Capstone Project
Carbon Emissions Forecasting

Final Report MIT ADSP

Sylvia H. Florez

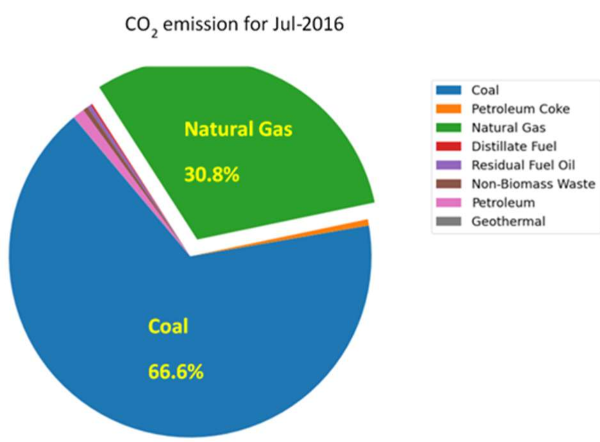
Sylvia.Florez@wdc.com

Executive Summary

Control and reduction of Carbon Dioxide (CO_2) emissions, together with other greenhouse gasses is known to be a vital to reduce the effects of global warming. The urgency is high as the effects of climate change are showing severe consequences such as water shortage, increased wildfires, intense storms and wildlife habitat destruction among many others. In the US, a large contribution of CO_2 emissions comes from the Electricity Generation Sector. Within this sector, coal and natural gas fuel sources produce the largest amounts of CO_2 emissions. In this work we focus on building time-series forecasting models that can accurately predict CO_2 emissions for the natural gas case, from the Electricity Sector. After evaluation of several models, two models, Seasonal Arima and FB Prophet were selected based on their capability of handling data with high seasonality, as was observed in the data. For the case of Seasonal ARIMA, automated optimization (autoarima function) was used to identify the best hyperparameters for the model. Both optimized Seasonal Arima and FB Prophet models were then studied in depth and compared in accuracy using "rolling window" cross-validation technique. The metrics for model evaluation were rmse and mape. Both models showed very good forecasting performance but the Seasonal Arima model with optimized parameters. SARIMA(1,1,1,)x(1,0,1,12) was found to produce most accurate predictions. Forecasted CO_2 emissions from this model can be used as a benchmark for future policy evaluation by comparing actual future data to the forecast(based on historical conditions without the policy). The SARIMAX model has the capability of adding exogenous variables to the model. In this problem, for instance, a separate parallel time series relevant to policy can be added to the model. This would allow forecast of CO_2 emissions with and without the policy in place.

Data Insight

U.S. Energy Information Administration (EIA) time series data for Electric Sector CO_2 emissions from various fuel sources was provided for the time period between 1973-2016. As can be seen in fig.1 below, coal and natural gas show highest contributions to CO_2 emissions. In the time series plot, fig.2 we see that the *coal* CO_2 emissions over time show a steep increase from 1973 to 2007, followed by a reduction between 2008 and 2016. For natural gas, the CO_2 emissions show a low and flat contribution until the 1990's, time from which a steady increase in the trend is observed. This change was likely induced by technology or policy change. The tradeoff between coal and natural gas sources is desirable given the higher efficiency of natural gas and may have influenced the opposing trend observed in the data. Alternative energy technology and policies are likely driving down the CO_2 emissions from coal.



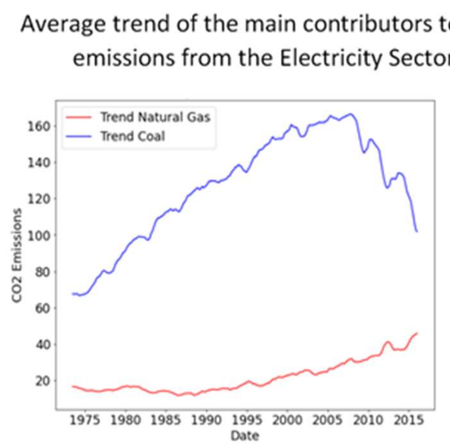
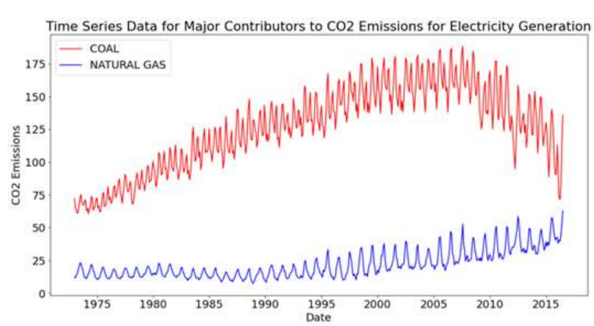


