Fall 2020

DSC 551 Final Project

Ditching Coal: How the United States Is Moving Away from One Fossil Fuel

Project done by

Mukila Rajasekar

Introduction

According to a Union of Concerned Scientists article, “If the United States continues burning coal the way it does today, it will be impossible to achieve the reductions in heat-trapping emissions needed to prevent dangerous levels of global warming.” Scientists also say that we need to cut down carbon emissions by 45% by 2030, and reach net zero by 2050 or we might face irreversible and destructive climate crisis due to global warming. Coal is one of the most affordable and largest domestically produced sources of energy in the U.S. Although in the recent years, energy production using coal is on a decline, is this enough to battle global warming?

In this project, I will be exploring the U.S National Net Electricity Generation using Coal from 1995 to 2020 dataset available on the EIA website. I will be using various time series models learned in the DSC 551 course on this dataset. The goal is to evaluate their accuracy measures and select the best method to forecast the values for next 5 years.

Data Exploration

The dataset used in this project is taken from the official U.S. Energy Information Administration (EIA) website. Source: https://www.eia.gov/totalenergy/data/monthly/. The dataset shows the U.S. National Monthly Net Electricity Generation from Coal in Million kWh from January 1995 to June 2020. Some exploratory data analysis is performed to understand the structure of the dataset.

```
 Time-Series [1:306] from 1995 to 2020: 147220 132900 131430 123076 130418 ...
[1] "Frequency: 12"
[1] "Is it timeseries? TRUE"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40576  125610  148494  140902  162399  190135
```

From the above statistics, we can confirm that the dataset is a time series and it has a frequency of 12, meaning it is a monthly data.

Figure 1.1 shows the basic plot of the time series. From this, we can see that the time series has an overall decreasing trend since the last decade. There also seems to be some seasonality within each year.

The season plot in Figure 1.2 shows visible seasonality within each year. The electricity generation seems to increase during the summer and the winter months. Electricity generation is lowest during the spring months.

Thus, the time series dataset has both trend and seasonality.

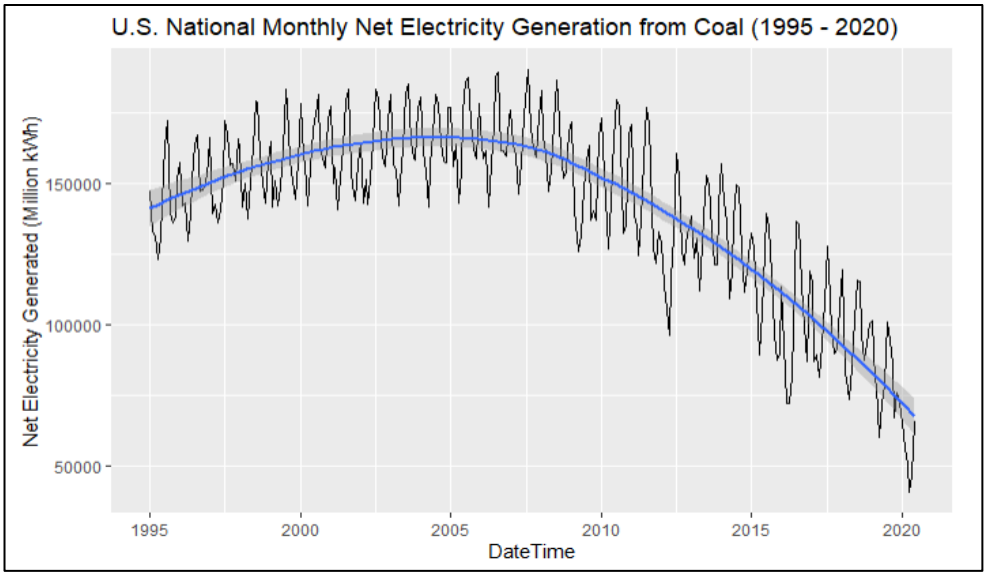

Figure 1.1

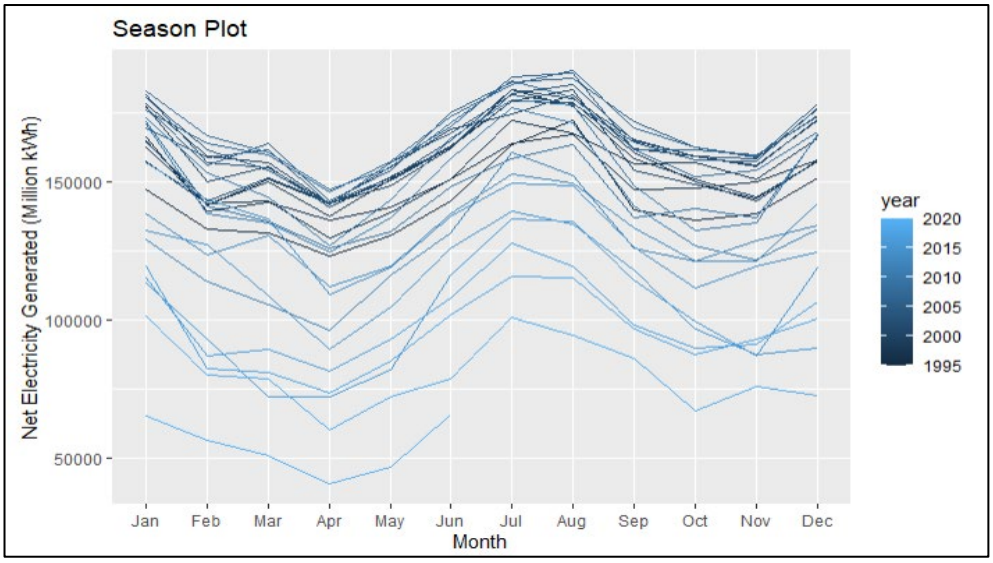

Figure 1.2

A simple box plot on the time series as seen on Figure 1.3 shows that the mean and variance seem to change from month to month. July and August have the highest means and March and April have the biggest variance. This shows that the time series has a strong seasonal effect.

