Predicting CO_2 Concentration using SARIMA Models

Forecasting Methods - Professors Bruno Damásio and Carolina Vasconcelos

Dinis Gaspar 20221869, Dinis Fernandes 20221848, Luis Mendes 20221949

Introduction

Rising CO2 concentration is one of the most worrying problems our generation and future generations face. This increase leads to multiple changes in climate factors, which in turn leads to an increase in violent natural disasters among other consequences. For this reason, it is crucial to predict the evolution of this concentration in future years. To this end, we will show how SARIMA models can be used to predict CO2 concentration from past values using R.

Data

The dataset we will use during this project includes monthly data regarding CO2 concentration in parts per million (PPM) and percentage change in concentration from March 1958 to December 2023 [1]. For our models, we will use the concentration in PPM because it will give us more valuable, absolute, and interpretable predictions.

Methodology

Stationarity:

Firstly, we're going to start by analyzing our time series.

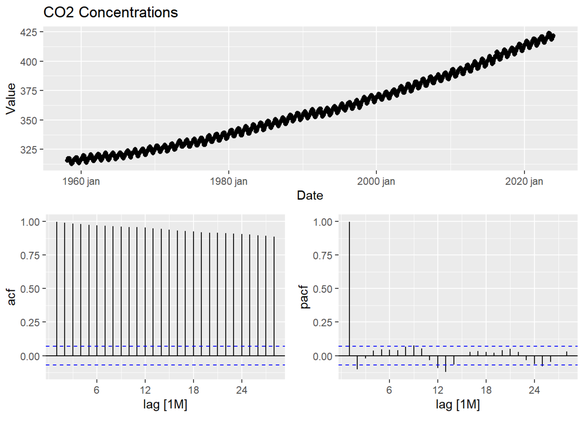

Figure 1: CO2 Concentrations Time Series

We can clearly see that there is trend and seasonality in our data. For these reasons, we can conclude that our series is not stationary which is confirmed by applying an ADF test with 46 lags (starting from 48 and cutting until a significant p-value is reached on the final lag) where our observed test statistic was -0.1601 and the critical value was -3.41, for a 5% significance level.

To achieve a stationary series, we started by applying seasonal differences to our data, looking at the resulting series and its correlogram we could see that the series was still not stationary, we confirmed this using through an ADF test with 46 lags (starting from 48 again), where our observed test statistic was -3.1787 and the critical value was -3.41, for a 5% significance level.

The next transformation we applied was a simple difference. After applying this difference, we once again analyzed our series and its correlogram which appeared to indicate that it was now stationary, we then applied another ADF test with 47 lags (starting from 48), where our observed test statistic was -7.3139 and the critical value was -1.95, for a 5% significance level. This allowed us to conclude that our series was now stationary.

Model Specification and Evaluation:

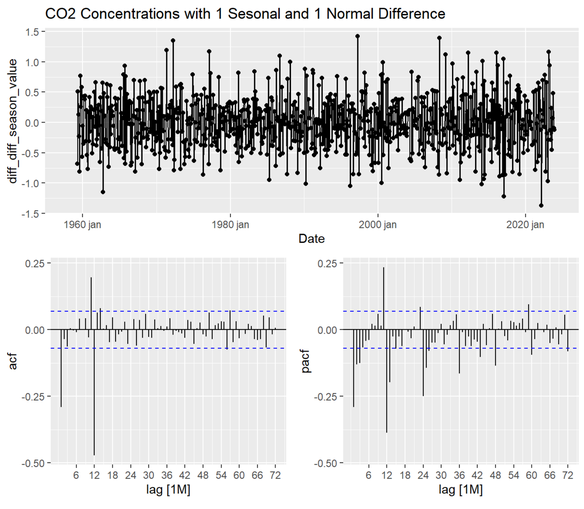

Figure 2: CO2 Stationary Time Series

Now we once again analyzed the correlogram of our final series to look for candidate models, we selected two models that we believed could be adequate, these models were a SARIMA(0,1,1)(0,1,1) and a SARIMA(0,1,1)(0,1,2), because looking at the ACF and PACF it was clear to us that the non-seasonal component was a Moving Average of order 1 (with a difference), looking and the seasonal component we believed it was a Moving Average of order 1 (with a difference), because the PACF (normal and the seasonal part) are decaying exponentially after the first lag and the ACF is cutting off after the first relevant lag, but we decided to also include a Moving Average of order 2 (with a difference) as a possibility to make sure we had multiple candidate models.

