Logo

Description automatically generated

**Oklahoma State University**

**MS, Business Analytics and Data Science, 2021-23**

**ANALYSING AND FORECASTING THE CARBON FOOTPRINT IN THE UNITED STATES USING A TIME SERIES DATA MINING MODEL**

**Rupom Bhattacherjee**

**Kodjo Opoku Botchway**

**April 11, 2022**

Contents

[ABSTRACT 3](#_Toc100611995)

[INTRODUCTION 3](#_Toc100611996)

[METHODOLOGY 5](#_Toc100611997)

[Business/ Data Understanding 5](#_Toc100611998)

[Dataset Details 5](#_Toc100611999)

[Data Transformation 8](#_Toc100612000)

[Exploratory Data Analysis 9](#_Toc100612001)

[Transforming data to stationary 11](#_Toc100612002)

[Moving average smoothing 11](#_Toc100612003)

[Differencing 11](#_Toc100612004)

[Modelling 12](#_Toc100612005)

[Optimal parameters selection for modelling 12](#_Toc100612006)

[RESULTS 14](#_Toc100612007)

[CONCLUSION 16](#_Toc100612008)

[APPENDIX 17](#_Toc100612009)

[REFERENCES 17](#_Toc100612010)

# ABSTRACT

The ever-increasing concentration of the quantity of greenhouse gasses in the atmosphere continues to drive alarming environmental and associated consequences in relation to global warming and climate change. The threat of global climate change and its effects has seen a large variation of public ideals and perspectives regarding the perception that that only measurable is manageable. Hence, the intensity with which governmental bodies, people and procedures view the carbon footprint differ across. A myriad of approaches has been developed and explored to predict and forecast the footprint. The carbon footprint is a definite and indefinite measure of the exclusive total amount of the CO2 emissions that is directly and indirectly caused by the accumulated activities or occurrences over the life stages of a given product.

This study incorporates various time series forecasting models to forecast the carbon footprint for the United States. CO2 emission data from different sectors (e.g., buildings, industries, transport, and electricity and heat) over 1995 to 2018 were used to make the five-year and ten-year forecasts of CO2 emission. Different matrices such as RMSE, MAPE, and MAE were used to compare the performance of the models and select the best one. To forecast CO2 emission from energy usage in residential and commercial buildings, Auto Regressor is found to be the best model. Auto Regressive Moving Average (ARIMA) performed better for the industries data and the transport. The predictions of the emissions from electricity and heat were also best modelled by the Auto Regressor model.

The results are indicative of the increasing carbon emissions in the electricity and heat sector. All the other sectors show a relatively downward forecast or a stagnant trend. This information is important because the implications of this confirmed forecasting would elaborate the threat on our environment given the worst of the forecast. In focusing on the reduction of the carbon emissions across all sectors, the electricity and heat sector would be a great place to begin for solution evaluation.

# INTRODUCTION

Humans generally benefit a lot from the processes that cause the emission of greenhouse gases (GHG) since these produce several goods and services. The benefits of these and with the liability that comes with the emissions differ by purpose or the category under which consumption is done. The responsibility for this, though spread and distributed unevenly across nations and organizations, also falls on individuals to contribute their quota of effort to make this as controlled as possible. Businesses, consumers, and policy makers are all drawn in with interest by the concept of a carbon footprint. The carbon footprint of an investor's portfolio is monitored as a risk indicator. Purchasing managers are interested in learning more about their supply chains' carbon footprints, and customers are increasingly being offered carbon-labelled items. Even though the phrase "carbon footprint" is somewhat an inaccurate representation or misleading term, it refers to the mass of CO2 emissions accumulated over time, such as through a supply chain or a product's life cycle, rather than some form of area measurement. The carbon footprint is defined by(Wiedmann & Minx, 2008) as the definite and indefinite measures of the total amount of CO2 emissions caused directly or indirectly by cumulative activities or occurrences over the given life cycle of a product. On a more global scale, majority of the greenhouse gas emissions come from home consumption constituting about 72 percent, while government consumption accounts for about 10 percent and investments about 18 percent. GHG emissions for food contributes about 20%, whereas residential operations and maintenance accounts for 19% and mobility, 17%.(Hertwich, 2009) In developing and underdeveloped nations, food and services are more significant, but in developed countries, mobility and manufactured goods increase rapidly with money and take priority. The value of public services and produced commodities has not been adequately recognized in policy. As a result, policy priorities are determined by the state of development and country-level features. The average national per capita footprint ranges from 1 tCO2e/y in Africa to 30 tCO2e/y in Luxembourg and the United States (Dunn, 2010; Ritchie, 2019).

Several regulations have been set in place in the United States to monitor and control the future emissions of these gasses in the atmosphere. Monitoring the current emission and forecasting it for the near future would help identify the potential sources of CO2 emission, both globally and within different private sectors. Given CO2 is emitted from any sectors, implementing this practice would make the companies understand the current situation and develop proper carbon measuring and reducing programs for better sustainability. This is imperative not only because the companies should care about the climate, but also because quantifying CO2 emission would help identify excessive energy usage and make the business process more cost effective.

There have been several works done in past to forecast CO2 emission globally, country wise or even sector-wise using time series. Akyol and Uçar (2021) forecasted CO2 emissions in Turkey from 2018 to 2030 using a time series forecasting algorithm considering population, gross domestic product, energy production, and energy consumption as the independent variables (Akyol & Uçar, 2021). The result indicated a gradual increase in current greenhouse gas emissions. The authors also suggested some courses of action (e.g., structural, and technological transformation) to reduce the emissions.

