Design and Analisys of ARMA, ARIMA, SARIMAX Models on Carbon Emissions Dataset

Authors : Prerana Nimhan, Vaishnavi Tale & Aditi Kandarkar

Post Graduate Diploma in Data Science & Artificial Intelligence from COEP

nimhanpk.pgddsai@coep.ac.in, talevr.pgddsai@coep.ac.in, kandarkarad.pgddsai@coep.ac.in

Motivation

- Global warming is become the growing concern in recent time . The root cause of problem directs towards the carbon emission.
- The idea behind our project is to run a prediction , analysis and forecasting over dataset related to carbon emission.
- The dataset used is a public dataset of monthly carbon dioxide emissions from electricity generation available at the Energy Information Administration and Jason McNeill. The dataset includes CO2 emissions from each energy resource starting January 1973 to 2014.
- The predictive analysis of carbon emission using different time series models such as ARMA ,ARIMA, SARIMAX.

Objective

- The objective of this research is to propose the appropriate Time Series forecasting models for the analysis of carbon dioxide emissions from electricity generation available at the Energy Information Administration and Jason McNeill.

Dataset

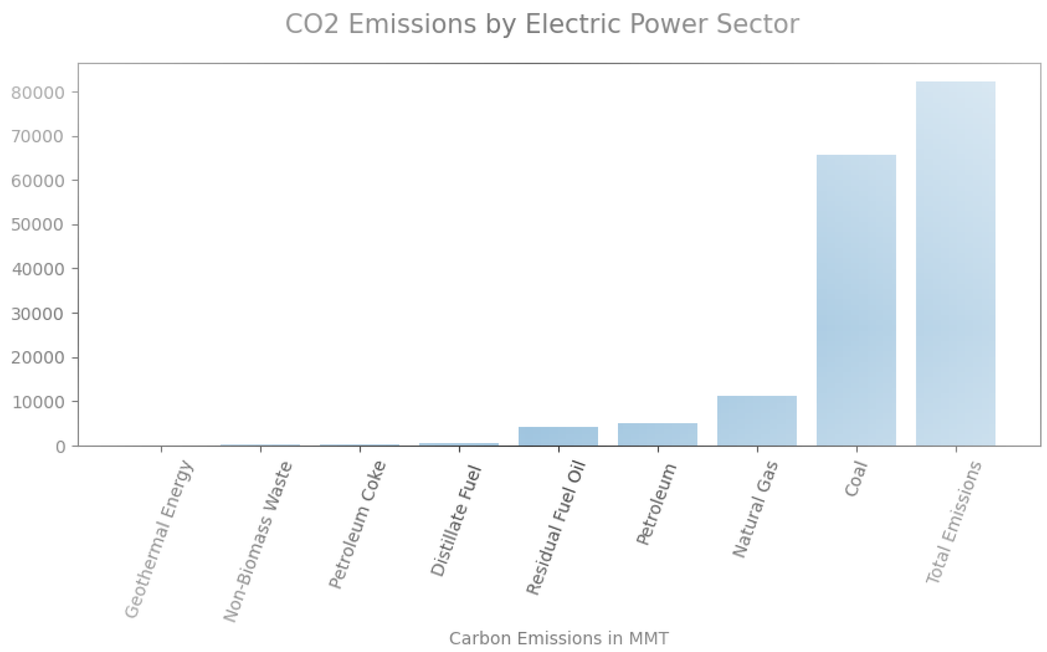
- The dataset includes CO2 emissions from each energy resource starting from January 1973 to 2014.
- Mainly it has 6 columns
  - 2 - integer data type
  - 4 - objective data type
  - with total 5096 observations
- The given dataset does not represent the time series,we first need to convert it into time series dataset using:
  - `parse_dates`: This is a key to identify the date time column.
  - `Index_col`: This is a key that forces pandas to use the date time column as index.
  - `date_parser`: Converts an input string into datetime variable
- Dataset contain the information about 8 energy source carbon emission. The carbon emission across the different sectors.

