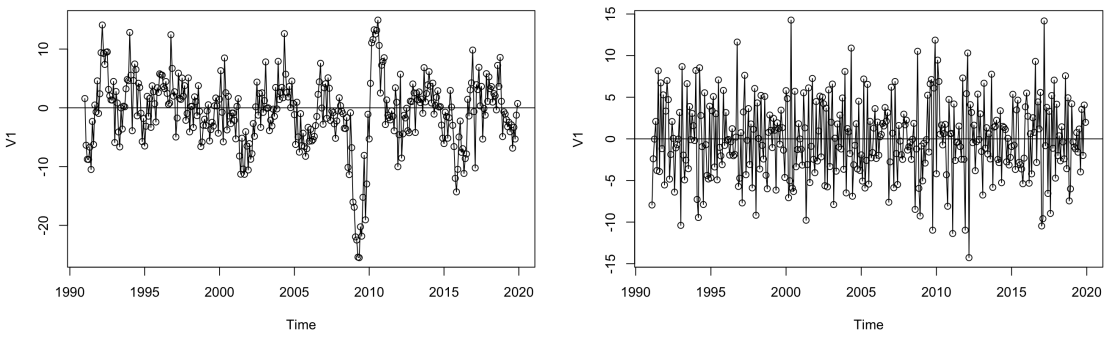


Figure 3: Monthplot of series

	serie	d12serie	d1d12serie	d1d1d12serie
Variance	150.0063	37.52583	22.08084	58.26977

Table 1: Variance of different transformed series

The next step in order to identify a suitable model is to plot the ACF and PACF of the transformed series (see Figure 5). For the seasonal part there is a decreasing pattern for the PACF. The ACF has one red spike outside the confidence band besides lag 0, suggesting a $MA(1)$ model for the seasonal part. For the regular part we observe a decreasing pattern for the ACF plot before the first seasonal lag. Two spikes lie outside the confidence band in the PACF plot which suggests an $AR(2)$ model. We also try the usual $ARMA(1,1)$.



(a) Times series after one seasonal difference plot(b) Time series after one regular difference and seasonal difference

Figure 4

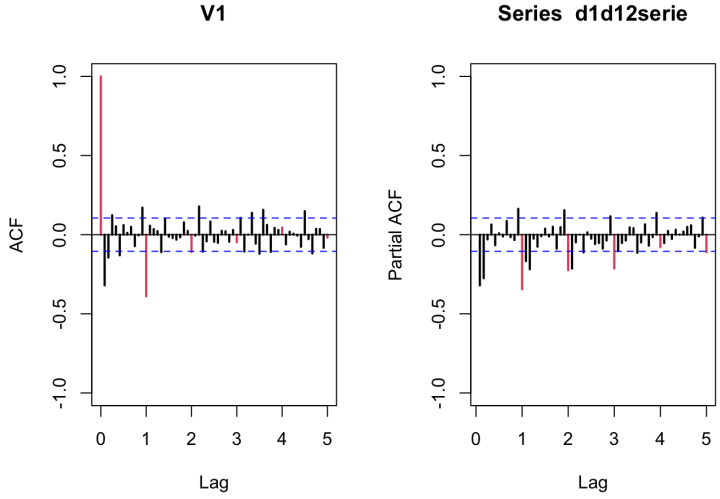


Figure 5: ACF and PACF for transformed series.

3 Estimation

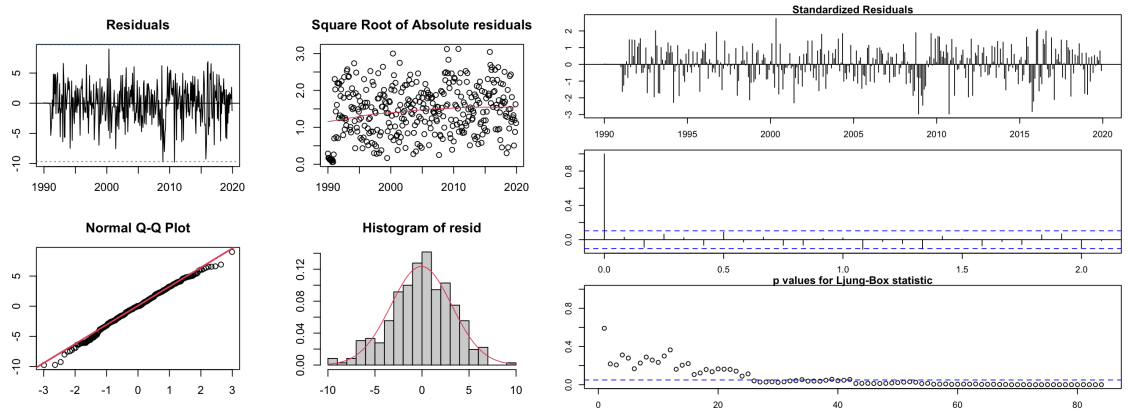
The two different models are estimated in R (R Core Team, 2021). First we fit $(1 - \phi_1 B)W_t = \alpha + (1 + \theta_1 B)(1 + \theta_1 B^{12})Z_t$, which is an $ARIMA(1,0,1)(0,0,1)_{12}$ model with a constant. When fitting this model both the coefficients for the AR_1 part and the mean are non-significant.
Instead the $ARIMA(0,1,1)(0,1,1)_{12}$ without a constant is fitted. The coefficients are now significant and the value for the AIC is 1836.73.
We then fit the other proposed model $(1 - \phi_1 B - \phi_2 B^2)W_t = \alpha + (1 + \theta_1 B^{12})Z_t$, which is an $ARIMA(2,0,0)(0,0,1)_{12}$ model with a constant. Since the mean is not significant the following model is fitted instead $(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{12})X_t = (1 + \theta_1 B^{12})Z_t$, which is an $ARIMA(2,1,0)(0,1,1)_{12}$ model.
With this model the values for all coefficients are significant. The AIC is 1833.74 which is a lower and thus better AIC compared to the first model.
The next step is to validate the two models with significant parameters.

