**Forecasting Energy Demand**

**Introduction:**

Rising socio-economic factors like increase in population, urbanization, industrialization etc, leads to an increased need for electricity and taking into account the past trends in energy consumption can significantly improve the accuracy of power demand forecasting.

**Objective:**

The objective of this project is to identify the best forecasting model to forecast energy demand for the next 50 periods (September 2018 - October 2022). Power demand forecasting helps modern power utility companies make informed decisions about operations, supply. The dataset we have chosen contains the American Electric Power consumption (MW) observations from 2004 to 2018.

**Data Collection and cleansing:**

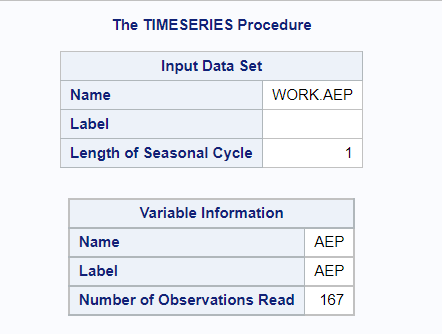
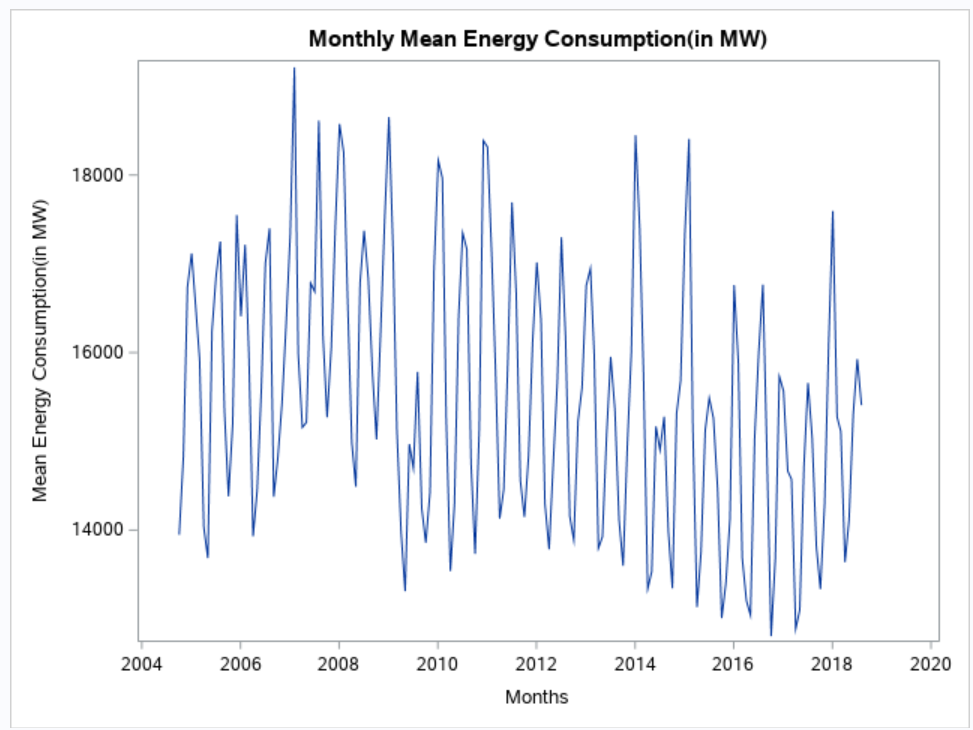
The dataset used in this project is sourced from Kaggle website[1]. The source file consists of hourly energy consumption in Megawatts from a major power Utility company, American Electric Power, that supplies electricity to around 5.5 million customers in 11 states[2]. The data ranges from October 2004 till August 2018. The source file consists of 2 columns - Datetime and AEP\_MW.