Baseline Approach

- Retrieve csv time series dataset, visualize time series datasets, transform dateset into time series data.
- Test if the time series is stationary or not using graphical representation and Dickey-Fuller test statistics method
- If time series is not stationary transform it to stationary using methods such as differencing,etc.
- For performing evaluation plot the ACF (Auto-correlation Function) and PACF (Partial Auto-correlation Function) plot
- To analyse data build ARMA,ARIMA,SARIMAX model on dataset.
- Model checking using RMSE-Root Mean Squared Error and AIC.

Feature extraction

- Set the date column(YYYYMM) as the `index_col`
- Visualize the factors or components that contribute to the total carbon emissions.

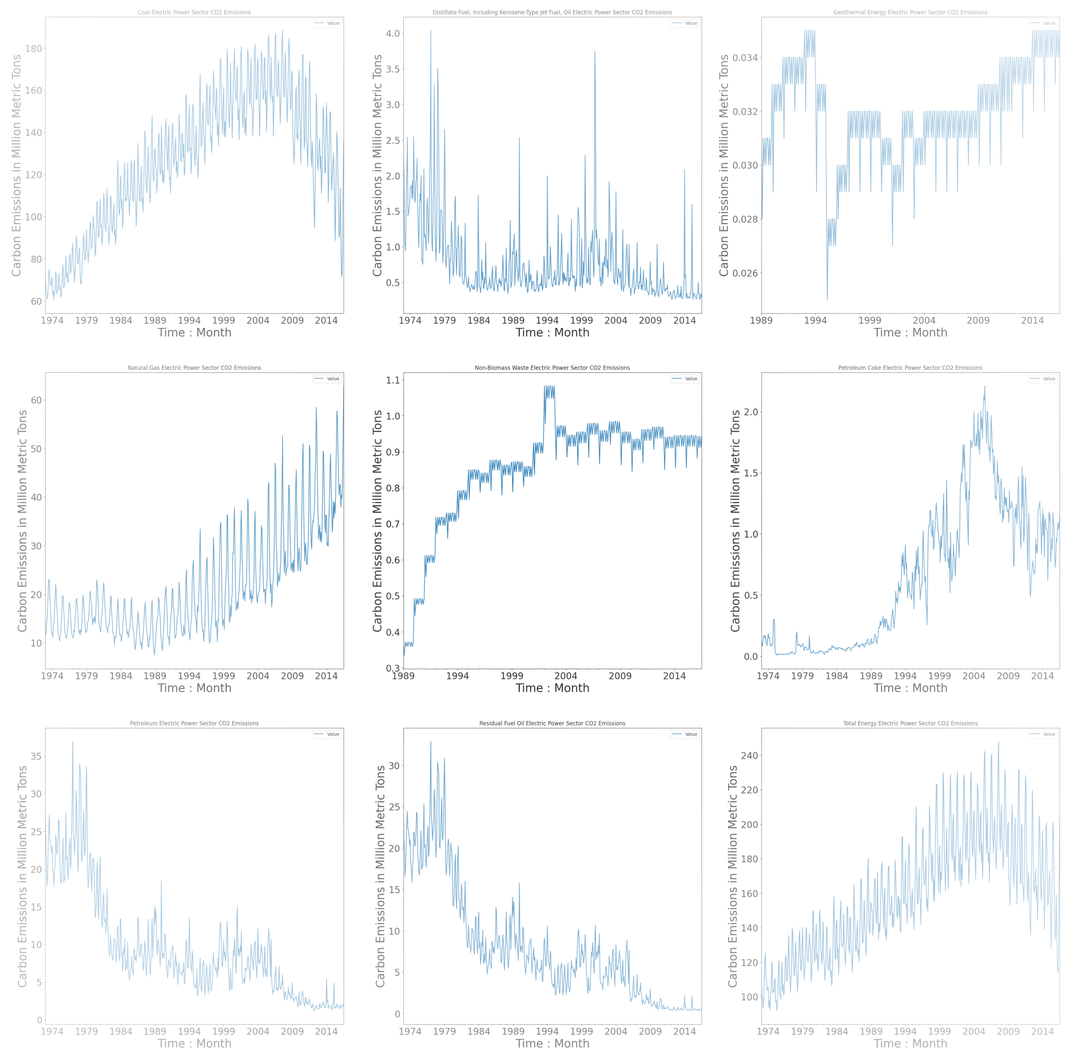


From above bar chart we can see that contribution of Coal is significant to total carbon emission followed by Natural gas. So we will make forecasting for Coal Electric Power Sector CO2 emission.