4 Validation

When we do the validation of the two models we check if the residuals satisfy the following assumptions for the residuals:

1. normality
2. constant variance
3. independence

We start by validating the $ARIMA(1,1,1)(0,1,1)_{12}$ model. From Figure 6a we can confirm from both the Normal QQ-plot and the Histogram that the residuals are normally distributed. From the scatter plot of the square root of absolute residuals we can also confirm that they have constant mean and variance. But when we look at the Ljung-Box test in Figure 6b the model fails severly, i.e there are p-values lower than 0.05 which tell us that we reject the null hypothesis indicating independence of the residuals. Due to this, we choose to continue the validation with the $ARIMA(2,1,0)(0,1,1)_{12}$ model.



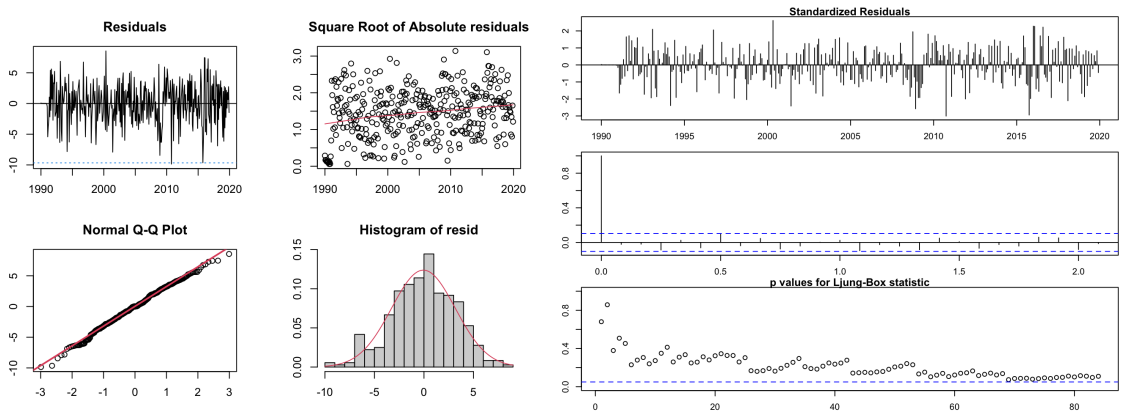
(a) (b)

Figure 6: Validation of $ARIMA(0,1,1)(0,1,1)_{12}$

As for the previous model we can confirm that the residuals are normally distributed and have a constant mean and variance from Figure 7a. This time we also notice from Figure 7b that there are no significant p-values for the Ljung-Box test. Thus we cannot reject the hypothesis that the residuals are independent.

Now that we know that the $ARIMA(2,1,0)(0,1,1)_{12}$ model fulfills the assumptions for the residuals we continue to analyse the expressions of the AR(∞) and MA(∞) models. As the AR Characteristic polynomial roots are both 1.854004 and thus above one we know that the model is stationary. For the MA Characteristic polynomial Roots which all 12 are 1.013068 which also are above 1 we know that the model is invertible.

The next step is to check the stability of the model. To do this we remove the last 12 observations and fit the model for both the whole series and the reduced series and compare the signs, magnitude and significance of the parameters of the models for the different series. As we can see in Table 2 and 3, the magnitude and the signs of the coefficients for the two series are similar and all parameters are significantly different from zero. Therefore we can conclude that the model is stable and can be used for predictions.



(a) (b)

Figure 7: Validation of $ARIMA(2,1,0)(0,1,1)_{12}$

Coeff.	ar1	ar2	sma1
	-0.4478	-0.2909	-0.8557
s.e.	0.0513	0.0516	0.0350

Table 2: Coefficients for whole series

Coeff.	ar1	ar2	sma1
	-0.4292	-0.3068	-0.8673
s.e.	0.0523	0.0524	0.0356

Table 3: Coefficients for reduced series

5 Predicitons

First we want to predict the last 12 months of the series with the $ARIMA(2,1,0)(0,1,1)_{12}$ model. We remove the last 12 observations from 2019, calculate a prediction and a confidence interval for those observations and plot them together with the original series. The plot can be seen in Figure 8.