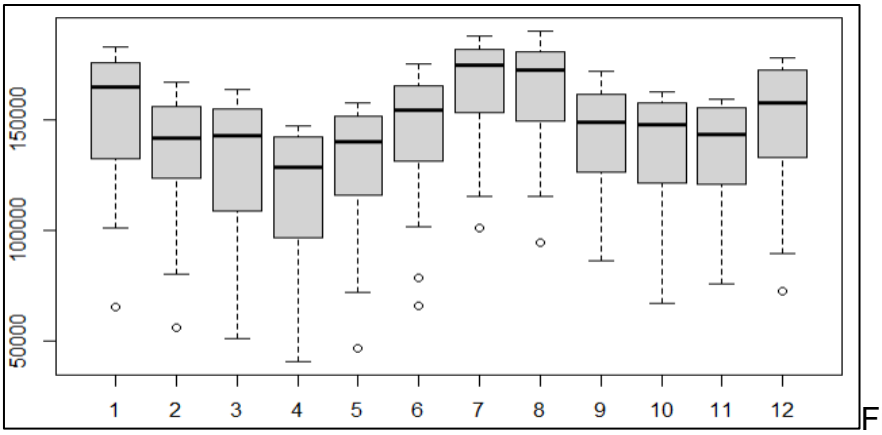

Figure 1.3

An ACF plot of the time series as seen on Figure 1.4 confirms that the series is not merely white noise and has strong trend and seasonality. The mean of the time series is 140902.44 and variance is 975070534.21.

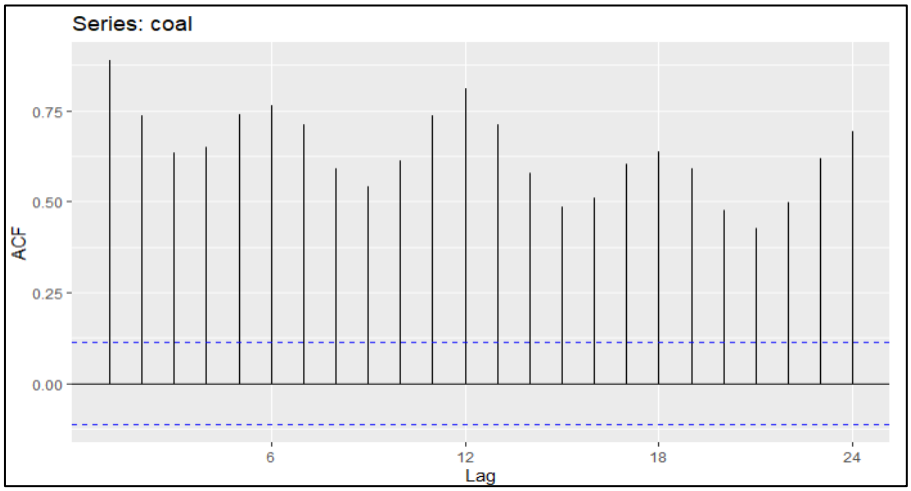

Figure 1.4

A simple gglagplot() on the time series also shows that there is a strong positive relationship at lag 12, indicating the seasonality of the series.

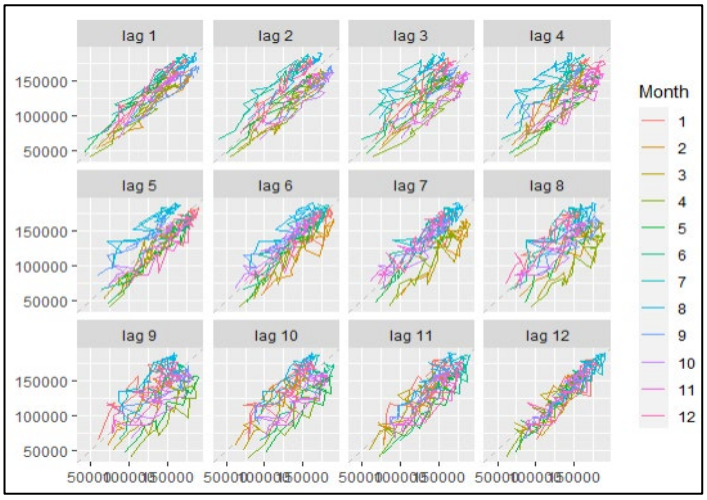

Figure 1.5

Basic Forecasting

Now that a clear picture of what the time series looks like and its general trend and seasonality are established, it's time to do some basic forecasting to see if the models are a good fit. Before proceeding further, it is better to split the dataset into a training set and a testing set so that the test set can be used to get the accuracy of the training models. A common way to split the dataset is using the **80/20** method. i.e. 80% of the data is used for training and 20% is retained as testing set. There are about 25yrs in this dataset, so I am going to keep the last 5 years as a testing set and use the remaining data for training the time series models.