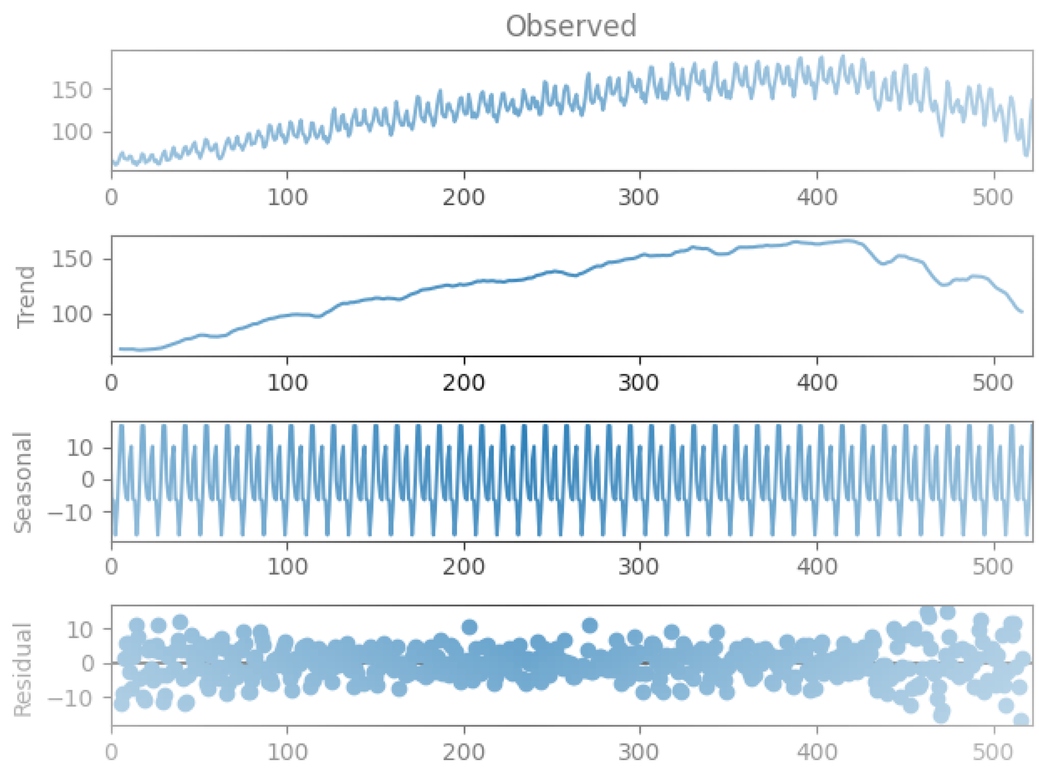
Experiments

- Stationary check:-
  - AUGMENTED-DICKEY-FULLER(ADF) Test:
    1. `Ho`- Fail to reject null hypothesis, therefore data is stationary
    2. `H1`- Reject the null hypothesis, therefore data is not stationary.
    3. If 'p-value' is less than threshold value is '0.05' we say fail to reject the null hypothesis and hence data is stationary.
    4. If 'p-value' is more than threshold value '0.05' we say reject the null hypothesis & accept alternative hypothesis hence data is not stationary.

- The CO2 emissions, classified by Electricity Generation sectors - Geothermal Energy, Non-Biomass Waste, Petroleum Coke, Distillate Fuel,Residual Fuel Oil , Petroleum, Natural Gas, Coal , are used to trend the 40 years of time series recording to forecast CO2 emissions.