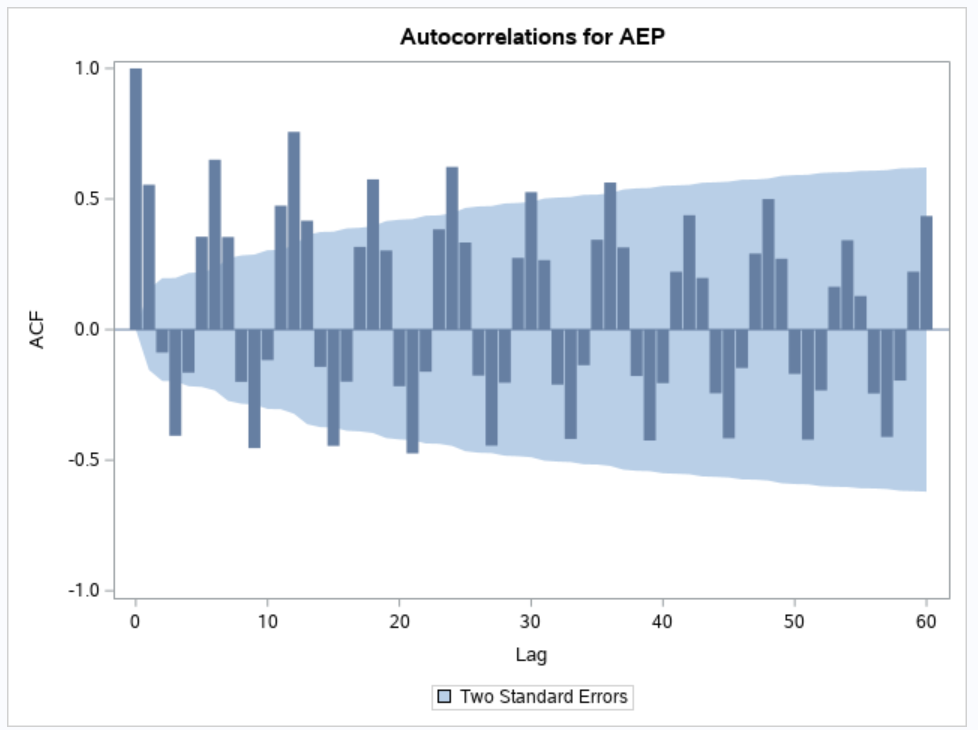
The dataset is aggregated to monthly data with 167 records using pivot table in excel.

The dependent variable, variable that needs to be forecasted is AEP. The independent variable is the time period in months from October 2004 till August 2018.

**Identifying the time series components:**

Based on the time series plot, we could see that the data exhibits seasonality,a slight overall increasing trend and a potential cyclic component. The ACF plot has significant peaks at lags 12,24,36 etc which confirms the seasonality. The autocorrelations quickly dropped to 0 by 3rd lag which indicates that there is no trend.





**Model Selection:**

Based on the time series components identified above, we have tried the below forecasting models:

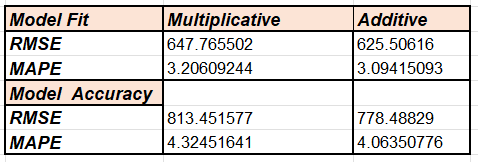
1. Holt-Winters Exponential Smoothing (Additive and Multiplicative)
2. Regression model with Seasonality (Deseasonalizing the data and using dummy variables)
3. Classical and X11 Decomposition
4. ARIMA

**Forecasting Models and Data Analysis:**

WINTER’S : Additive and Multiplicative:

From the timeseries plot, we can see that the magnitude of the seasonal component is not changing drastically. This indicates that Holts-Winter Additive would be a better model for the data. However, we confirmed the same by comparing the measures from both Additive and Multiplicative methods.

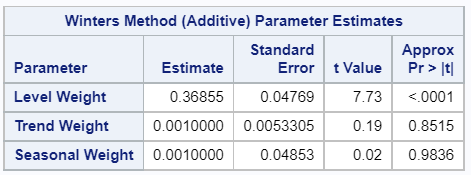
Below are the error measures we obtained ,



We can see from the above values that error measures for model accuracy for Additive is less when compared to multiplicative.As our goal is to generate better forecasts , we proceeded with Holt’s Winter Additive Model.

**Winters Additive:**

Using the Additive Winter’s Exponential Smoothing , the values of AEP for the next 24 months were predicted.