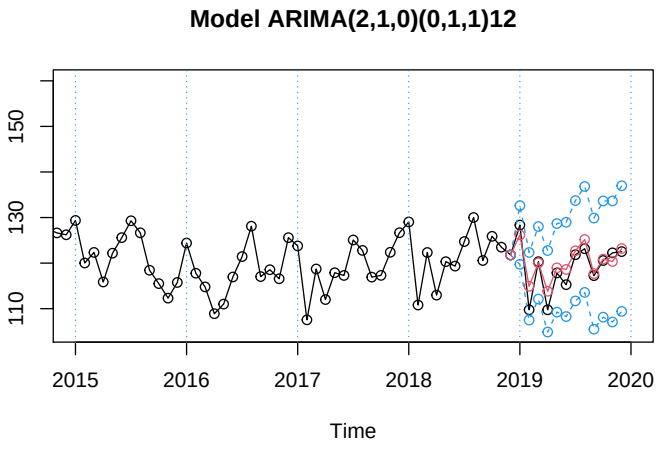


Figure 8: Prediction of observed data 2019

To check and later compare it with different models the root mean square percentage error (RMSPE) and the mean absolute percentage error (MAPE) are calculated. In this case the values are RMSPE = 0.021 and MAPE = 0.016.

We notice that the original series is always inside the confidence interval and the predicted series is quite similar to the original one which is requested for a reasonable model. Next we forecast the 12 unobserved time points for 2020 and calculate a confidence interval for them as well. The result is displayed in figure 11.

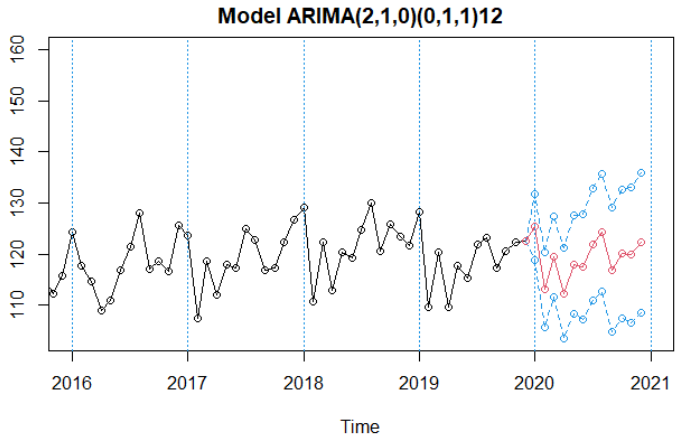


Figure 9: Prediction of future data 2020

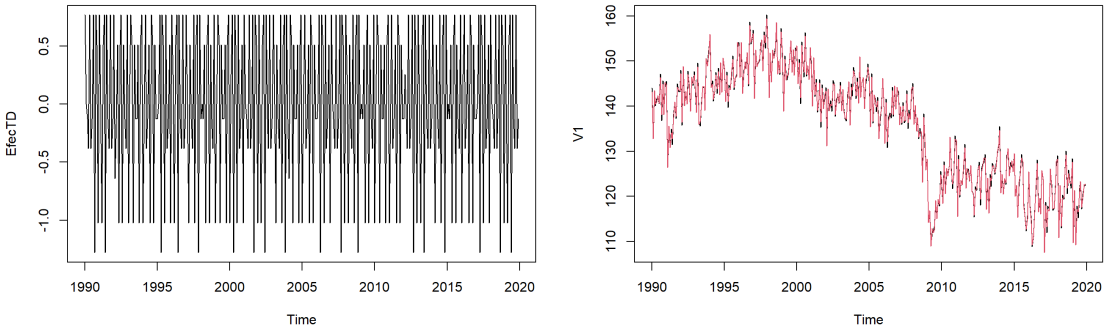
6 Calendar Effect and Outlier Treatment

After we have made the predictions we want to see whether there are any calendar effects or outliers that we have to take care of in order to make better predictions. We start by analyzing the effect of trading days and Easter. The trading days are the number of days in the month that the stock market is open. The difference in the number of trading days from year to year can have an impact on the emissions and therefore on the time series. Easter can be located in either March or April which could also make a difference in the series.

We fit the model to the series but taking the trading days and the Easter effect into account. The model with both effects separately and together are considered. The coefficient for the Easter effect is non-significant in the model including only the Easter effect and including both effects. Therefore, the Easter effect is not taken into account. The model including trading days as a calendar effect has a significant coefficient for that effect. Figure 10a shows the profile of the effect of the trading days.

A new time series without the effect of trading days is created and a new model has to be estimated and validated. Following the same steps as in Chapter 3 and 4 the new model is an $ARIMA(0,1,1)(0,1,1)_{12}$. The plots leading to that model can be found in the Appendix in Figure 16, 17 and 18. Figure 10b shows the original series (black) and the series where the effect of the trading days is removed (red). The next step is the outlier detection.