- **Forecasting Only for Coal Electric Power Sector CO2 Emissions from the various other generation sectors**
- **Seasonal Decomposition Analysis of Coal Emissions: Trend, Seasonality, and Residuals**

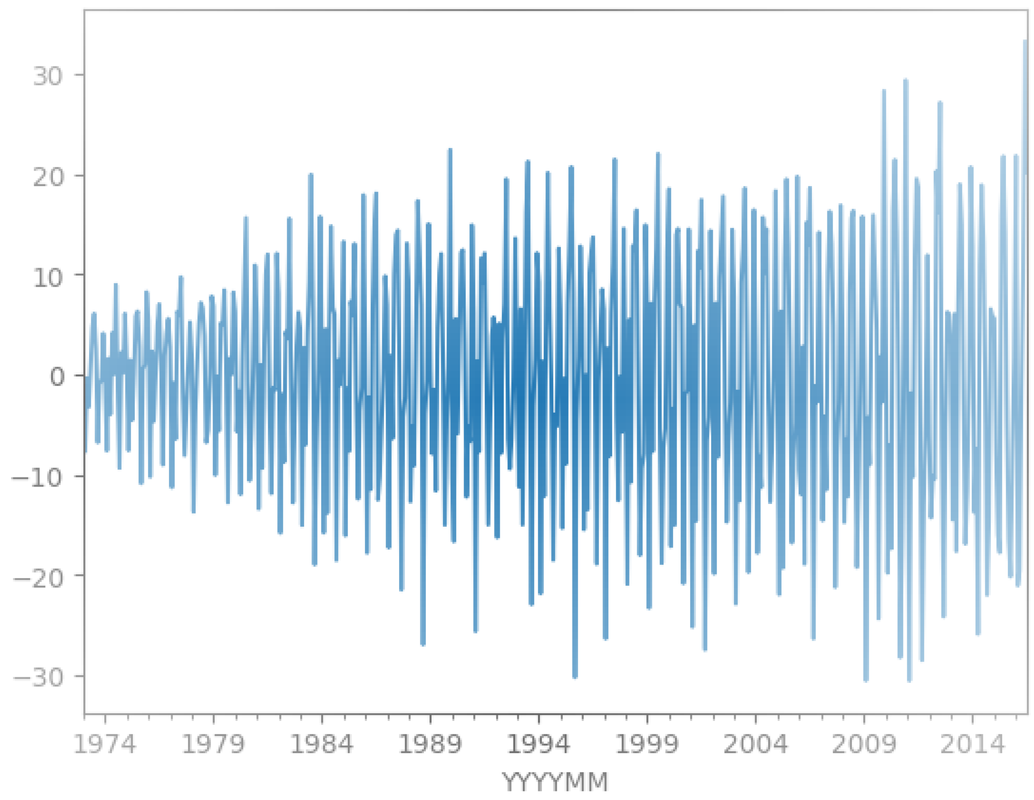


  - Decomposition is primarily used for time series analysis, and as an analysis tool it can be used to inform forecasting models on your problem.
  - Used here is A additive model which suggests that the components are added together as follows:

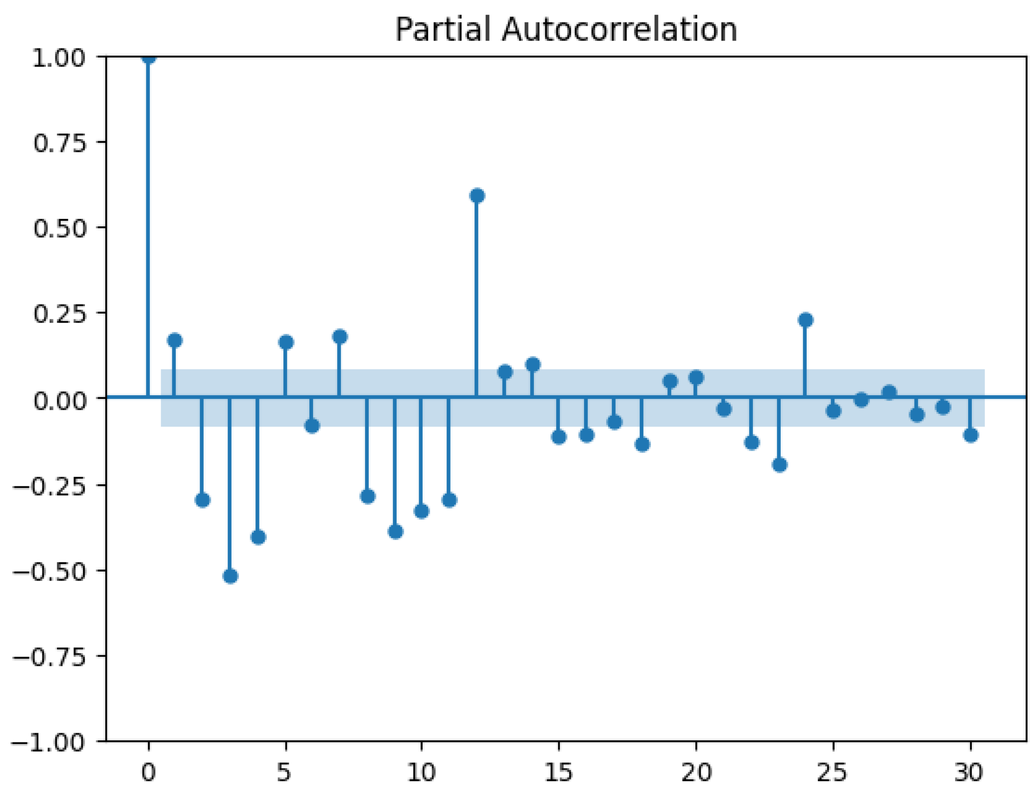
$$y(t) = Level + Trend + Seasonality + Noise \tag{1}$$