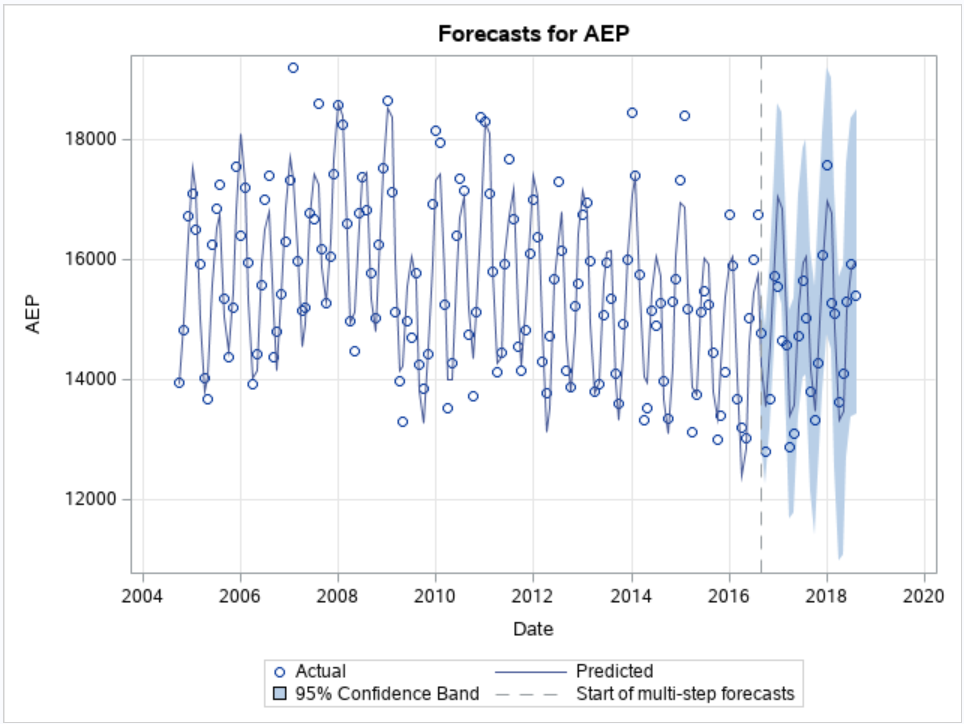
The parameter estimates obtained are as follows,

Level Weight - 0.36855 , indicates low values are assigned to most recent observations.

Trend Weight - 0.001000 ,indicates slope is barely changing.

Seasonal Weight - 0.00100 , indicates the seasonal component is barely changing over time.

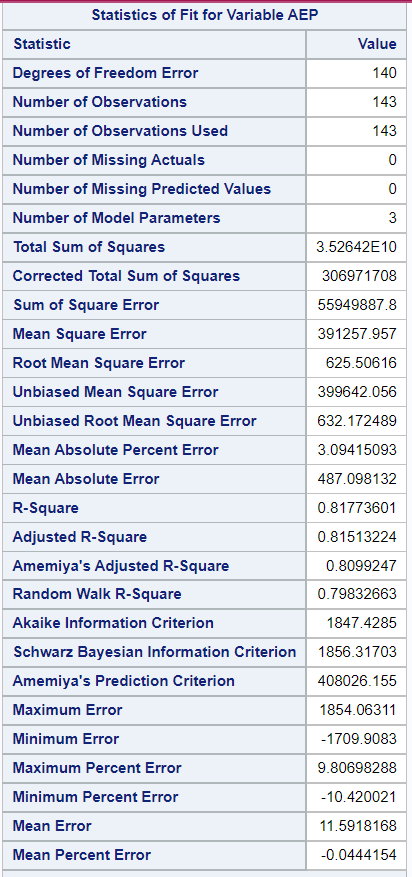
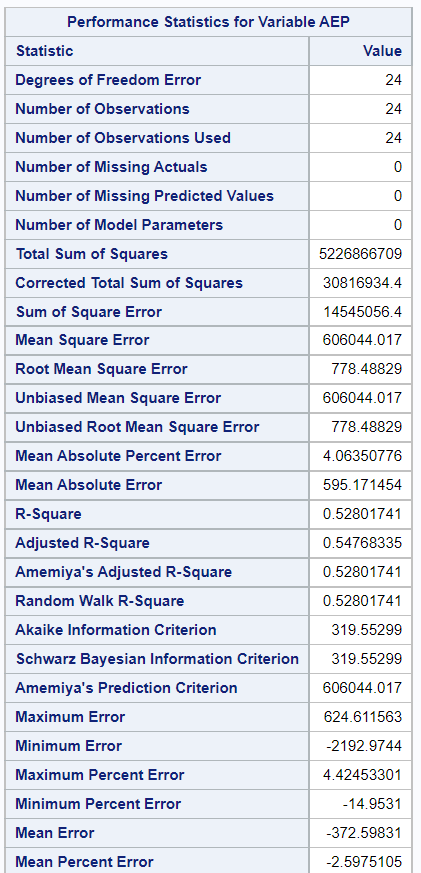
Below is the graph showing Actual and the predicted values.