Hosseini et al. (2019) did the forecasting of CO2 emissions in Iran from 2015 to 2030 using multiple linear regression (MLR) and multiple polynomial regression (MPR) considering population, CO2 intensity, GDP per capita, the share of fossil fuel in electricity production, and per capita energy use as the primary sources to emit CO2 in Iran (Hosseini et al., 2019). Several time series algorithms were used to forecast the predictors from 2015 to 2030 using the data from 1971 to 2014. The result indicated a steady increase in current CO2 emissions if there were no significant changes in the government's current policies.

There was a research paper that focused on the carbon footprint of global tourism, (Lenzen, et al., 2018), and this paper investigated the increasing dangers in the carbon footprint that is left in the wake of global tourism. Using some significant contributors selected, they modelled the forecast based on the industry and their results indicated a significant increase/ build-up of the emissions from tourism.

Li et al. (2020) forecasted CO2 emissions in China for the years of 2019 to 2030 using four time series forecasting models: autoregressive integrated moving average (ARIMA), traditional grey model GM (1,1), discrete grey model (DGM) and rolling DGM (RDGM), considering the emission intensity data of China from 1990 to 2018 (Li et al., 2020). ARIMA was the best performing forecaster, however mean MAPE value was less than 2% across the four forecasting models. The result indicated a gradual decrease in current greenhouse gas emissions in China. However, the gradient of reduction is lower than the government goal of 60-65%.

However, there has not been much work done to forecast CO2 emission in the USA using time series. Zhao et al. (2018) forecasted CO2 emission for United States using mixed data sampling regression model and back propagation neural network (Zhao et al., 2018). Ikram et al. (2021) used optimized discrete grey model (ODGM), nonhomogeneous discrete grey model (NDGM), and variable speed and adaptive structure grey model (VSSGM) to predict future trend of CO2 emission for the USA and China, the top two countries contributing most to the emission (Ikram et al., 2021).

Moreover, the study of CO2 emission from energy usage in the USA is also very limited. Almost three-quarters of CO2 emission in the USA come from energy usage in different sectors. Therefore, it is imperative to see how things will be shaping soon if fossil fuel remains the primary source of energy in the USA as it is now.

This study used different time series algorithms such as ARIMA, Auto Regressor, Moving Average, and Auto Arima to forecast the CO2 emission from energy usage in different sectors (e.g., buildings, transportation, electricity and heat generation, and industrial processes). Those models perform better in static data; that is trend in the data is constant over time. Therefore, in some cases where the data was non-stationary, transformed to stationary to improve the prediction accuracy of the models. Several metrics such as RMSE, MAPE, MAE were used to compare the models. The model that exhibited least errors and captured more information from the training set was used to forecast the emission for next five and ten years.

The results showed different trends across the forecasted periods of 10 years. The electricity and heat sector showed a drastic increase in the carbon emission predicted while the industries and transport sectors had a slow decline in the carbon emissions. The building sector showed a stable forecast across the forecasts.

# METHODOLOGY

This section will discuss detailed information on the dataset, methods, and model evaluation criteria used in this study.

## Business/ Data Understanding

We broke down the overall concentration into different sections to properly exhaust the topic at hand. In the initial phase of the project, the aim was to explore and furthermore identify the section of the carbon footprint. This was followed by the concept building. The means for acquiring and extracting the data were to identify several relevant sources from which the data might be correlated, and effective relationships drawn between the respectively obtained data. For this project's scope, all inferences and data collected were made for just the United States. This was primarily selected based on the availability of the data and the volume of data to test and compare different models for maximum efficiency. This also offers the opportunity to gain deeper understanding of the data among other sectors and factors alike.

## Dataset Details

CO2 emissions from sectors such as energy usage in residential and commercial buildings, direct industrial, and chemical processes (e.g., production of cement, chemicals, and petrochemicals), transportation (e.g., direct emission from burning fossil fuels), and the generation of electricity and heat were used to train the model and make the forecast. The data is sourced from “ourworldindata.org” website and used for the forecasting of CO2 emissions for next five and ten years based on the 1990-2018 period. Figure 1 shows the breakdown of CO2 emission from four different sectors over the years of 1998 to 2018. Electricity and heat production dominates the CO2 emission, followed by transportation and buildings. Figure 2 shows the per capita breakdown of CO2 emission in 2018 from different sectors including buildings, industries, transportation, and electricity and heat. Similar to the overall CO2 emission as displayed in figure 1, the per capita CO2 emission is also dominated by electricity and heat generation, followed by transportation and buildings. Industry processes are one of the least contributors of CO2 emissions.

Figure 1: United States CO2 emissions by sector (reproduced from the data of ourworldindata.org)