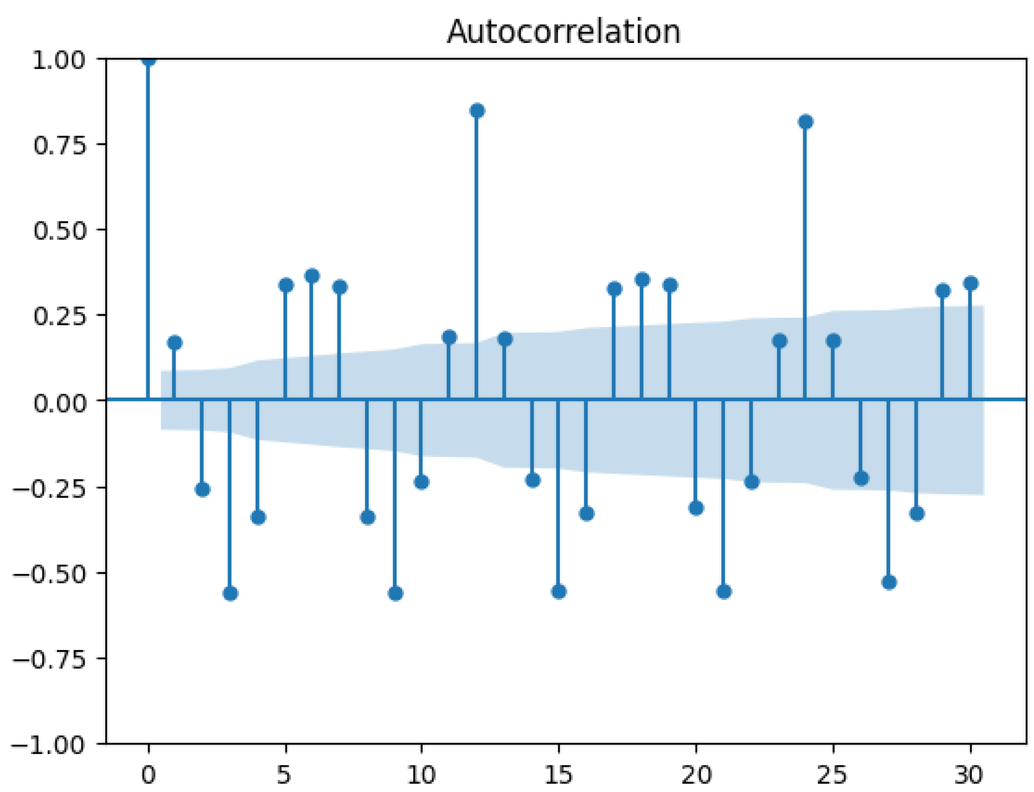
  - `Level`: The average value in the series.
  - `Trends`: The trend component describes how the variable – Coal Emissions in this case – changes over long periods of time. We see that the emissions of coal have substantially increased during the time period.
  - `Seasons`: The seasonal component showcases each years wavelike changes in emissions patterns.We here see monthly or quarterly patterns for emissions.
  - `Residual`: It represents Irregular components which appear due to unexpected events, like cataclysms, or are simply representative of noise in the data.