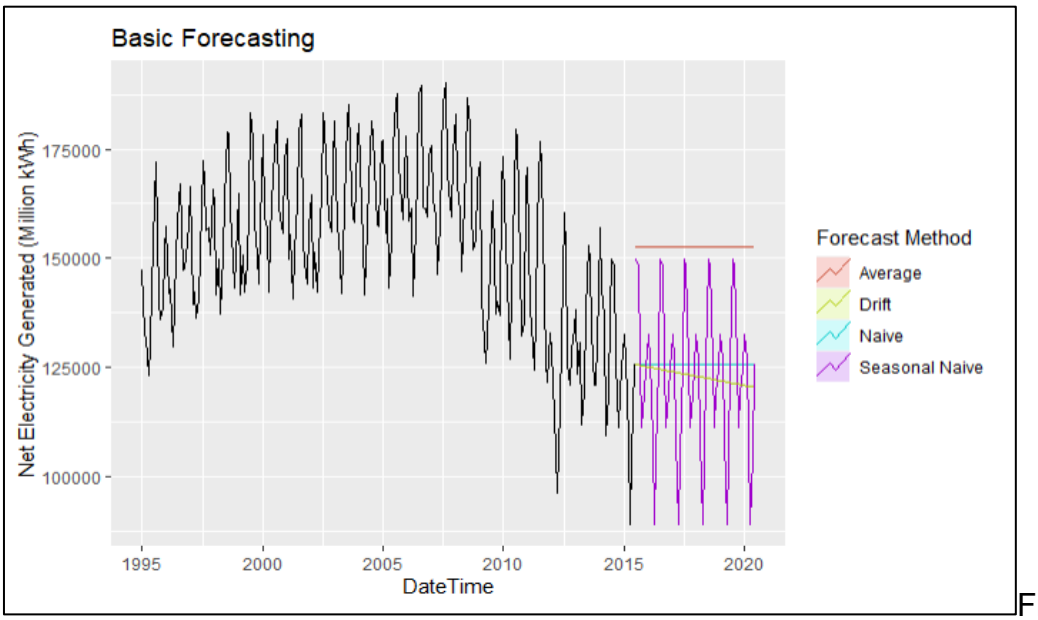

Figure 2.1

Sometimes, even simple forecasting methods can be very effective. The 4 benchmark forecasting methods are,

1. Average Method
2. Naïve Method
3. Seasonal Naïve Method
4. Drift Method

Figure 2.1 shows the forecasting results from these 4 methods. Their accuracy values are measured by testing the trained model on the test set.

	RMSE <dbl>	MAE <dbl>
Mean	64256.42	60203.30
Naive	40058.20	34671.19
Seasonal Naive	33647.10	29695.81
Drift	37338.28	32150.00

From the graph and the above table, we can see that the **Seasonal Naïve** method performs much better than the other 3 methods. It has the lowest error value among all 4 methods.

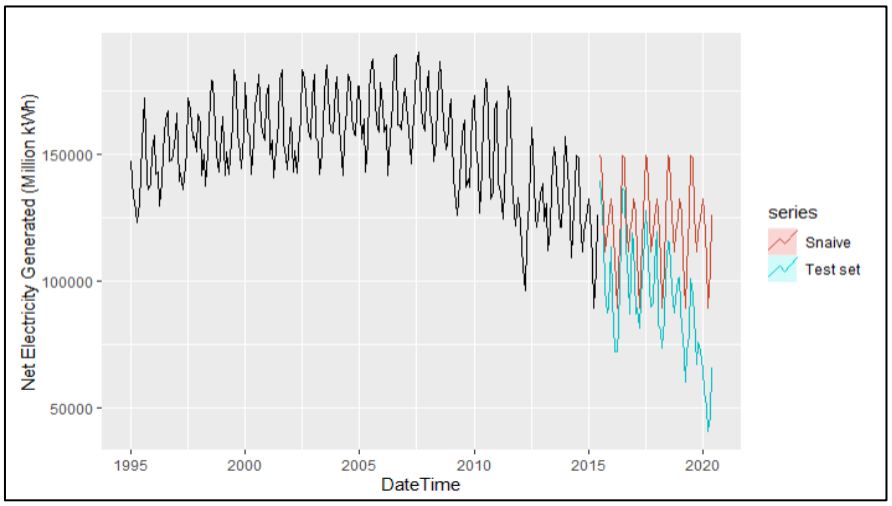

Figure 2.2

From figure 2.2, we can see that the seasonal naïve forecast values does not exactly match the test set. This is further proved in Figure 2.3, where residuals for the seasonal naïve method are plotted.

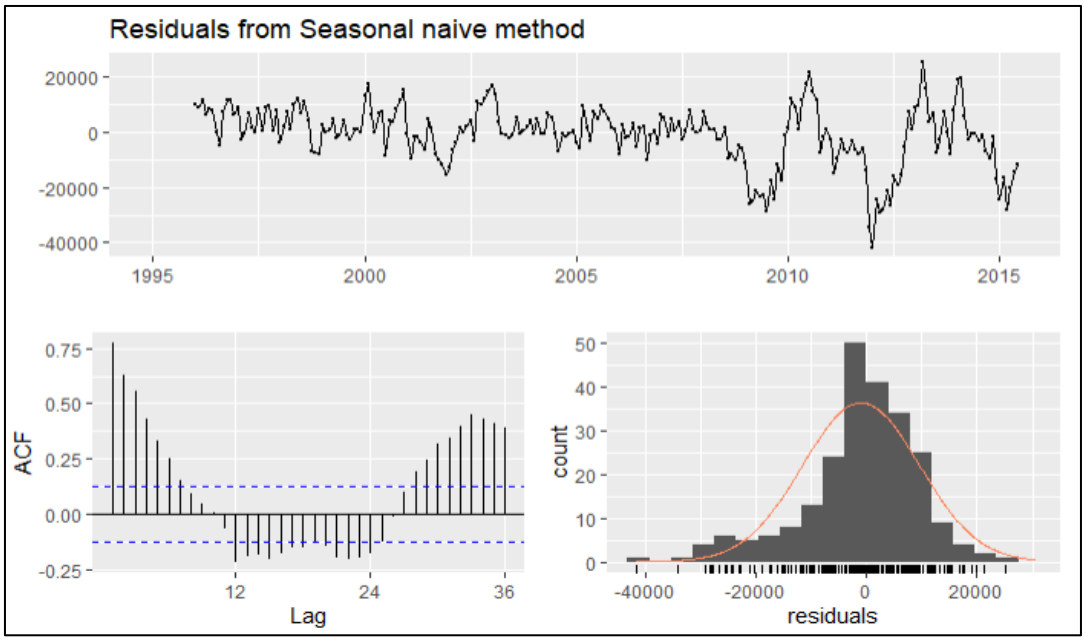

Figure 2.3

The residuals in the above ACF graph do not resemble a white noise meaning a lot of the seasonality is still unaccounted in the Seasonal Naïve forecasting method. So, it would be better to do some decomposition to separate the trend and seasonal components.