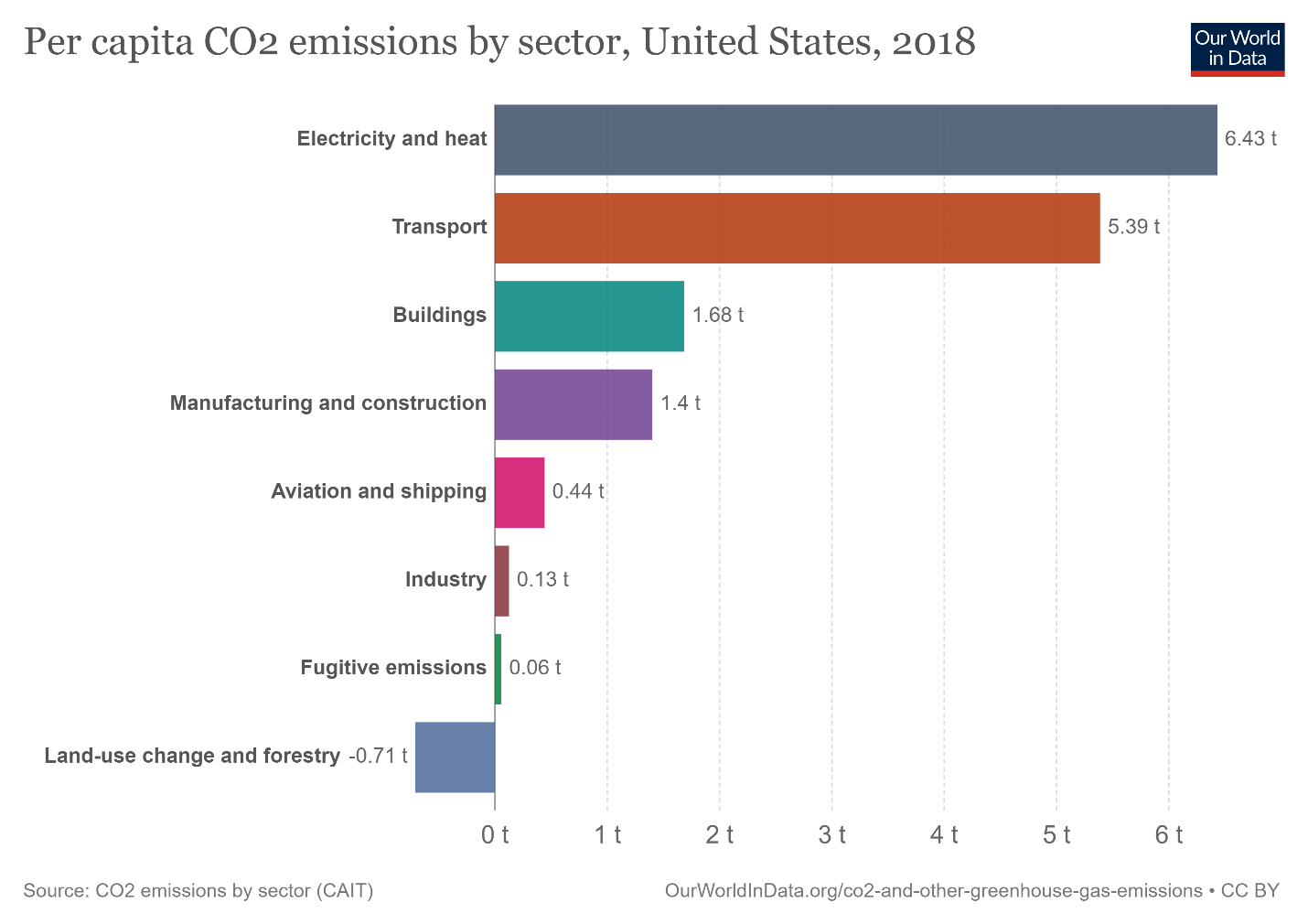


Figure 2: United States per capita CO2 emission, 2018 in tonnes (Courtesy: ourworldindata.org)

Chart, histogram

Description automatically generated

Figure 3: Descriptive Statistics of the buildings, industry, transport, electricity and heat.

## Data Transformation

To prepare the data for the time series modelling the data were first converted to a time series format. The distributions of the data were checked and none of the sectors were showing data normally distributed (Fig. 3). Five different kinds of transformations were used to improve the normality of the data: Log Transformation, Reciprocal Transformation, Square-Root Transformation, Exponential Transformation, and Box-cox Transformation. However, none of the transformation made any conclusive evidence of showing normal distribution. Therefore, in this study untransformed raw data were used for the time series modelling. Figure 4 shows the distribution of CO2 emission from buildings with and without transformations.

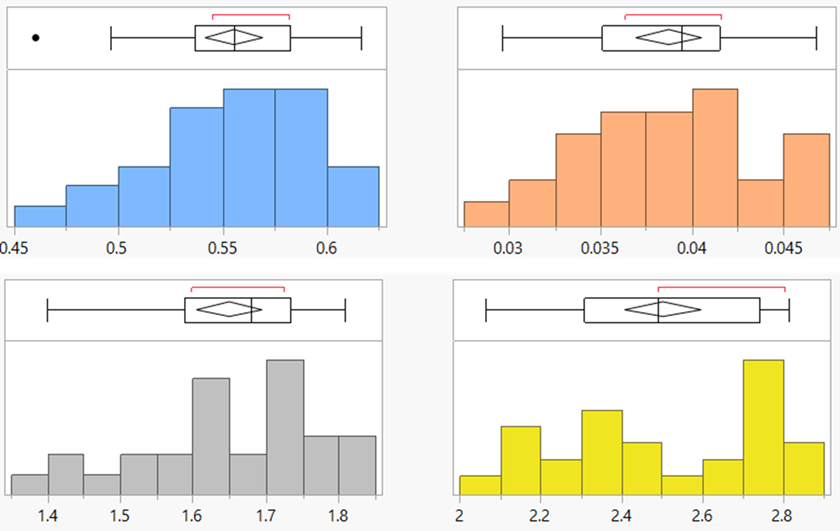


Figure 4: Distribution of CO2 emission from (left to right) buildings, industry, transport, and electricity and heat