Fig. 1

Fig. 2

Strong seasonality observed in the cases CO_2 emissions from both coal and natural gas as can be seen in fig.3a. For the case of CO_2 emissions from natural gas source, fig. 3b shows 12-month seasonality with peaks in the hotter months of July and Aug.



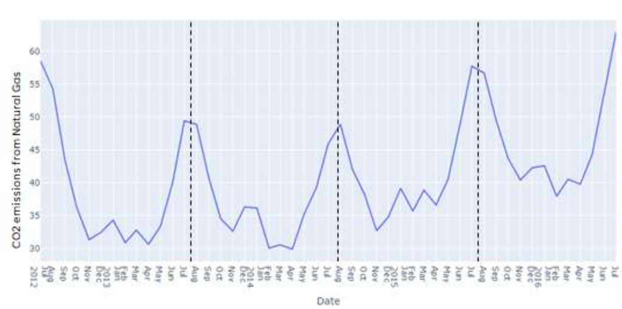


Fig. 3 (a,b)

Problem and solution summary

The steps for solution were

1. Optimize hyperparameters for SARIMA model using automated auto-arima.
2. Build the optimized SARIMA Model
3. Build the FB Prophet Model
4. Use a moving" window based" cross- validation technique to determine which has better performance i.e. lowest average prediction error (see below).
5. Select model with best performance (lowest error), using average from cross-validation horizon time windows.
6. Forecast natural gas CO_2 emissions into the future.

Model 1 : SARIMA (1,1,1)x(1,0,1,12)

SARIMA : Seasonal autoregressive integrated moving average model with 12 month seasonality.

Model parameters, obtained through automated optimization procedure auto-arima.