There are three types of outliers and they all contribute to the series in different ways. A *Level shift* (LS) shifts the whole series either up or down. The series stays around that



(a) profile of trading days

(b) series with (black) and without (red) the effect of the trading days.

Figure 10: Trading day effects

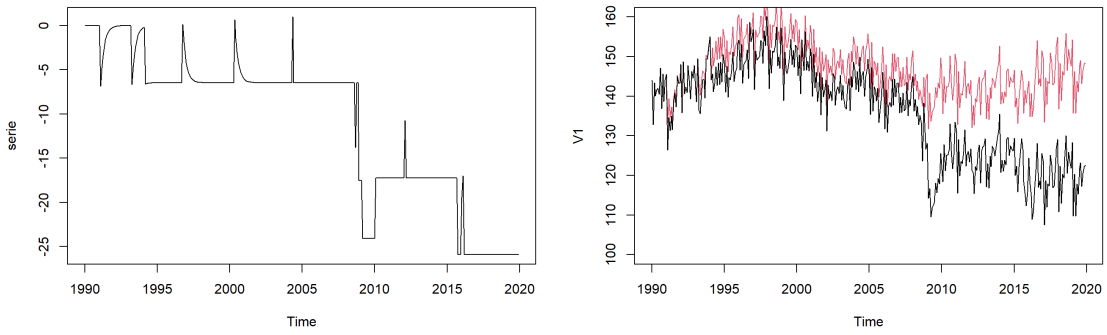
	Obs	type_detected	W_coeff	ABS_L_Ratio	date
9	14	TC	-6.867291	2.869206	Feb 1991
13	40	TC	-6.658615	2.913762	Apr 1993
12	51	LS	-6.464667	2.895178	Mar 1994
14	82	TC	6.577298	2.911797	Oct 1996
6	125	TC	7.149050	2.886624	May 2000
4	173	AO	7.442664	3.131095	May 2004
5	225	AO	-7.331548	3.122850	Sep 2008
1	228	LS	-11.060704	4.296762	Dec 2008
10	231	LS	-6.587483	2.877064	Mar 2009
7	242	LS	6.881041	2.903492	Feb 2010
8	266	AO	6.485825	2.849909	Feb 2012
3	310	LS	-8.688132	3.459232	Oct 2015
11	313	AO	6.495974	2.925275	Jan 2016
2	314	AO	8.895463	3.582324	Feb 2016

Figure 11: Different outliers detected

level. *Additive outlier* (AO) affect the series only at one data point. This is common if there is something special happening on one day that affects the result for that day only. *Transitory change* (TC) affects the series for multiple time points but with decreasing intensity over time. The type and the date of the outliers can be found in Figure 11.

The outlier with the highest impact (highest value in w_coeff) is a level shift in December 2008. This might be due to the fact that the first commitment of the Kyoto protocol started in 2008. It had a negative impact meaning that there is less CO_2 emission after that date than before. The negative additive outlier in September 2008 might be due to the financial crisis which had an impact on the economy and therefore possibly also on the CO_2 emission.

After finding the outliers the time series can be linearized, meaning that the effect of the outliers are removed for better predictions. The function showing the linearization is: $X_t = \sum_{i=1}^{h} \beta_i V_{i,t} + \tilde{X}_t$, with X_t being the observed series, $V_{i,t}$ the exogenous variable and $\tilde{X}_t$ the theoretical series without effect of the outliers. In Figure 12a the profile of the outliers is displayed showing the different impact of the three types. Figure 12b presents the original time series X_t (black) and the one where the outliers are removed $\tilde{X}_t$ (red).



(a) The impact of the outliers

(b) The series with and without outliers.

Figure 12: Outlier effects.

We fit a new model to the series without the effects of the outliers. With the ACF and the PACF (Appendix: Figure 19) we discover several possible models. When checking validation and significance for all the models we choose to continue with the $ARIMA(0,1,1)(0,1,1)_{12}$-model since it had the lowest AIC-value, all parameters were significant and it passed the Ljung-Box test (Appendix: Figure 20 and 21). We check the stability of the coefficients of the model in the same way as before and we consider this model to be stable because the coefficients are similar in sign, magnitude and significance.

Comparing the first found model and the one treated for the calendar effect and the outliers in Figure 13 we can see that the profile of the predictions are similar. But the confidence intervals are narrower for the latter. So we get more precise predictions with the last model. For a more in depth comparison of the two models different measures are used. They are displayed in Figure 14. We can see that the AIC and BIC are smaller for the model with treatment of outliers and calendar effect, showing a better model fit. Also the root mean square percentage error (RMSPE) and the mean absolute percentage error (MAPE) being smaller for the last model shows that this model is giving more accurate predictions. Lastly, the mean length of the confidence intervals (meanLength) is smaller. Therefore, all measures lead to the conclusion that the $ARIMA(0,1,1)(0,1,1)_{12}$ with a treatment for trading days as calendar effect and treatment for outliers is to prefer over the $ARIMA(2,1,0)(0,1,1)_{12}$ for forecasting.

