# ANLY 565 Time Series and Forecasting (Project Carbon Emission).

Rahul Awale, Sonachi Mogbogu, Uttara Suryavanshi[1]

[1] Harrisburg University of Science and Technology

Author Note

ANLY 565-90- O-2021/Spring - Time Series and Forecasting

Correspondence concerning this article should be addressed to Rahul Awale, Sonachi Mogbogu, Uttara Suryavanshi, 326 Market St, Harrisburg, PA 17101. E-mail: USuryavanshi@myharrisburgu.edu, rawale@my.harrisburgu.edu, smogbogu@my.harrisburgu.edu

## Introduction

Global warming has been the most alarming concern for everyone on this planet in the past decades. The sharp rise in the global temperatures due to ever increasing greenhouse gases' emissions across the planet present a unique and formidable challenge. The effects of this climate change due to global warming can already be seen in the ecological, physical and socioeconomic impact that it has had, driving extreme weather and climate events across the globe. This study concentrates on the Carbon Dioxide Emissions, by developing a analytical and statistical models to predict the rise of this particular greenhouse gas in our atmosphere in the United States. "Carbon dioxide is invisible- no color, no odor, no taste. It puts out fires, puts the fizz in seltzer, and is to plants what oxygen is to us. It is hard to think of it as a poison." (Verhulst 2007). The effects of global warming due to rise of greenhouse gases, including CO2 primarily, is a global concern which is affecting the climate and weather patterns in the far north and far south of the Arctic and the Antarctic Oceans. CO2 is a very dangerous gas in this aspect, since it has the capability to trap the heat generated by the planet and its occupants, via burning fossil fuels, deforestation, and natural causes such as volcanic activities and respiration of all living organisms. Humans have increased the atmospheric CO2 levels by 47% since the Industrial Revolution began,the genesis of climate change.

The consumption of all fossil fuels such as coal, natural gas, and oil for electricity and heat for the betterment of every economy is the single largest source of global greenhouse gas emissions. "Global carbon emissions from fossil fuels have significantly increased since 1900. Since 1970, CO2 emissions have increased by about 90%, with emissions from fossil fuel combustion and industrial processes contributing about 78% of the total greenhouse gas emissions increase from 1970 to 2011. Agriculture, deforestation, and other land-use changes have been the second-largest contributors" (Global Greenhouse Gas Emissions Data. (n.d.), 2021).

Time series forecasting is the method of finding out and analyzing time-series data. gathered over a predetermined time period. This technique is widely used for forecasting values and predict future trends. A study in time series is frequently conducted in stock market predictions, weather forecasting, earthquake prediction, Web Traffic management, and among many others.

## Motivation

The two articles “Apple Commits to be 100 percent carbon neutral by 2030” published on Apple’s website and “Samsung to Offset Lifetime Carbon Footprint of All Washing Machines” published on Samsung’s website were the main motivation behind our research. The headlines grabbed our attention because emphasized just how important it is to control carbon emissions. Both these articles talk about the companies reducing their carbon emission by using renewable energy and not producing devices that lead to more carbon waste in order to clean the environment. Therefore, for our research, we want focus on carbon emission generated from electricity generation to help stop global warming. We can do this by predicting which carbon emission source type is the main contributor for carbon emission in the United States. The prediction can further be used for awareness purposes, so that other companies and people follow Apple and Samsung.

## Research Question

Our main research question is will carbon emission increase in the next four years? We want to be able to predict four years of carbon emission generate by electricity using the best forecasting method. Some other questions we are trying answer are: is there any correlation between type of carbon emission source and carbon emission value? What is the main source type of carbon emission? And what is the trend for carbon emission generated by electricity?

## Data

We are using a public dataset of monthly carbon dioxide emissions from electricity generation at the Energy Information Administration and Jason McNeill. The dataset includes CO2 emissions from each energy resource starting January 1973 to July 2016.

First, we retrieved the monthly CO2 emissions dataset to visualize the dataset to decide. The dataset has six columns where 2 of them are integer data types and four objects, and 5096 observations. We converted the "YYYYMM" column to date format and the Value column to numeric. Then we looked for missing values in the data set, and there were 387 missing values for the YYYYMM field,416 missing values for the Value field. The data was divided into two sets: the missing values fields and the non-categorical data. The mice function resolved the missing values. After resolving the missing values, we bound the subset into a single set.

We used a histogram to check the distribution for the normality test. From this, we performed skewness, kurtosis, and Shapiro test. The p-value was less than 0.05 that implied not normally distributed. We transformed the data by applying log transformation, which resulted in equally scaling the data. After the transformation, we can see that data was multinominal distributed and ready for analysis to be carried out on.