- **Visualizing the Stationary Nature of the Data through Plots**
- Stationary data refers to the time series data that mean and variance do not vary across time
- Here our Data is made Stationary through ADF Test
  - Original data:
    - `p-value` :0.36683253565668955
  - After 1st Order Differencing
    - `p-value` :5.409945916192542e-06
- Hence we use order of differencing d=1 for building models



- **Pacf and acf plots to determine q and p values respectively**



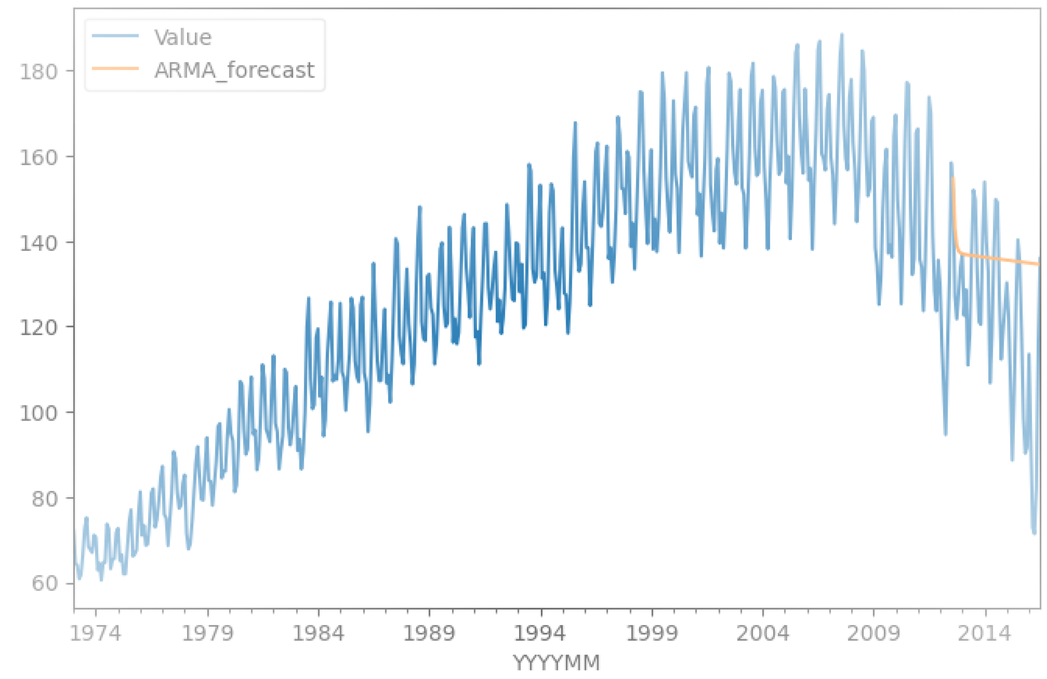


- Autocorrelation function (ACF) and Partial Autocorrelation Function (PACF)are important functions in analyzing a time series.
- They generally produce plots that are useful in finding the values p, q for Autoregressive (AR) and Moving Average (MA),ARMA,ARIMA,SARIMAX,etc. models.
- An ACF measures and plots the average correlation between data points in time series and previous values of the series measured for different lag lengths.
- A PACF is similar to an ACF except that each partial correlation controls for any correlation between observations of a shorter lag length.