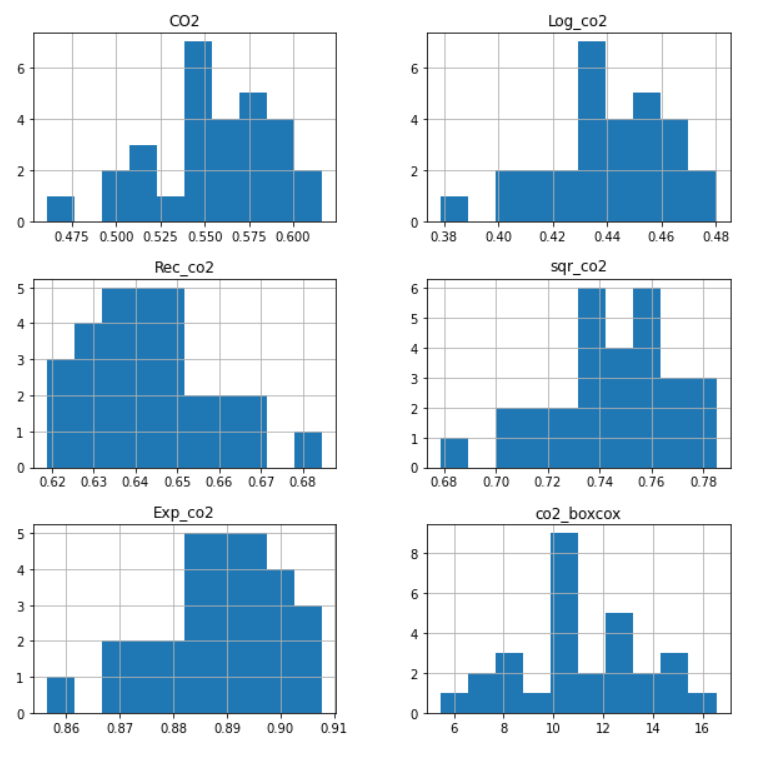


Figure 5: CO2 emission from buildings with and without transformations

## Exploratory Data Analysis

Prior to the modeling, the CO2 emission data from each section were analysed to check whether the data were stationary. Three different methods were used to test the stationarity of the data: Naïve test which divides the data into two section and compares the mean and variance of two sections; Augmented Dickey-Fuller (ADF) test that hypothesizes that the data is not stationary (consists of a unit root) and uses p-statistics to accept or reject the hypothesis (Mushtaq, 2011); Rolling statistics that visualizes how the rolling mean and standard deviations vary with time. Five years were selected as the rolling window. Table 1 shows the results from Naïve test and ADF test. Buildings and Industry both were found as having stationary time series from the Naïve test (the numbers are not varying drastically) and ADF hypothesis test (p-value is not significant). However, based on our subjective assessment of stationarity (Fig. 6), it appears that the mean and standard deviation of CO2 emissions from industries clearly shift with time. This demonstrates that the series follows a pattern. So, the time series is not stationary. However, for buildings the mean and standard deviation was flat, indicating stationarity of data.

Table 1: Results from the objective test of the stationarity of data

|  |  |  |
| --- | --- | --- |
| Sectors | Naïve Test Result | ADF test result |
| Buildings | mean1=0.579223, mean2=0.532992.  variance1=0.000427, variance2=0.001010 | p-value 0.163488. Failed to accept null hypothesis. Data is stationary |
| Industries | mean1=0.038041, mean2=0.039406.  variance1=0.000012, variance2=0.000028 | p-value 0.145290. Failed to accept null hypothesis. Data is stationary |
| Electricity and heat | mean1=2.539316, mean2=2.467653  variance1=0.055894, variance2=0.060989 | p-value 0.964459. Failed to accept null hypothesis. Data is stationary |
| Transport | mean1=1.586307, mean2=1.708987 variance1=0.015405, variance2=0.004880 | p-value 0.058953. Failed to accept null hypothesis. Data is stationary |