Before moving on to decomposition, the above 4 basic methods are used with cross-validation to see if it improves the accuracy.

Without Cross-Validation

	RMSE <dbl>
Mean	64256.42
Naive	40058.20
Seasonal Naive	33647.10
Drift	37338.28

With Cross-Validation

```
"Drift RSME: 52372.7742751471"
"Naive RSME: 26082.6538591161"
"Snaive RSME: 20076.8544210978"
"Average RSME: 35292.8015733204"
```

The above tables show that the models perform much better when using cross-validation and Seasonal Naïve method is still the best out of all 4 methods. But there are still a lot of improvements to be made on the forecasting process as the residuals do not resemble a white noise series.

Decomposition

A seasonal time series consists of a trend component, a seasonal component and an irregular component. Decomposing the time series means separating the time series into these three components and estimating their values.

Electricity Generation using Coal has changed drastically over time in our time series and Classical Decomposition methods are unable to capture seasonal changes over time. So, other sophisticated decomposition models like X11, SEATS and STL methods might be better suited for this time series.

After applying X11, SEATS and STL decompositions with multiple s.window and t.window values, the **STL** decomposition method with s.window=10 and t.window=10 seems to capture the trend of the time series better than the other methods. This is shown in Figure 3.1.

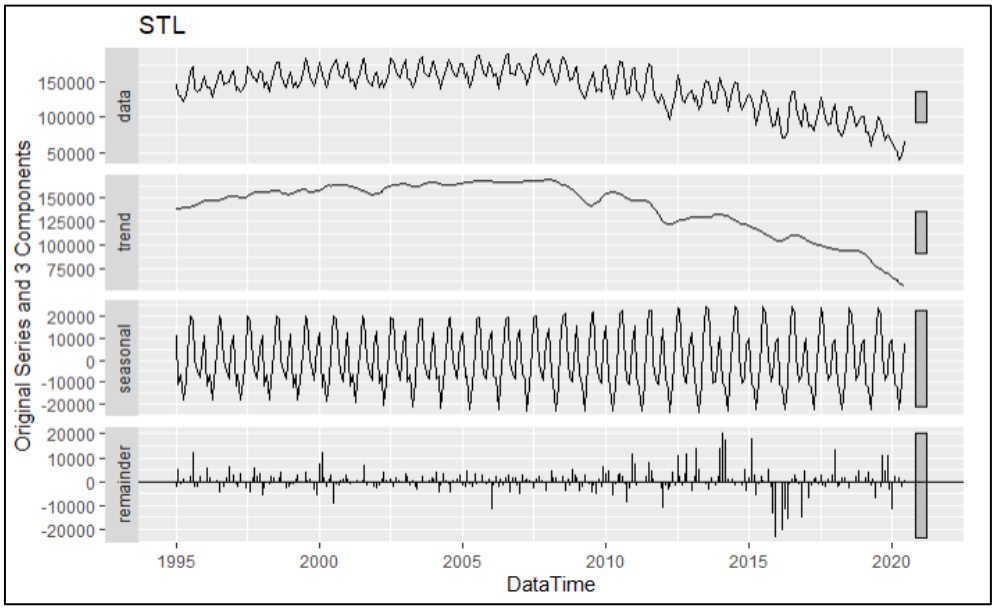

Figure 3.1

Next step is to do some forecasting on the decomposed series. For this purpose, Holt's Linear Method can be used.

Holt's Linear Method extends simple exponential smoothing to allow the forecasting of data with a trend. Since, Holt's method alone does not provide a good fit, a combination of Holt's Linear and Seasonal Naïve method are used on the decomposed series. The decomposed series is split into seasonal component and seasonally adjusted component (seasonality removed).

After applying Holt's linear method with damped trend on seasonally adjusted series and Seasonal Naïve method on the seasonal component, the series is re-seasonalized by adding in the forecasts of the seasonal component. The resulting forecast is shown in the figure 3.2.

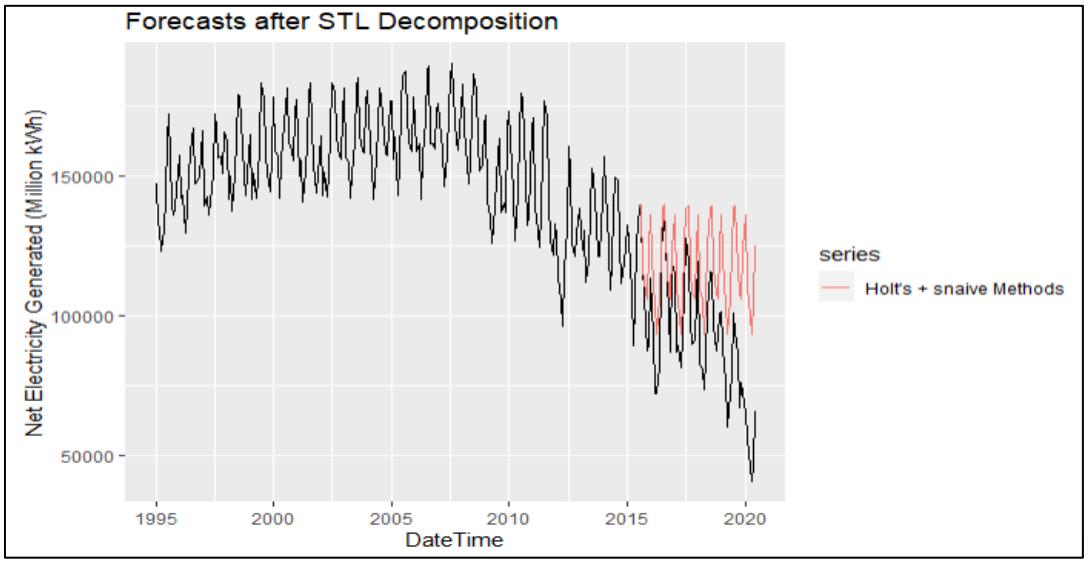

Figure 3.2

From the above graph and the table below showing the accuracy values for all the methods done so far, we can see that decomposing the time series prior to forecasting provides better results.