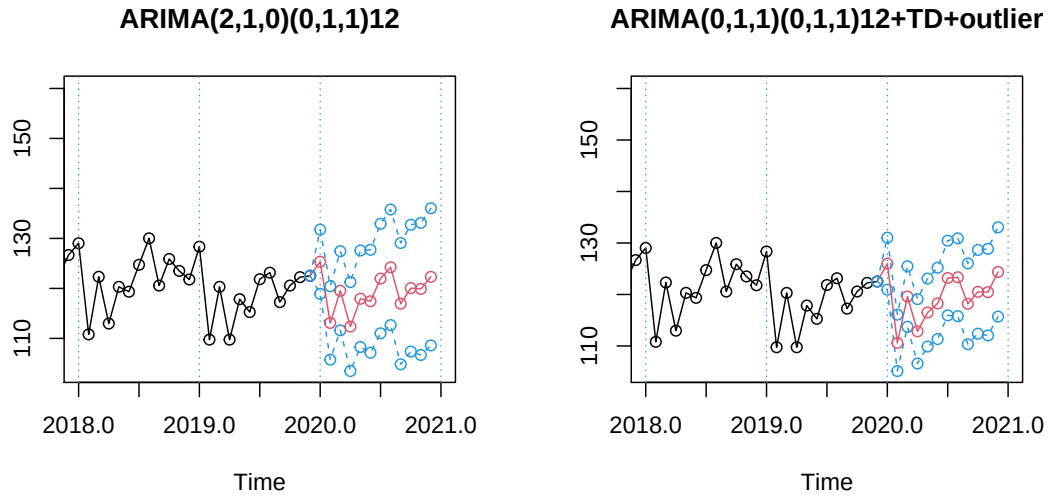


Figure 13: Comparison of the first model and the model considering the calender effects and outliers.

	par	Sigma2Z	AIC	BIC	RMSPE	MAPE	meanLength
ARIMA(2,1,0)(0,1,1)12	3	10.745501	1832.807	1848.204	0.02125953	0.01604009	20.79713
ARIMA(0,1,1)(0,1,1)12+TD+outlier	17	6.640788	1691.693	1760.981	0.01342432	0.01150628	14.00928

Figure 14: Comparison of the different models.

7 Discussion

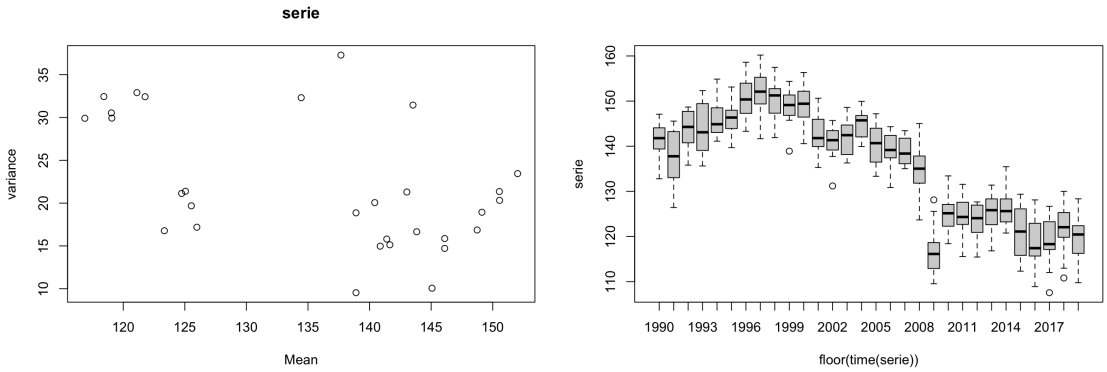
We propose an $ARIMA(0,1,1)(0,1,1)_{12}$ model with treatment for trading days effect and outliers as the best found model to perform predictions for the CO_2 emissions from the industrial sector in the USA. A problem with predictions for 2020 is the Covid-19 pandemic which is not taken into account in the modeling process. It is reasonable to expect that the predictions for 2020 are not accurate. In a few years the pandemic can be included as an outlier. But since it is a recent event and the full effect is not yet known, we cannot adjust for it.

References

Box, George EP, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung (2008). *Time series analysis: forecasting and control*. John Wiley & Sons.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

A Appendix



(a) Mean Variance plot (b) Box plot.