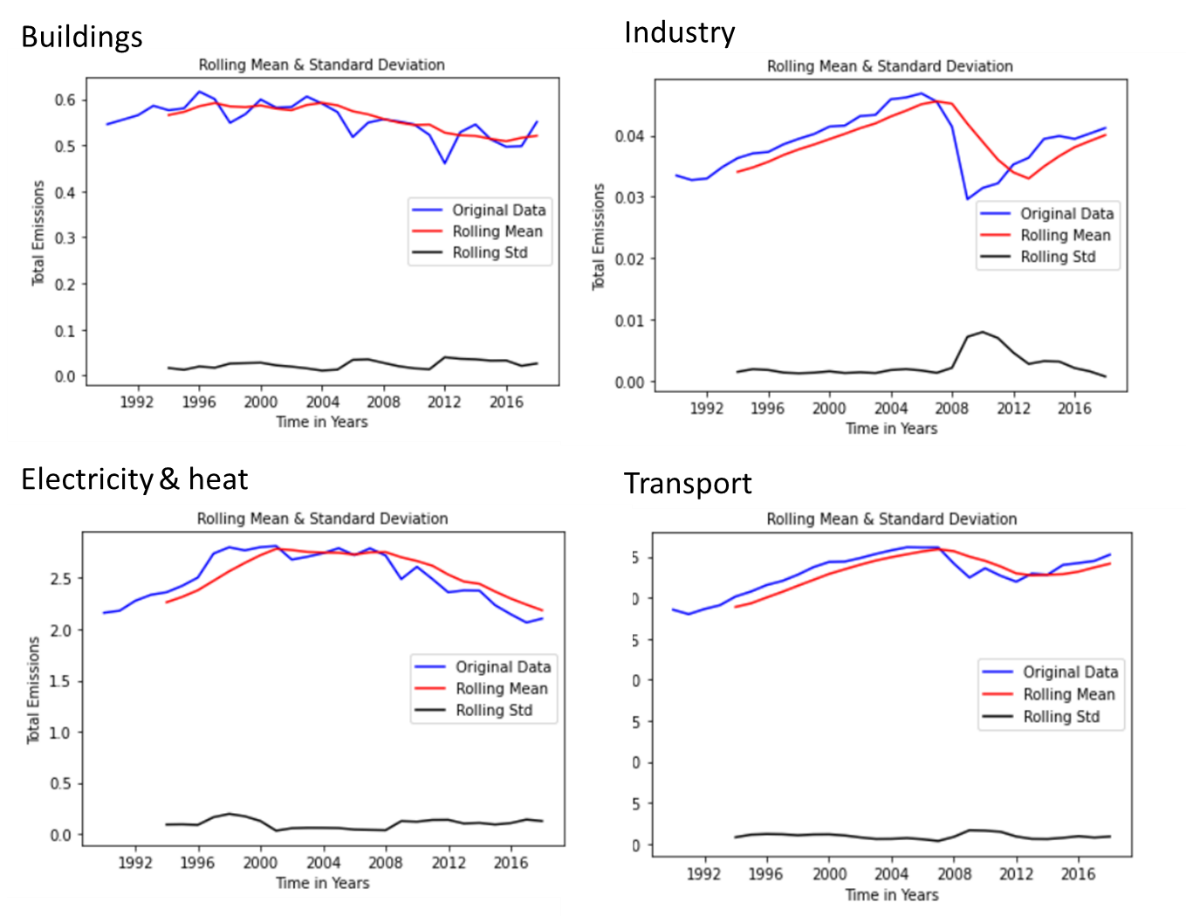


Figure 6: Subjective assessment of stationarity based on rolling statistics

## Transforming data to stationary

The dataset that exhibited non stationarity were transformed to stationary using two methods: Moving average smoothing and differencing.

### Moving average smoothing

Moving average smoothens time series data by eliminating non-stationarity and fine-grained volatility between time increments. Smoothing removes noise and exposes the signal of the underlying causal processes. Moving average is the simplest and extensively used smoothing technique in time series analysis and forecasting. A moving average is calculated by creating a new series with values equal to the average of the original time series' raw data. A window size needs to be defined to specify how many raw observations were used to calculate the moving average value. The window will slide along the time series to calculate the new series' average values (Wei, 2006). In this study a window size of five years is selected. Figure 7 is showing the time series of the CO2 emission from three different sources after applying the moving average smoothing technique. The blue line shows the raw data, and the orange line is showing the averaged data for five years. Even though the moving average smoothed the time series, the ups and downs in the plot exhibit that the data still has some non-stationarity in it.

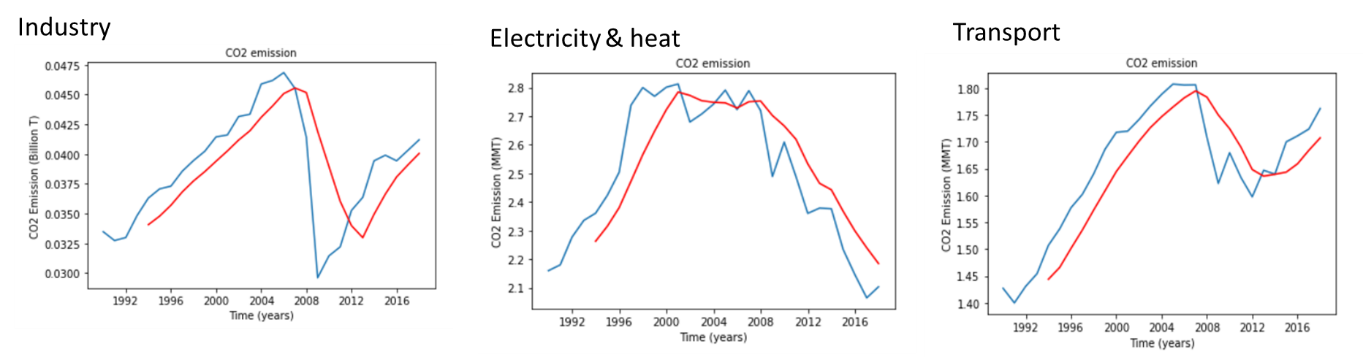
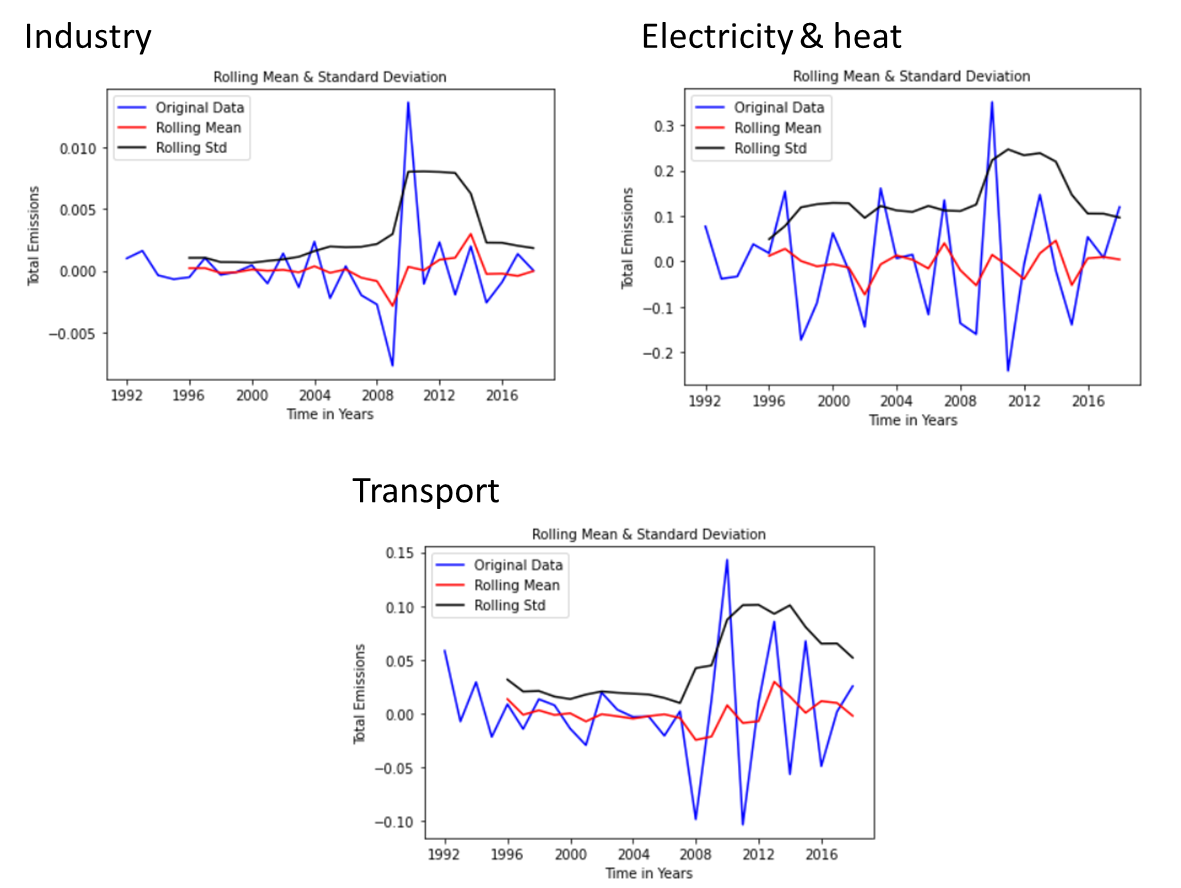