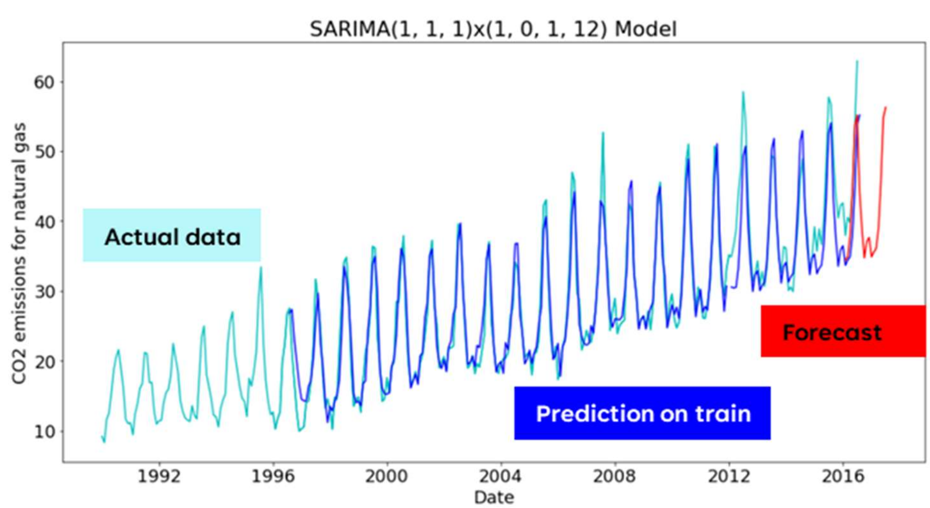


Fig.4

Model 2 : FB Prophet with additive seasonality

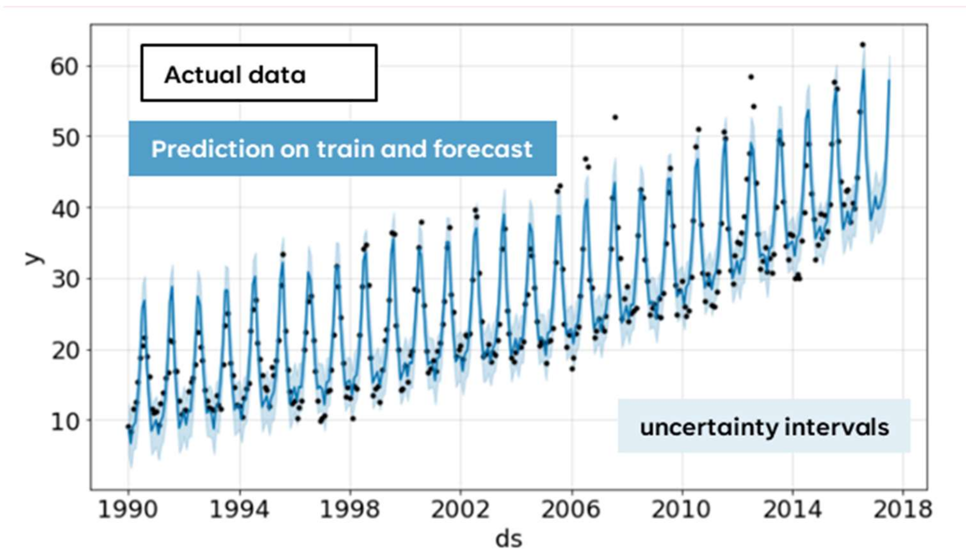


Fig. 5

Comparison of the residuals over a time range with existing data is shown below in figure 6 and suggests good predictability for both models.

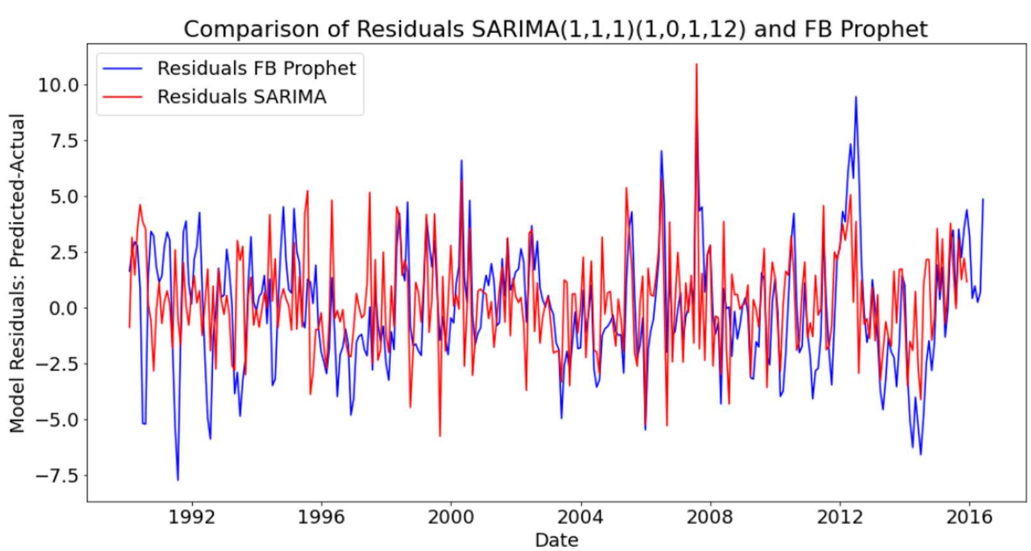


Fig. 6

Cross-Validation with timesplit n= 5

Validation over unseen data in various time windows was performed through cross-validation for both models.

In cross-validation technique(see figure 7) we select an early subset of data to train the model and then evaluate the error between model forecast into the test period and the actual data in the test period. The test window or “horizon” is then moved forward in time and the model is retrained with a larger data set and the error evaluated. This process repeats for all horizons and the final error is the average of the errors corresponding to the different time frames.   Figure 7b shows the time windows used in this particular problem.

Fig. 7(a,b)

Cross-Validation Results SARIMA and PROPHET models

The error metrics used in the cross-validation are mape (mean absolute percentage error) and rmse (root mean square error). These errors are calculated between the model predicted data and the unseen data, both in the test time frame. The smaller the value the better the predictability of model.