	RMSE <dbl>
Mean	64256.42
Naive	40058.20
Seasonal Naive	33647.10
Drift	37338.28
STL on Holt's + snaive	29480.68

Exponential Smoothing

Holt-Winters Method extends Holt's method to capture seasonality. The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations — level, trend and seasonal component with corresponding smoothing parameters α, β and γ. The additive method is preferred when the seasonal variations are roughly constant through the series, while the multiplicative method is preferred when the seasonal variations are changing proportional to the level of the series.

Since, this time series does not have constant seasonality, Holt-Winter's method with multiplicative seasonality and damped trend is used. Figure 3.3 shows the predictions from this method and we can clearly see that it has huge prediction intervals.

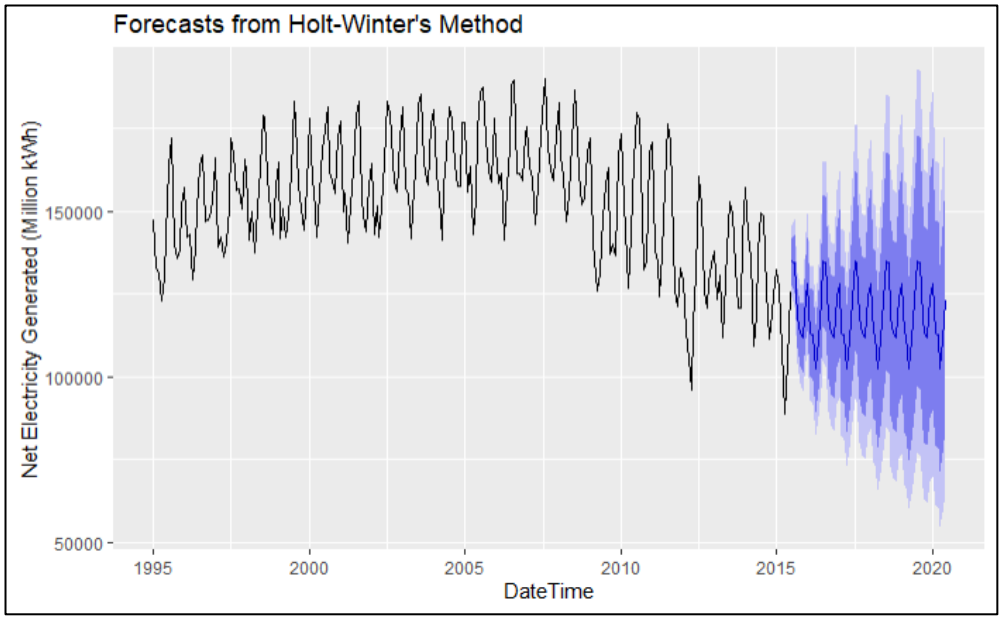

Figure 3.3

Having huge predication intervals does not mean the model is a good one. This is proved by the accuracy measures from the model as seen below.

	RMSE <dbl>
Mean	64256.42
Naive	40058.20
Seasonal Naive	33647.10
Drift	37338.28
STL on Holt's + snaive	29480.68
Holt-Winters	31325.66

ETS Model

A great advantage of the ETS statistical framework is that information criteria can be used for model selection. Using 'AIC' as the criteria, the best ETS model is selected by using the automatic ETS model selection function available in R.

The best ETS model picked based on 'AIC' value is **ETS(A,Ad,A)**. This model has Additive error, Damped Trend and Additive Seasonality.

```
ETS(A,Ad,A)

Call:
 ets(y = coal_train, damped = TRUE, ic = "aic", restrict = FALSE)

  Smoothing parameters:
    alpha = 0.7256
    beta  = 1e-04
    gamma = 1e-04
    phi   = 0.9749

  Initial states:
    l = 140159.9642
    b = 607.7708
    s = 7066.521 -8686.596 -7264.047 -1127.986 20405.34 20562.92
           4622.358 -10544.96 -21585.94 -7705.628 -7596.084 11854.1

  sigma:  5547.904

     AIC     AICc      BIC
5614.314 5617.327 5677.410
```

The small β and γ values show that the slope and seasonal components change very little over time. However, from figure 4.1 we can see that the model has broad prediction intervals. This shows that the model is struggling to forecast the time series. This is also reflected in the accuracy measure table and the ACF of the residuals in figure 4.2 tells us that there is still some seasonality in the series that it left out by the ETS model.

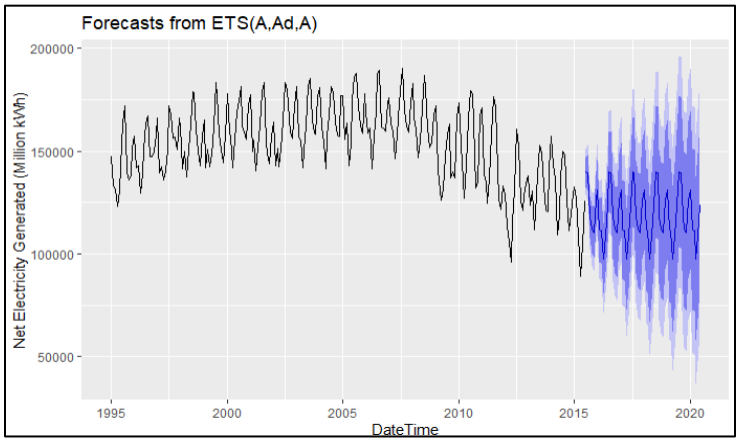

Figure 4.1

	RMSE <dbl>
Mean	64256.42
Naive	40058.20
Seasonal Naive	33647.10
Drift	37338.28
STL on Holt's + snaive	29480.68
Holt-Winters	31325.66
ETS(A,Ad,A)	31134.06