Then we recorded the description data since the description was too long. After the description was reduced, it was easier to plot the distribution of CO2 emissions of different energy types. We also checked the correlation between emission value and different energy types using "Pearson," "Kendall," and "Spearman". We didn't find any significant correlation between them.

## Analysis and Methodology

This section contains detailed information about the exploratory data analysis and time series analysis for this project.

### Exploratory Data Analysis:

We used ggplot2 package in R to complete our exploratory analysis. First, we started by creating a boxplot. The boxplot allowed us to summarize the variations in our dataset. For this research, we filtered our box plot by using different types of carbon emissions. Doing this, we were able to summarize how each carbon emission variable type in our dataset was distributed based on their value, mean, max, and min.

Second, we created histograms. We decided on three different histograms with different values for x. The first shows the distribution of data based on date. The second histogram shows data distribution based on carbon emission value. The third histogram shows the data distribution based on the type of emissions.

Third, we created a correlation plot. To start, we had to create a new individual variable for “Value” as emission and another new individual variable for “Column_Order” as type_no. Next, we created a new data frame using the two individual variables and named it “emission_type.” We tested emission_type using Pearson’s correlation method to see any correlation between emission (carbon emission value) and type_no (the number assigned to each type of emission source). The result for the correlation was -0.058, which showed that there was no significant relationship between the two variables. Lastly, we created a correlation plot and performance matrix to visualize the data.

Fourth, we created two bar charts. The first bar chart shows the carbon emission value based on date, where fill= type of emission. This bar chart shows how each type of carbon emission has changed over the last forty-three years. The second bar chart is a horizontal bar chart showing how much each emission source type contributes to the total carbon emission. Using this bar chart, we can see which is the main contributor to carbon emission generated by electricity in the United States.

Fifth, we created a scatterplot using the date as the independent variable(x) and carbon emission as the dependent variable(y). We further used the emission type as a color fill to improve our visualization. Using the scatterplot, we were able to see the distribution of all the points in our data, where each color represented a type of carbon emission source.

Finally, we created a line chart. We again used the date as the independent variable(x) and carbon emission as the dependent variable(y). We used the same color fill to differentiate the types of carbon emission sources. Using the line chart, we were able to see the trends for different types of carbon emission sources over time.

## Time Series Analysis

We used time-series analysis to predict future values for carbon emission because we realized our dataset was a time data structure. Our entire approach involved looking at events in the past, studying the present pattern, the time series analysis part, and then applying them to the future to forecast and make business decisions. We conducted a univariate time series, and we focused on linear models. In the linear family, we used the ARIMA models and exponential smoothing, which are vitals in the time series analysis and commonly used. Also, we put into other model considerations such as naive methods, holt's trend methods - which is like an extension of the exponential smoothing method but focuses on the trend component while producing forecasts. Last is the TBATs model, which is just a combination of several components; it was useful to add because we have seen in recent research works to be a good choice for forecasting.