Figure 15: Checking for constant variance in original series

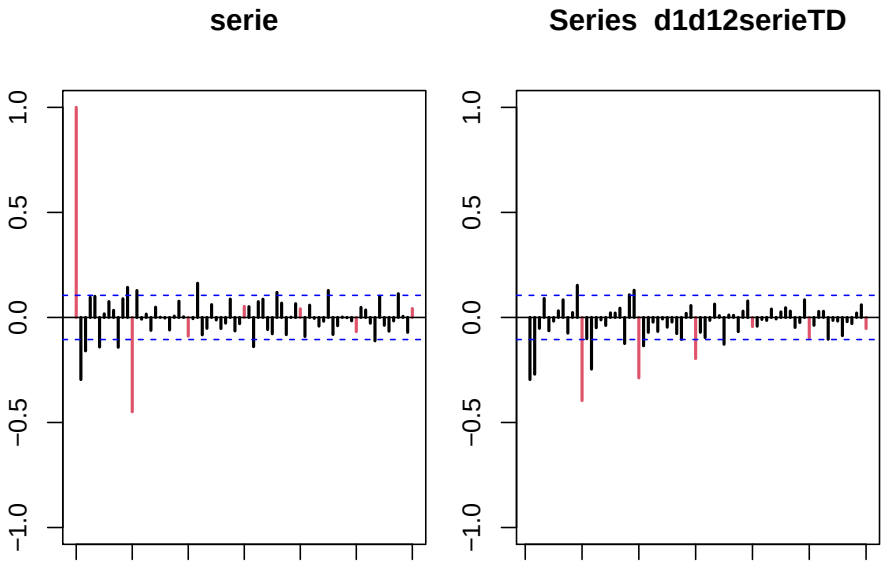


Figure 16: ACF and PACF to find models for the series treated for trading days

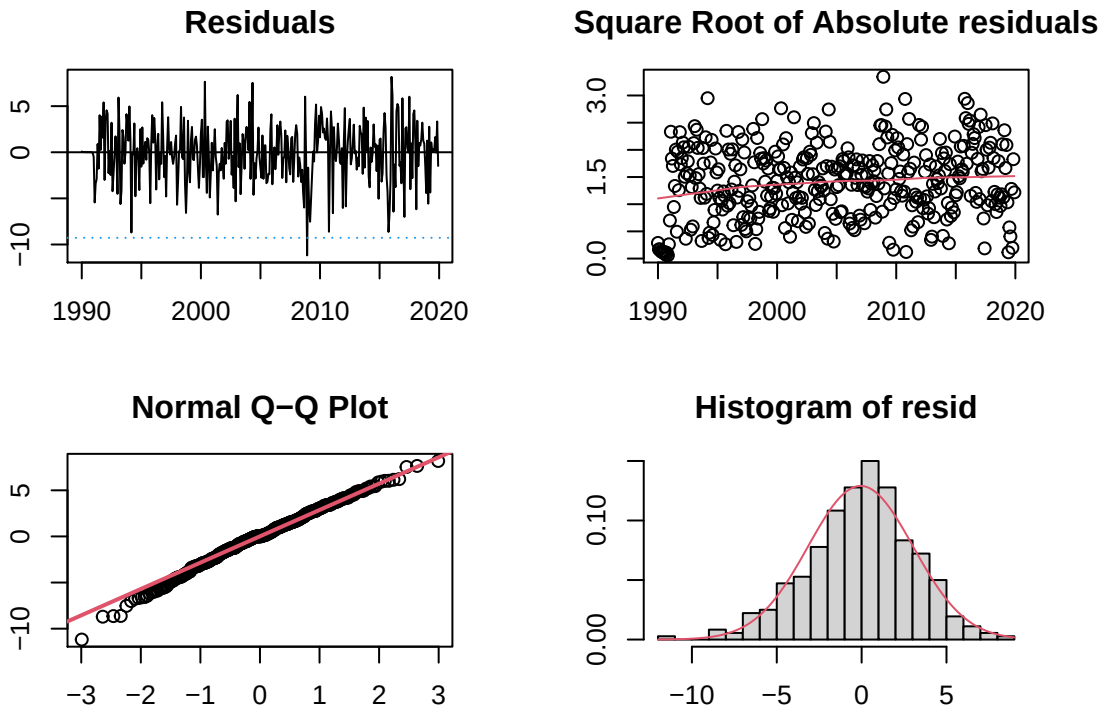


Figure 17: Validation of the $ARIMA(0,1,1)(0,1,1)_{12}$ for the series treated for trading days