To analyze which of our candidate models performed better we decided to split our dataset into train, which included data up to 2018, and test, which included the data after 2018. Then, we fitted our models to our train set. Firstly, we looked at the Information Criteria where the SARIMA(0,1,1)(0,1,1) had slightly better results. As the results were not overwhelmingly convincing, we decided to predict the 5 years that corresponded to our test set and compare the predictions with the real values.

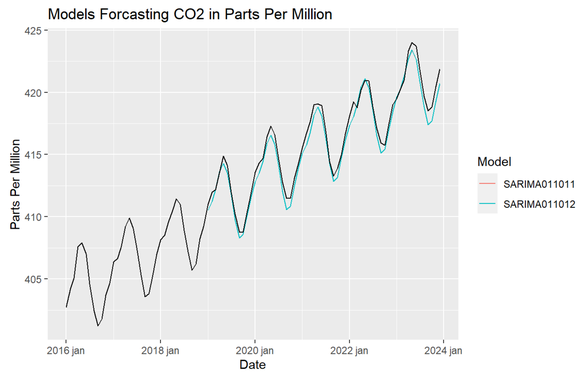

Figure 3: Model Comparison

We concluded that the models were making identical predictions and were both quite accurate to the real values, using multiple accuracy metrics (RMSE= 0.641, MAE= 0.562, MAPE= 0.135 and MASE= 0.360, equal for both). With all this in mind the model we chose as a final model was the **SARIMA(0,1,1)(0,1,1)** because it had slightly better information criteria (with AIC= 377 vs 379, AICc= 377 vs 379 and BIC= 391 vs 398) and it is less parameterized.

Residuals:

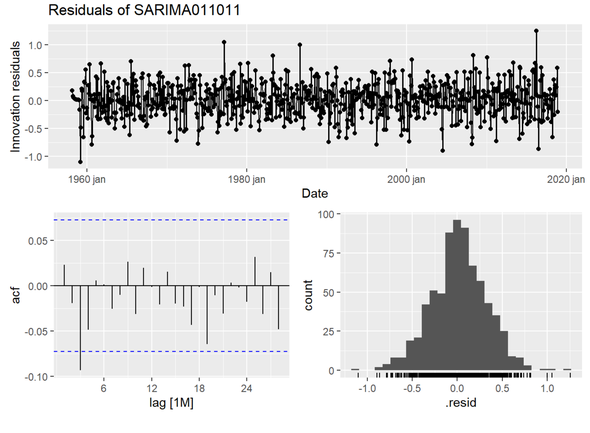

Figure 4: Residuals of SARIMA(0,1,1)(0,1,1)

Now that we have our final model, we will look at its residuals and we can see that they are like white noise and follow a normal distribution. We then run the ljung_box test to confirm that our residuals are not autocorrelated, with a p-value of 0.531, H0 is not rejected for all relevant significance levels, and thus the residuals are not autocorrelated, so we have a reliable model.

Conclusion

Now that we have a final model, we will predict the next 5 years to see if there are signs of the increase in concentration slowing down or if it is predicted to keep increasing.

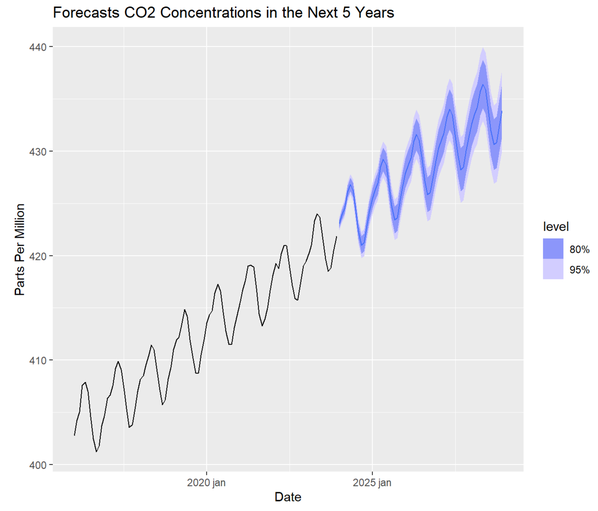

Figure 5: Next 5 Years Forecast

Looking at the graph above, we can clearly see that there is no sign of the increase in CO2 concentration slowing down. This is very worrying, because this means that the consequences and disasters caused by climate change, which is heavily affected by this increase, are likely to remain present and even become more frequent and destructive.

References

- 1: Atmospheric CO_2 Concentrations, by IMF.