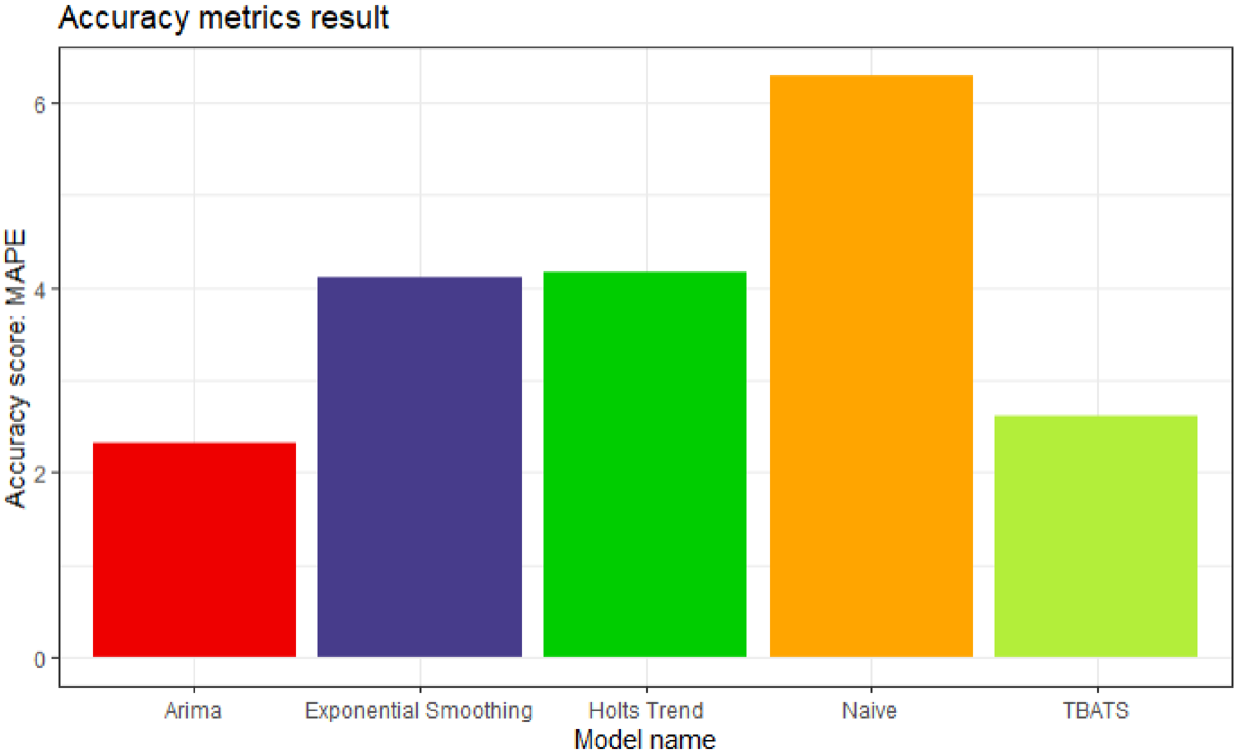
The time series analysis started with converting our simple vector into a time series vector and plotting to see if it is stationary or not. It was evident that from the autocorrelation plot, one can see the box plots leaving the 95 percent confidence boundary crossing the threshold from the autocorrelation. From this, it is to note that there is autocorrelation and resulting in data being non-stationary. To make the data to be stationary, enforcing some techniques will be the best approach, and from there, we can carry out more operations. We would be using the total_emission object without filtering on a specific carbon emission type. We used the Augmented Dickey-Fuller test to verify the structure of our time series. At first glance, we saw that the p-value from the test was 0.9828, which is more than the significance level of 0.05. Therefore we fail to reject the null hypothesis.

To transform the non-stationary data to stationary, we attempted the log differencing by even adjusting the differencing in lags. We noticed it was yielding an outrageous value for our model accuracy. In the effort to completely transform the non-stationary data to stationary, since we did not continue with differencing technique, we decided to use one of the most commonly used techniques: Detrending, which we just performed by removing the trend component from the time series and that made it stationary with a p-value of 0.01 which is less than the significance level; therefore, the null hypothesis is rejected with a higher significance level, and we noticed the test-statistic was smaller. For once, you could see that there was a constant mean and variation. However, autocorrelation is still present since some box plots leave the 95 percent confidence boundary crossing the threshold from the autocorrelation.

We executed an 80-20% data partitioning consisting of training and testing dataset for model validation. Then, we created a linear model 12 months ahead of the forecast of the carbon emission value, and we found out that using the mean absolute error to calculate the forecast that the model seems accurate enough with an error rate of 2.4856%. Then we decided to compare and contrast by utilizing the MAPE to evaluate the performance of the foresting models mentioned above. The result section provides more information about our discoveries.