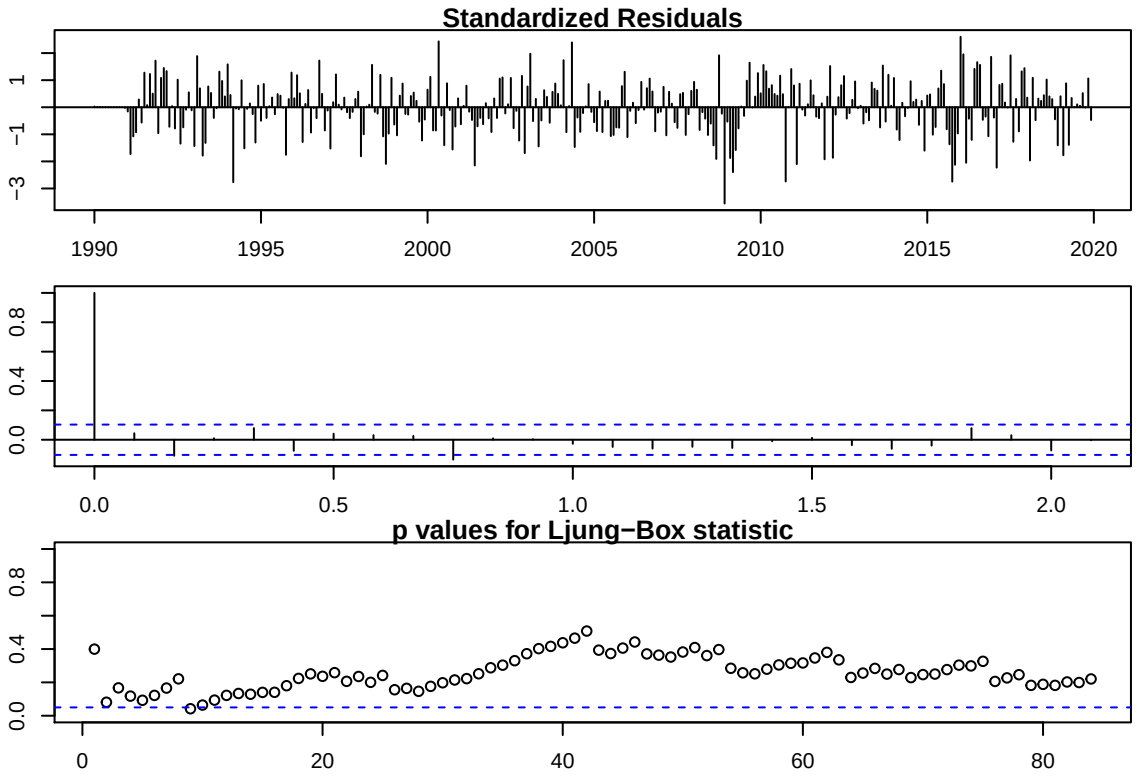


Figure 18: Validation of the $ARIMA(0,1,1)(0,1,1)_{12}$ for the series treated for trading days