Figure 7: The time series of CO2 emission after applying moving average.

### Differencing

Differencing is another frequent approach of dealing with the non-stationarity of the data. It is the transformation of the original time series to a new time series where instead of raw observation, the differences between the original observation at one instant and the preceding observation at another instant are used. The procedure can be applied more than once and known as first order differences, second order differences, etc. In this study we used both first order and second order differencing to remove both the trend and seasonality from the data. The second order differencing improved the stationarity better than the first order one, hence used in the further analysis. Figure 8 shows the rolling statistics of the data after applying second order differencing. Looking at all three charts with the red lines representing the respective rolling means, it is noticeable that this rolling mean variable, although not constant, does not deviate much from a straight line. This suggests that the differenced data has no effect of time on the flow through the years.



*Figure 8: The rolling statistics of the data after applying second order differencing*

## Modeling

For the forecasting predictions, the models that were used to obtain the final model results were the Autoregressive Integrated Moving Average (ARIMA), Auto Regressive (AR), Moving Averages (MA) and the Auto ARIMA models. The ARIMA model is a statistical analysis model that uses the effect of the variables as related to a time series to better understand the given data sets or to predict the future trends of the already present relationships. These statistical models would be termed autoregressive if the predictions are done based on the effects of the past values. Moving averages are also calculated to smoothen out the effects of the outliers on the predictions. The combination of these in the ARIMA model can consider the trends, cycles, and other static and non-static types of data for the forecast. The previous steps, being identifying the trends within the lags from the ACF and the PACF, and the testing for stationarity is an important step to properly understanding or forecasting these relationships from the already obtained data.

The next objective of the modelling process is data partitioning and hyperparameter tuning. A split of 70 30 percent was used for the training and the validation of the datasets respectively. This was to work towards obtaining the optimal values for the p, d and q values which were going to be used in our model building.

### Optimal parameters selection for modelling

Autocorrelation (ACF) and partial autocorrelation (PACF) charts were plotted to find out if the time series data has any correlation with itself. The ACF measures the correlation of the observation at each point with a lagged version of itself. The PACF does the same thing but controls the effect of covariates. Figure 9 shows the ACF and PACF plots. Only the first 2 or 3 lag effects extend beyond the 95% confidence interval, which means the data doesn’t have significant level of autocorrelation or partial correlation which is also an indication of stationarity. Since the data do not exhibit correlations, we identified the optimal parameters by hyper-parameter tuning.

Graphical user interface, application, table, Excel

Description automatically generated

Figure 9: ACF and PACF plots for the respective sectors

A series of ARIMA models with different values for levels (ranged from 0 to 5), terms (ranged from 0 to 3), and moving average (0 to 3) were fitted into the data and based on least RMSE value, the best performing model was selected. For the auto regressor model a lag of 6 was used for the CO2 emission data from building, electricity and heat, and transportation. For industries, lag of 2 was found to have the least RMSE value. Running the data through the different models allowed the comparison to the forecasting prediction accuracy. This was hierarchically ordered based on the Mean Absolute Percent Error (MAPE) and the Mean Absolute Error (MAE).