## Results

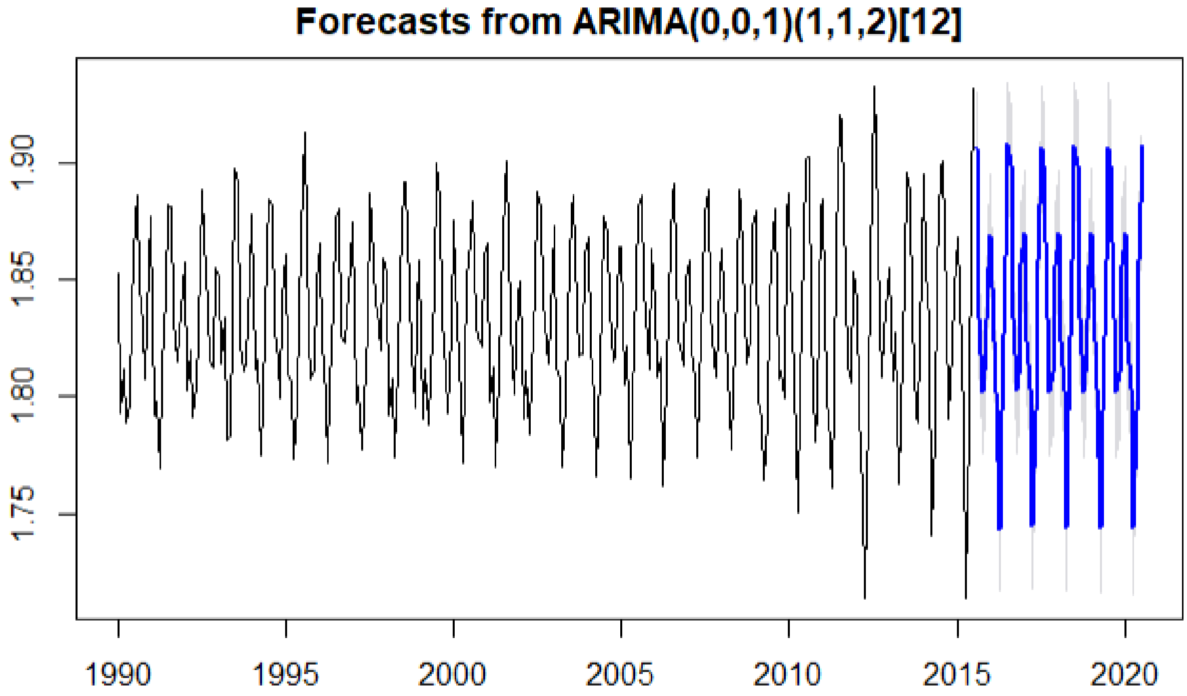
The figure below shows the models used and their accuracy metrics. We used the MAPE to measure each model's performance and see which model outperforms its ability to make accurate predictions. Although we could have considered other metrics like the Root squared mean error (RMSE) or the Mean absolute error (MAE), or the Mean absolute scale error (MASE), the MAPE is commonly used as the overall accuracy metrics in time series analysis because it is unit-free. With this one line of code: accuracy(model_name), we can see all the metrics results.



A Model Accuracy Result

The Naive, Simple Exponential Smoothing, and Holts Trend models did well at equally achieving a lower MAPE of 6.28 percent, 4.12 percent, and 4.16 percent, respectively. However, ARIMA and TBATS model emerged as the winner based on their performances on the test data with MAPE close to 2.31 percent and 2.60 percent, respectively. Therefore, we would use the Arima model to predict the next four years.

Next, we predicted the change in Total Carbon Emissions Using the ARIMA for the next four years. We compared the Carbon Emission Value in July 2016 to July 2020 and found that the percent change in Carbon Emission increased by 5.70 percent.



Forecast from Arima for next 5 years

## Conclusion

In conclusion, we found no trend between the type of Carbon Emission source and the Carbon Emission Value as the correlation was non-significant. Next, we discovered that Coal is the primary source of Carbon Emissions, followed by Natural Gas. Although the overall carbon emission has a downtrend, we need to focus on ways to reduce different types of carbon emissions, mainly Coal and Natural Gas, to reduce the impact on Global warming. Also, we tested various models such as Naive Forecasting, Simple Exponential Smoothing, Holt's Trend Method, ARIMA, and TBATS. We found ARIMA to be the best model with the lowest absolute mean percentage error of 2.31 percent.

Furthermore, we used ARIMA to predict total Carbon Emissions by Electricity

Generation for the next four years. Based on the results, we predicted that Carbon Emissions to increase by 5.7 percent between July of 2016 and July of 2020. But this is only expected if all the factors remain unchanged. However, large companies such as Apple and Samsung are starting Carbon Emission initiatives to lower their carbon footprint starting last year. Other companies also are following in their footsteps. Therefore, we can expect Carbon Emissions by Electricity Generation to decrease in the coming future.

## References

Apple commits to be 100 percent carbon neutral for its supply chain and products by 2030. (2021, March 26). Retrieved April 22, 2021, from https://www.apple.com/newsroom/2020/07/apple-commits-to-be-100-percent-carbon-neutral-for-its-supply-chain-and-products-by-2030/

Global Greenhouse Gas Emissions Data. (n.d.). Retrieved April 22, 2021, from https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

Samsung to Offset the Lifetime Carbon Footprint of All Washing Machine and Tumble Dryer Purchases. (2021, March 24). Retrieved April 22, 2021, from https://news.samsung.com/uk/samsung-to-offset-the-lifetime-carbon-footprint-of-all-washing-machine-and-tumble-dryer-purchases

Verhulst, J. (April 22, 2007). “Feeling The Heat”. St. Petersburg Times. St. Petersburg, Florida.