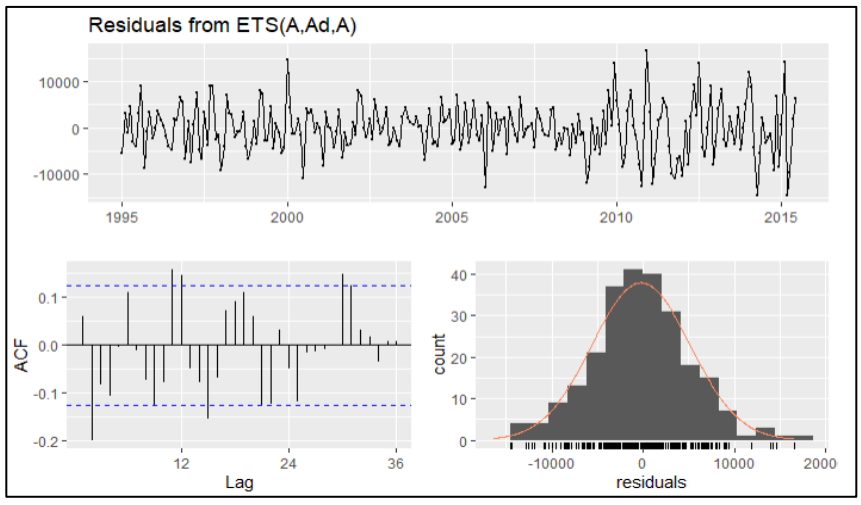

Figure 4.2

ARIMA Model

While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data. In order to select a good ARIMA model, we need to first check if the time series needs differencing or not. For this purpose, an ADF Test is performed on the time series.

```
        Augmented Dickey-Fuller Test

data:  coal_train
Dickey-Fuller = -2.5905, Lag order = 6, p-value = 0.3273
alternative hypothesis: stationary
```