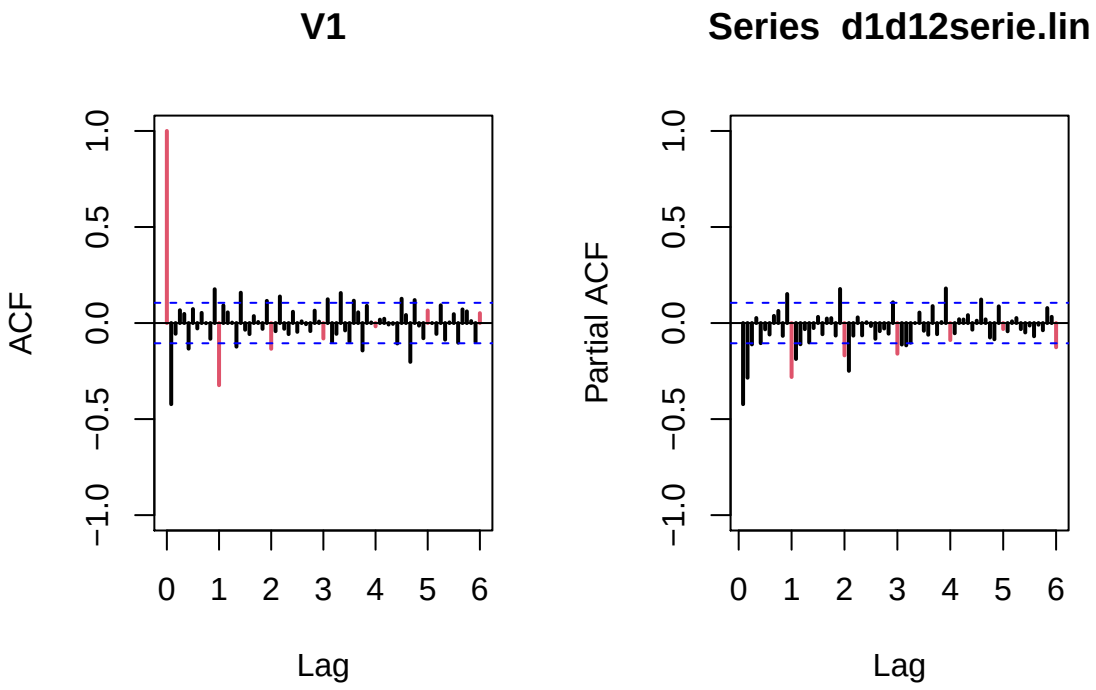


Figure 19: ACF and PACF of the linearized series

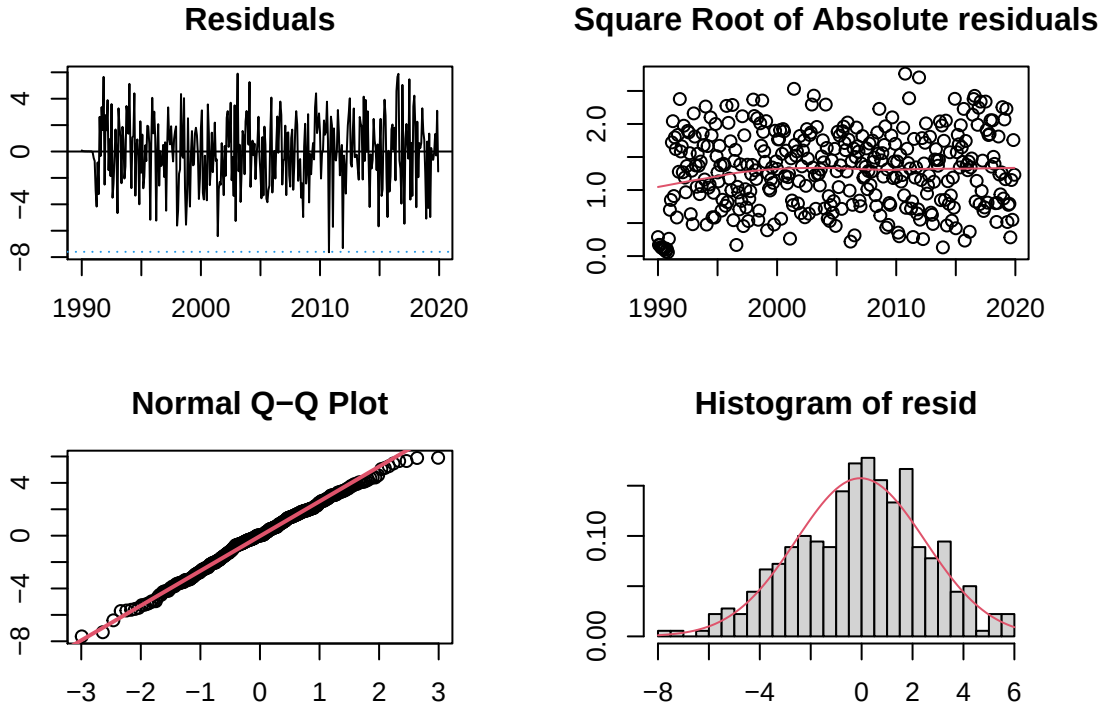


Figure 20: Validation of $ARIMA(0,1,1)(0,1,1)_{12}$ for linearized series

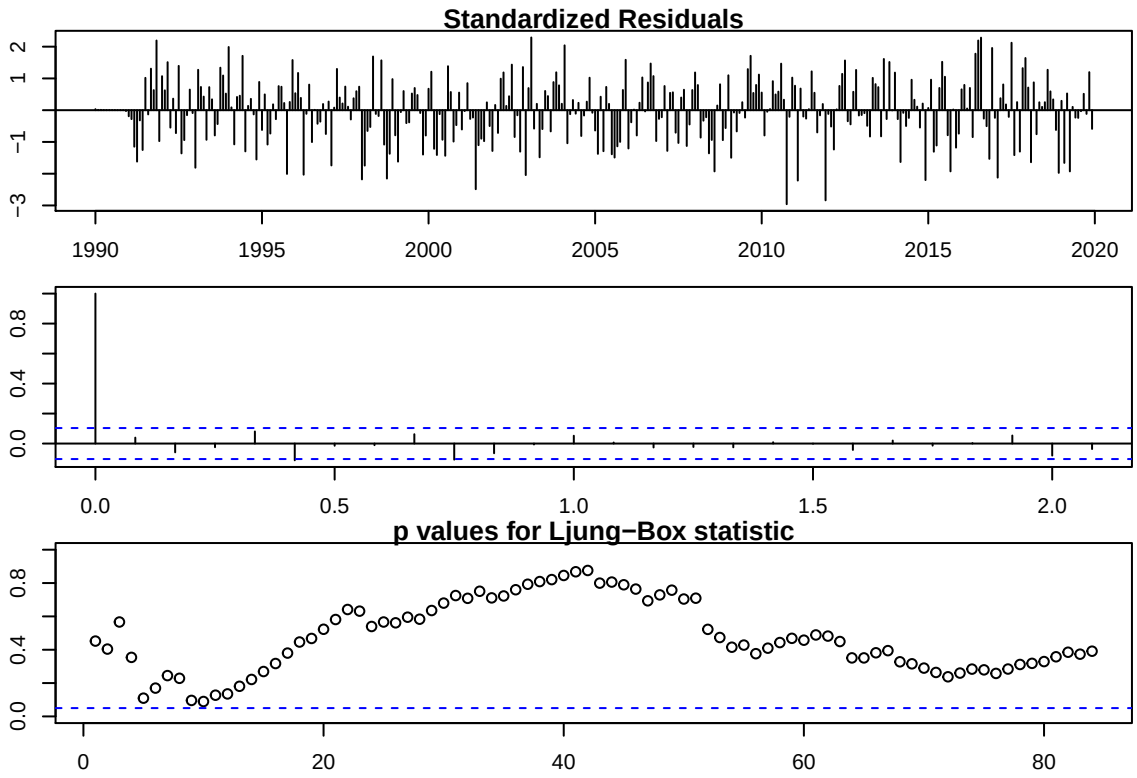


Figure 21: Validation of $ARIMA(0,1,1)(0,1,1)_{12}$ for linearized series