The breakdown of the selection of these models helps using the accuracy determiners is the basis for the forecasting analysis and prediction. The output for the forecasted values were saved and added to the data dictionaries as present in the collected data. This will help to visualize the future emission trends.

# RESULTS

The results derived from the models that were applied in the methodology of our paper were as an outcome of running the respective models through different time series statistical models and then assessing the accuracy of the results. As previously mentioned, the criteria selected to be the deciding factor for selecting the models were the Mean Absolute Percent Error and the Mean Absolute Error inferential quantities. Under each of the sectors of the CO2 emissions data, each variable was run through the ARIMA, AR, MA, and the Auto ARIMA models. Shown below are the results of the model accuracy and the associated interpretations:

Buildings

Graphical user interface, text, application

Description automatically generated

Figure 10: Model selection for CO2 emissions from buildings based on the MAPE values

The final model and the forecast/prediction was then built using the best model from this accuracy list. From the figure, even though the Auto Regressor was not the best model out of the 4, we intuitively went with that model because the forecasted vs Actual values graphs shown in Figure 14 captured more information when compared to other models.

Industries

Graphical user interface, application

Description automatically generated

Figure 11: Model selection for CO2 emissions from industries based on the MAPE values

For the variable industries, the iteration through the multiple ARIMA methods produced the best performing model for forecast and prediction. The ARIMA model and the Moving Average model seemed to have the same accuracy values. This was then determined to be an ARIMA model with no Auto Regressive component which makes the Moving Average the best model due to the principle of parsimony. However, for future ease of modelling we decided to select the ARIMA model since the input of the parameters were the same, but the build function would not need to be changed in the future, just altered.

Electricity and Heat

Graphical user interface, table

Description automatically generated with medium confidence

Figure 12: Model selection for CO2 emissions from electricity and heat based on the MAPE values

The Moving Average model had the best accuracy for the modelling the electricity and heat variable.

Transport

Table

Description automatically generated

Figure 13: Model selection for CO2 emissions from transport based on the MAPE values

Transport was best defined by the Auto ARIMA model. For this variable analysis, the Auto ARIMA and the ARIMA had very similar best predicted model, however the hyperparameter selection of the Auto ARIMA was more fine-tuned in line with obtaining the best model based on the overall prediction performance.

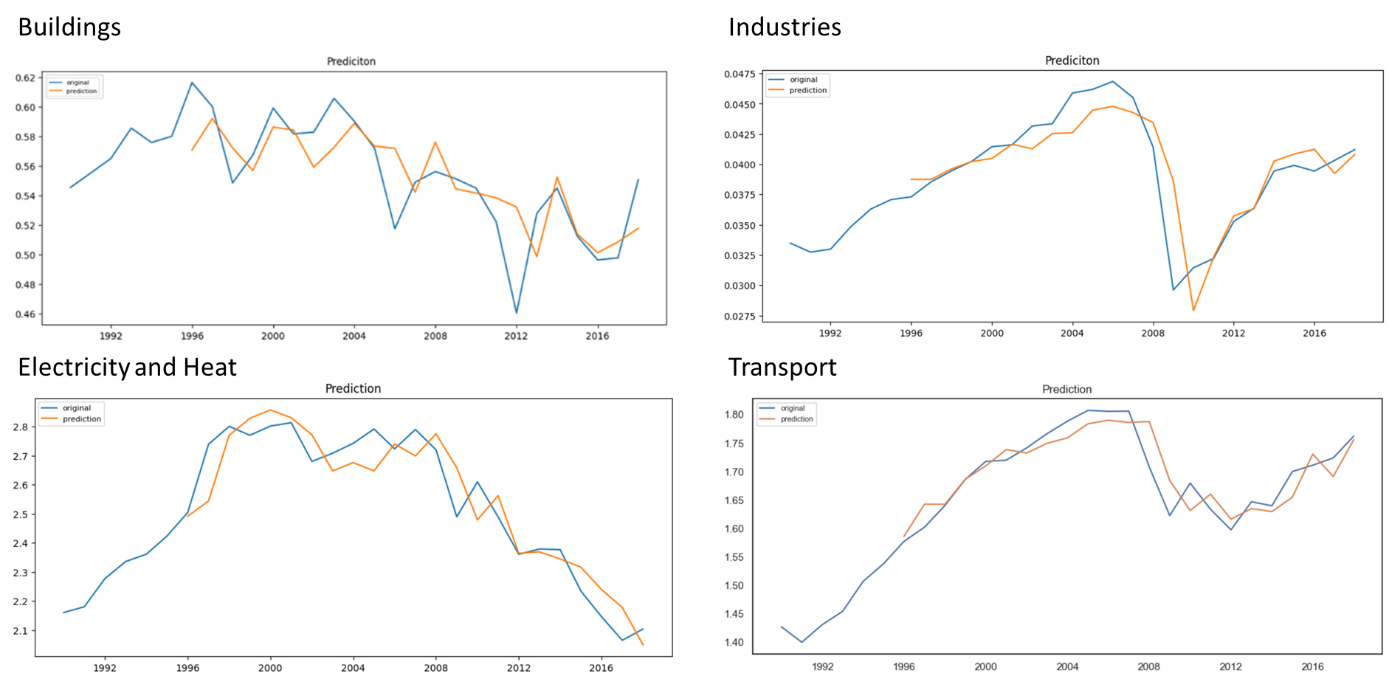


Figure 14: Cross validation of selected model

The information Figure 14 is illustrating is the validation of the model with the actual values from the datasets. The matching that is seen here is relatively consistent with the actual data that was used to build and train the model. This cross validation asserts the accuracy of the models chosen for the prediction and the consequently the forecast.