- **Model Building**

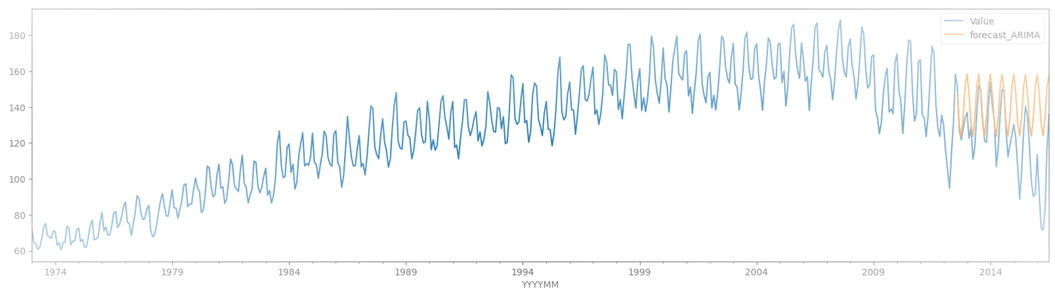
- The ARMA, ARIMA & SARIMAX forecast plot covers the time period from January 1973 to December 2014.
- It includes the historical data up to August 2012 and provides forecasted values for the subsequent months. The plot displays both the observed values and the forecasted values

- **(Auto Regressive Moving Average)ARMA Model Plot**



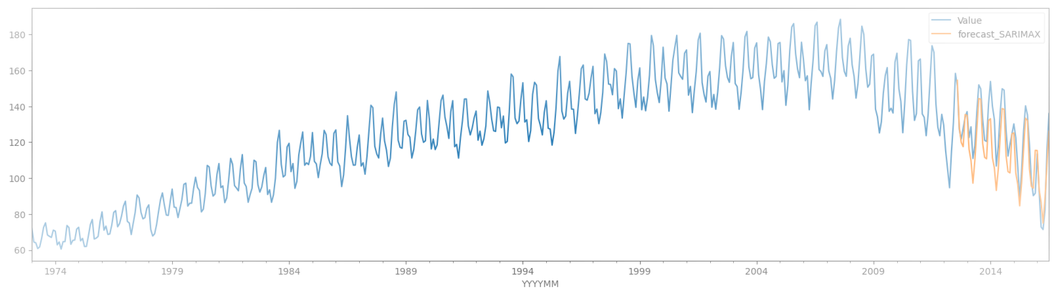
- The ARMA forecast plot provides insights into the future behavior of the time series using an Autoregressive Moving Average (ARMA) model.
- The historical data in the plot shows a clear upward trend from 1996 to 2009, followed by Notable seasonal patterns are observed with regular peaks during specific months.
- The forecasted values suggest a slight decline from 2012 to 2013 & series is further constant from January 2013 onwards

- **(Auto Regressive Integrated Moving Average)ARIMA Model Plot**



- ARIMA Model combines the concepts of autoregressive (AR), differencing (I), and moving average (MA) models to capture both the linear dependencies within the series and the impact of past observations on future values.
- The ARIMA model is typically denoted as ARIMA(p, d, q), where:
  p represents the order of the autoregressive component.
  d represents the degree of differencing required to achieve stationarity.
  q represents the order of the moving average component.
- Here the model captures the overall trend and seasonal patterns reasonably well.

- **(Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) SARIMAX Model Plot**



- The SARIMAX model (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) is an extension of the ARIMA model that incorporates both seasonal and exogenous factors in the analysis of time series data. It is particularly useful for capturing seasonal patterns and incorporating external variables that may influence the time series.
- The SARIMAX model is typically denoted as SARIMAX(p, d, q)(P, D, Q, s), where:
  (p, d, q) represents the non-seasonal order of the AR, I, and MA components.
  (P, D, Q, s) represents the seasonal order of the AR, I, MA, and the length of the seasonal period
- Overall, our model performs well in capturing the historical patterns and providing meaningful forecasts.

Results

- Root Mean Squared(RMSE) Values for different Models

ARMA	ARIMA	SARIMAX
24.464103	25.629771	11.719615

Conclusion

- Built ARMA(2,0,2) model, ARIMA(2,1,2) model & SARIMAX(2,1,2,12) model on the time series data
- The studies show that the SARIMAX(2,1,2,12) model is the optimal model to forecast the total amount of CO2 emissions
- Forecasted values through graphs for each model prove to be essential as they help to understand the future.
- Thus, our forecast using the models can be useful to researchers, environmentalists and others working towards the cause of saving the environment.

References

- https://www.kaggle.com/datasets/txtrouble/carbon-emissions
- https://towardsdatascience.com/time-series-from-scratch-moving-averages-ma-theory-and-implementation-a01b97b60a18
- https://pypi.org/project/statsmodels/ For statsmodel library information.