The p-value of the ADF Test is too large, meaning the time series is not stationary and needs differencing. After performing seasonal differencing and repeating the same steps for another ADF Test as mentioned above, it is clear that the time series needs first order differencing on both seasonal and non-seasonal components.

```
p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  diff(coal_train, lag = 12)
Dickey-Fuller = -4.7211, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

The graph in figure 5.1 shows the time series after first order seasonal differencing.

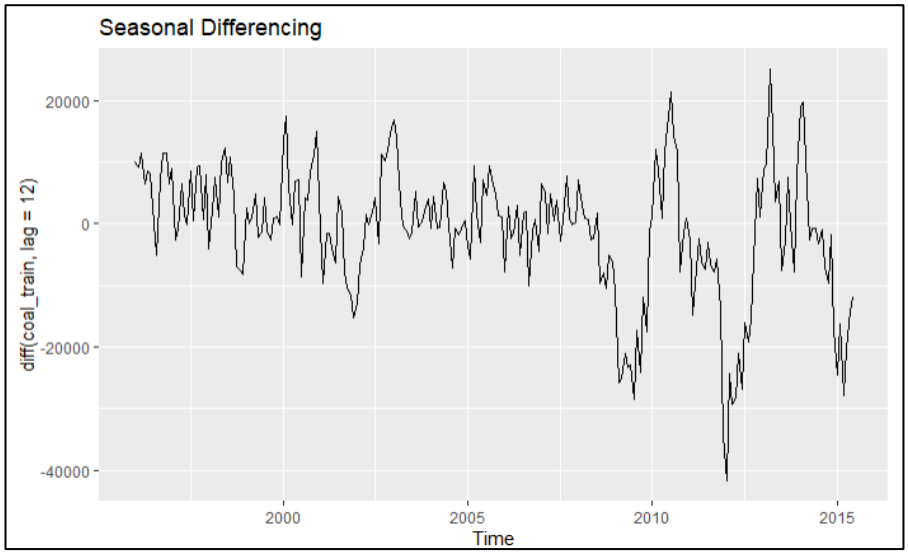

Figure 5.1

The ggtsdisplay() function on seasonally differenced series shows significant spikes at lag 12 of ACF plot and lags 12 and 24 of PACF plot. This proves the strong seasonality of the time series. This is shown in figure 5.2.

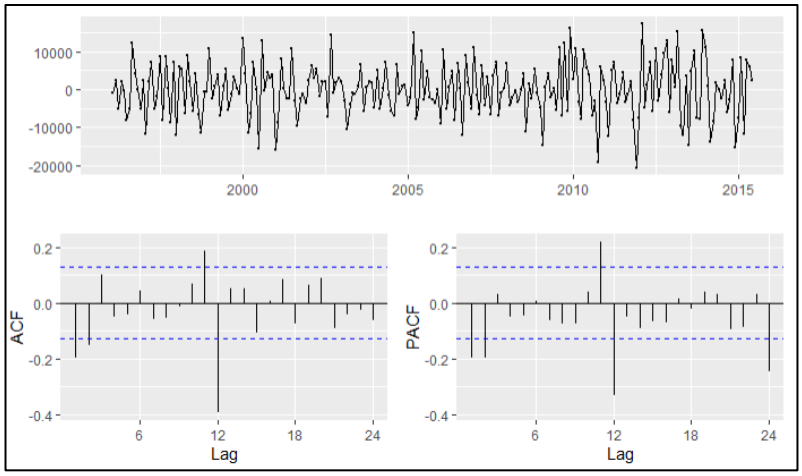

Figure 5.2

For the initial ARIMA model, P=1, D=1 and Q=2 are selected, and the resulting forecasting model seems to account for all the seasonality within the series as seen in Figure 5.3

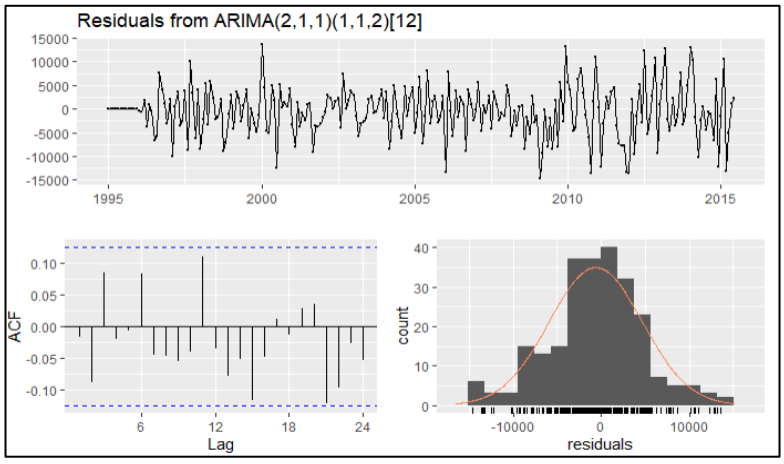

Figure 5.3

However, before finalizing on this model, I try a few other variations of this model to see if I can get better results. Out of all the models tested, **ARIMA(3,1,3)(1,1,2)[12]** provides the best forecasting result and lowest error value. The residuals in the ACF plot also clearly resemble a white noise series as shown in figure 5.4

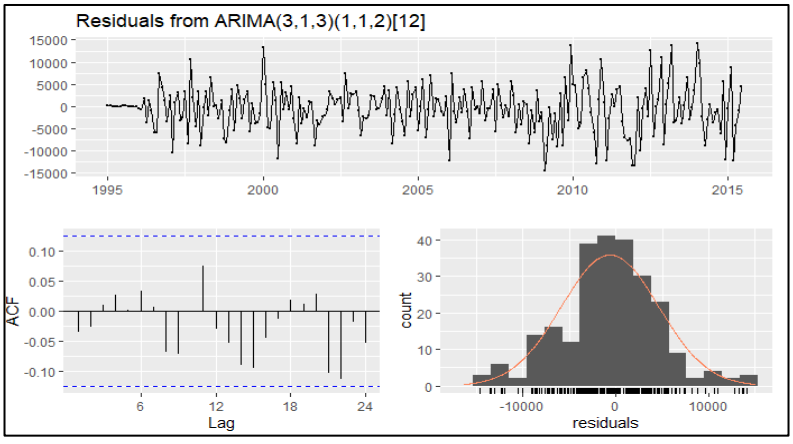

Figure 5.4

The forecasts from the best ARIMA model have narrow prediction intervals and lowest error value among all the forecasting models explored so far as seen in figure 5.5 and the accuracy table below.

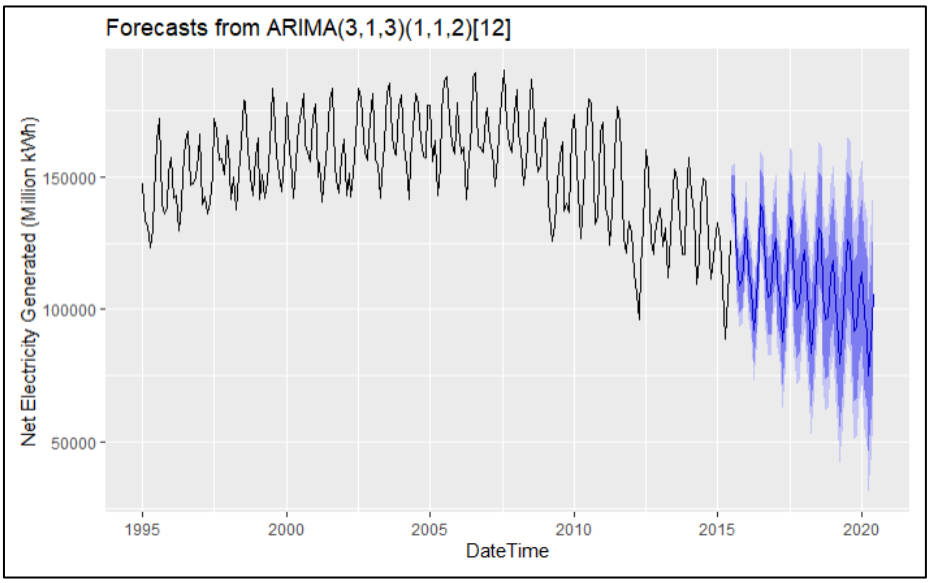

Figure 5.5

	RMSE <dbl>
Mean	64256.42
Naive	40058.20
Seasonal Naive	33647.10
Drift	37338.28
STL on Holt's + snaive	29480.68
Holt-Winters	31325.66
ETS(A,Ad,A)	31134.06
ARIMA(3,1,3)(1,1,2)[12]	20919.52

ETS vs ARIMA using Cross-Validation

Before concluding on the best forecasting model, cross-validation is performed on the best ETS and best ARIMA models for comparison.