The tables below list rmse and mape as obtained for the SARIMA and PROPHET models respectively, together with the corresponding test time (excluded in the model training data) frame. Data is summarized in figure 8.

	CV SARIMA(111)(10112) rmse	CV SARIMA(111)(10112) mape	Start horizon date	End horizon date	Middle of horizon period
0	3.94	14.31	1994-07-01	1998-11-01	1996-09-01
1	3.41	12.43	1998-12-01	2003-04-01	2001-02-01
2	3.14	8.39	2003-05-01	2007-09-01	2005-07-01
3	2.46	5.67	2007-10-01	2012-02-01	2009-12-01
4	4.45	8.20	2012-03-01	2016-07-01	2014-05-01

	CV Prophet rmse	CV Prophet mape	Start horizon date	End horizon date	Middle of horizon period
0	5.00	18.06	1994-07-01	1998-11-01	1996-09-01
1	4.52	13.11	1998-12-01	2003-04-01	2001-02-01
2	5.99	17.58	2003-05-01	2007-09-01	2005-07-01
3	5.34	12.55	2007-10-01	2012-02-01	2009-12-01
4	6.19	11.44	2012-03-01	2016-07-01	2014-05-01

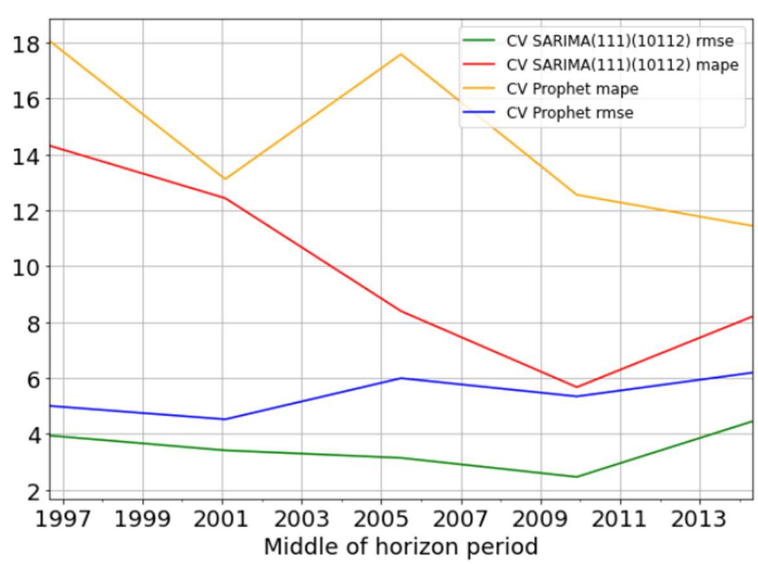


Fig. 8

Final average mape and rmse are summarized in the table below and showing lower errors for the SARIMA model.

Model	MAPE average over 5 horizon	RMSE average over 5 horizon
SARIMA	9.8%	3.5
FB Prophet	14.5%	5.4

Conclusions and Recommendations for Implementation

Based on this study, we recommend the more accurate SARIMA(1, 1, 1)x(1, 0, 1, 12) which showed 9.8% mean average error percent over rolling window validation. FB Prophet model also shows good performance in comparison to other models evaluated such as Prophet with multiplicative seasoning or ARIMA 6,0,6 (not shown).

Forecasted future CO_2 emissions from this model can be used as a benchmark for policy evaluation as a comparison reference for actual future data. This will help policy makers understand the effects of policy.

An added benefit of the used SARIMAX model is that it has the capability of the additional of exogenous variables. In this problem, for instance, a separate parallel time series relevant to policy can be added to the model. This would allow forecast of CO_2 emissions with and without the policy in place.

Yet another benefit is that the model is not very computationally demanding and can be run on a variety of cloud servers such as Google Colab (even without enabling GPU).

As a final note, we state that if strong trend changes are observed in more current data, the FB Prophet model would become beneficial as it has the capability to detect these changes change points.

References

https://www.eia.gov/electricity/data.php

https://www.eia.gov/electricity/