Even though the MAPE value from Auto Regression (AR) was not the least among the models for building, we chose AR as the best model because it was able to capture the variation in the validation data better than the other models and the MAPE value was also not significantly higher than that from other models. However, unlike industries, for buildings, electricity and heat, and transport data set, even though AR model continued to capture more information about the validation in comparison, it has MAPE significantly higher than the other models. Therefore, we have chosen the ARIMA model as the best performing model for those datasets solely based on the MAPE values.

The forecasting results for all the selected models and their predictions for the next 10 years are shown in the figure below.

Graphical user interface, chart, line chart

Description automatically generated

Figure 15: Forecasted values for CO2 emissions per sector.

The forecasts generated indicated the trends observed in 10 years. The forecast was selected for 10 years because of the increasing forecasting error and uncertainty of the models.

# CONCLUSION

The research study aimed to identify the trends of the forecasted CO2 emissions effects with respect to different sectors considered. The best models that were chosen gave the optimum predicted forecasts in comparison to the other used models with the highest accuracies. The results show that the building sector in the next 10 years would not be actively change. Considering the industry and the transport sectors, a steady downward trend was identified and for the sectors for the electricity and heat, the carbon emissions were predicted to rise very hugely.

Moving forward, a lot more focus must be directed towards alternative solutions in the electricity and heat sectors. Energy efficiency, fuel switching, combined heat and power, renewable energy, and more efficient use and recycling of materials are just a few of the ways the different sectors may incorporate to minimize the emissions of CO2. This information would be crucial for the monitoring and control of carbon emission to accepted standards.

# APPENDIX

TABLE OF FIGURES

[Figure 1: United States CO2 emissions by sector (reproduced from the data of ourworldindata.org) 6](#_Toc100610625)

[Figure 2: United States per capita CO2 emission, 2018 in tonnes (Courtesy: ourworldindata.org) 7](#_Toc100610626)

[Figure 3: Descriptive Statistics of the buildings, industry, transport, electricity and heat. 7](#_Toc100610627)

[Figure 4: Distribution of CO2 emission from (left to right) buildings, industry, transport, and electricity and heat 8](#_Toc100610628)

[Figure 5: CO2 emission from buildings with and without transformations 9](#_Toc100610629)

[Figure 6: Subjective assessment of stationarity based on rolling statistics 10](#_Toc100610630)

[Figure 7: The time series of CO2 emission after applying moving average. 11](#_Toc100610631)

[Figure 8: The rolling statistics of the data after applying second order differencing 12](#_Toc100610632)

[Figure 9: ACF and PACF plots for the respective sectors 13](#_Toc100610633)

[Figure 10: Model selection for CO2 emissions from buildings based on the MAPE values 14](#_Toc100610634)

[Figure 11: Model selection for CO2 emissions from industries based on the MAPE values 14](#_Toc100610635)

[Figure 12: Model selection for CO2 emissions from electricity and heat based on the MAPE values 15](#_Toc100610636)

[Figure 13: Model selection for CO2 emissions from transport based on the MAPE values 15](#_Toc100610637)

[Figure 14: Cross validation of selected model 15](#_Toc100610638)

[Figure 15: Forecasted values for CO2 emissions per sector. 16](#_Toc100610639)

#### Scripts Used

The coding scripts used for the entire research analysis alongside the experimented ideas for replication of the information presented are available upon request.

# REFERENCES

Akyol, M., & Uçar, E. (2021). Carbon footprint forecasting using time series data mining methods: the case of Turkey. *Environmental Science and Pollution Research*, *28*(29), 38552-38562.

Dunn, J. C. (2010). *Comparing Subnational Ecological Footprints for the United States*.

Hertwich, E. G., Peters, Glen P. (2009). Carbon Footprint of Nations: A Global, Trade-Linked Analysis. *Environ. Sci. Technol.*, 6414-6420.

Hosseini, S. M., Saifoddin, A., Shirmohammadi, R., & Aslani, A. (2019). Forecasting of CO2 emissions in Iran based on time series and regression analysis. *Energy Reports*, *5*, 619-631.

Ikram, M., Sroufe, R., Zhang, Q., & Ferasso, M. (2021). Assessment and prediction of environmental sustainability: novel grey models comparative analysis of China vs. the USA. *Environmental Science and Pollution Research*, *28*(14), 17891-17912.

Lenzen, M., Sun, Y.-Y., Faturay, F., Ting, Y.-P., Geschke, A., & Malik, A. (2018). The global footprint of tourism. Nature Climate Change.

Li, Y., Wei, Y., & Dong, Z. (2020). Will China achieve its ambitious goal?—Forecasting the CO2 emission intensity of China towards 2030. *Energies*, *13*(11), 2924.

Mushtaq, R. (2011). Augmented dickey fuller test.

Ritchie, H. (2019). *Per capita CO2 emissions*. <https://ourworldindata.org/per-capita-co2>

Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Wiedmann, T., & Minx, J. C. (2008). *A Definition of CArbon Footprint*.

Zhao, X., Han, M., Ding, L., & Calin, A. C. (2018). Forecasting carbon dioxide emissions based on a hybrid of mixed data sampling regression model and back propagation neural network in the USA. *Environmental Science and Pollution Research*, *25*(3), 2899-2910.