```
      ETS    ARIMA
7582.407 7295.512
```

The Root Mean Squared Error Values from cross-validation results of the best ETS and ARIMA models show that the seasonal ARIMA model ARIMA(3,1,3)(1,1,2)[12] is still the best model out of all the forecasting methods explored in this report.

Figure 5.6 shows a comparison graph between the forecasted values from the best ARIMA model and the actual test set values. The prediction intervals of the ARIMA model cover most of the test set values. This shows that the final model selected is a good one.

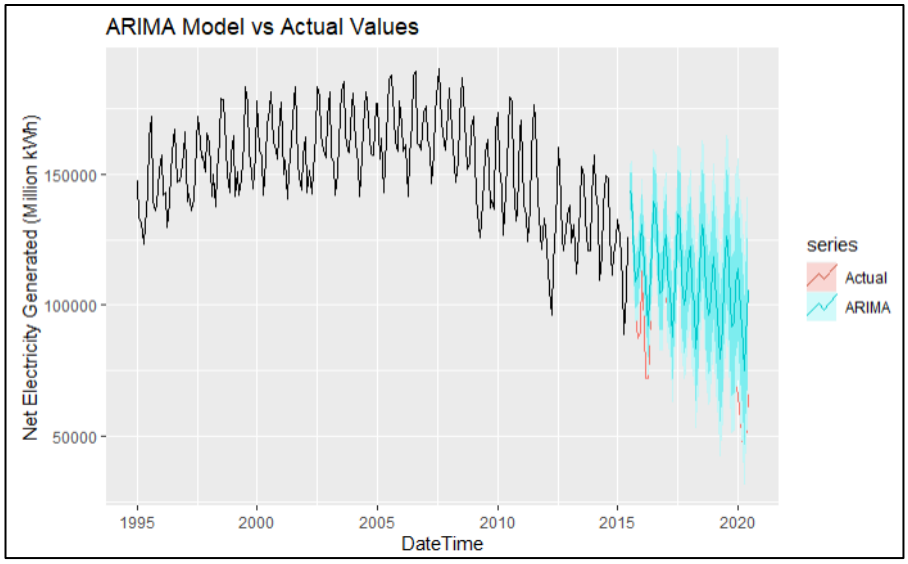

Figure 5.6

Predicting the future

Now that the best forecasting model is selected for this time series dataset, it is time to use this model on the whole dataset, instead of just the train set, and predict the Net Electricity Generation using Coal for the next 5 years.

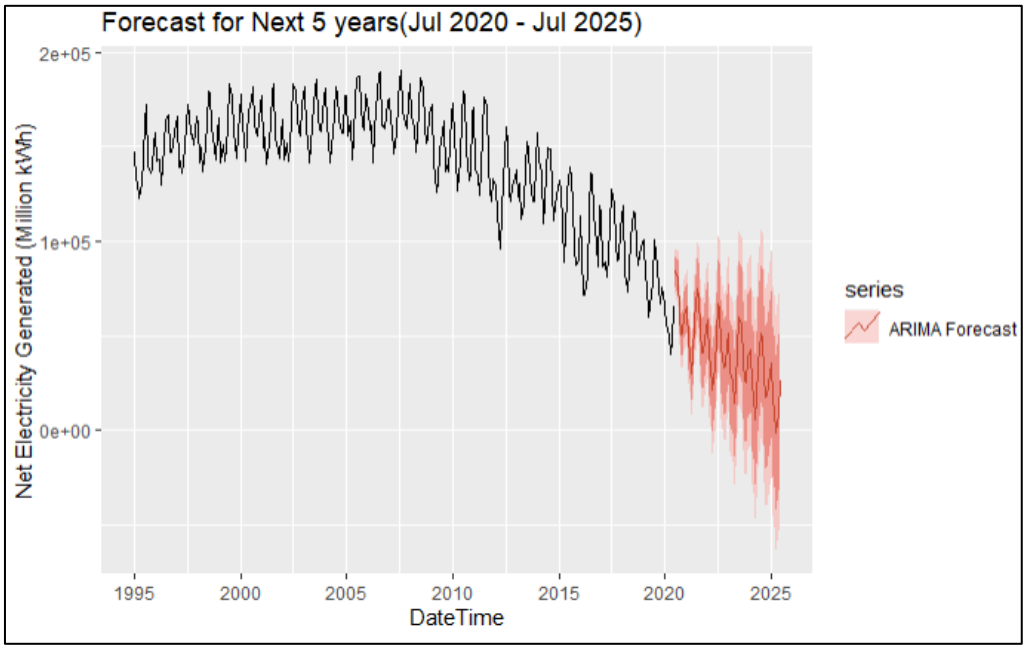

Figure 6.1

The above figure shows the projected national electricity generation values for the next 5 years based on the best ARIMA model selected previously.

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>
Sep 2024	29322.765	-7691.7662	66337.30
Oct 2024	17711.632	-19812.7955	55236.06
Nov 2024	18788.817	-19252.9259	56830.56
Dec 2024	30477.692	-8067.3100	69022.69
Jan 2025	35316.478	-3714.8311	74347.79
Feb 2025	15801.336	-23728.3820	55331.05
Mar 2025	10760.455	-29245.9372	50766.85
Apr 2025	-1175.532	-41655.6971	39304.63
May 2025	9337.205	-31622.3432	50296.75
Jun 2025	26330.779	-15085.1139	67746.67

The overall forecasted trend for electricity production using coal in the U.S seems to decrease continuously over the next 60 months with point forecast even hitting negative values in 2025.

Conclusion

Among the various time series forecasting models and methods explored, the seasonal ARIMA(3,1,3)(1,1,2)[12] is the best forecasting model for this time series dataset. The forecasting results from this model for the next 5 years show the national electricity production from coal to decrease steadily in the future. In an ideal world, if these forecasted results are reflected in the future, United States might achieve net zero electricity generation using coal by mid-2025.

References:

- Information about U.S Electricity Generation from EIA
  https://www.eia.gov/tools/faqs/faq.php?id=427&t=3

- “Coal’s Decline Continues with 13 Plant Closures Announced in 2020”
  https://www.scientificamerican.com/article/coals-decline-continues-with-13-plant-closures-announced-in-2020/

- “Coal Power in a Warming World”
  https://www.ucsusa.org/resources/coal-power-warming-world

- “Climate Change - How hot cities could be in 2050”
  https://www.bbc.com/news/newsbeat-48947573