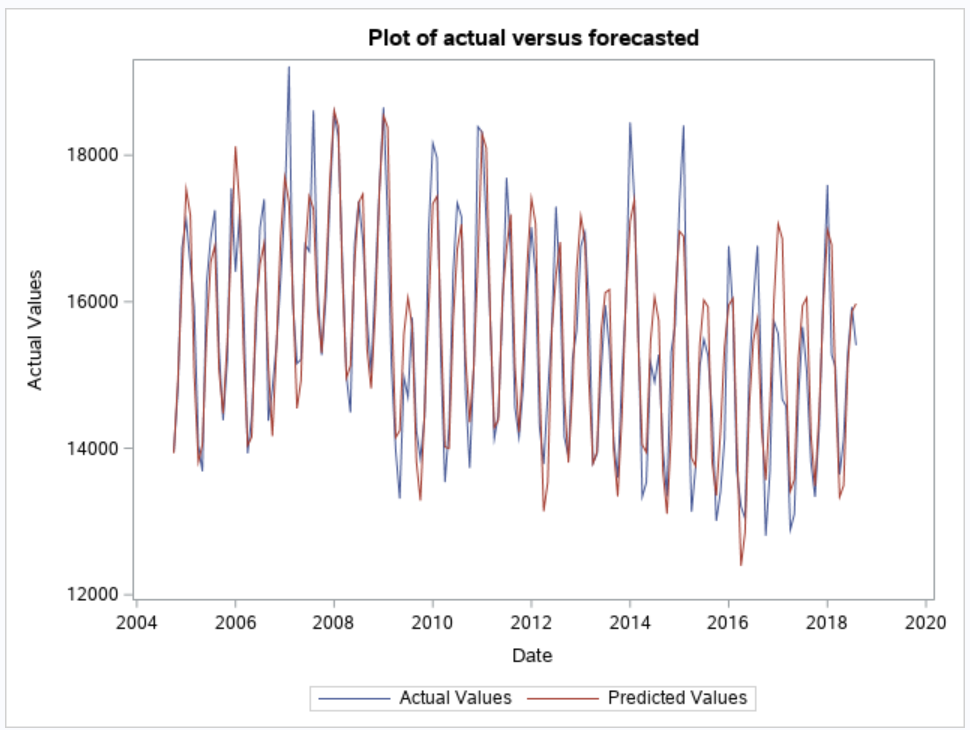
From the above plot, we can see that in the forecasted area , the seasonal component is not changing much.

Dividing the observations into training and test set (143 observations as training set and 24 observations as test set) to obtain model fit and model accuracy.

If we consider other error measures such as AIC and BIC , the values obtained for model accuracy are better which indicates a good fit .

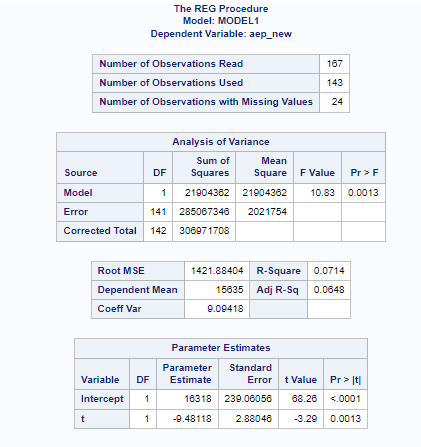
 

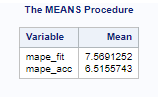
Plot of Actual v/s Forecasted



From the plot of actual versus forecasted value, we can see that the seasonal component is captured well except for the peaks indicating its a good model .

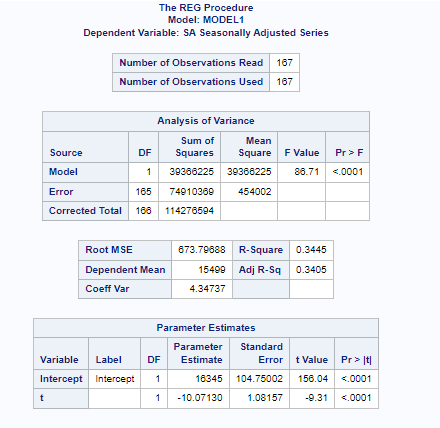
**Deseasonalizing using Seasonal Index :**

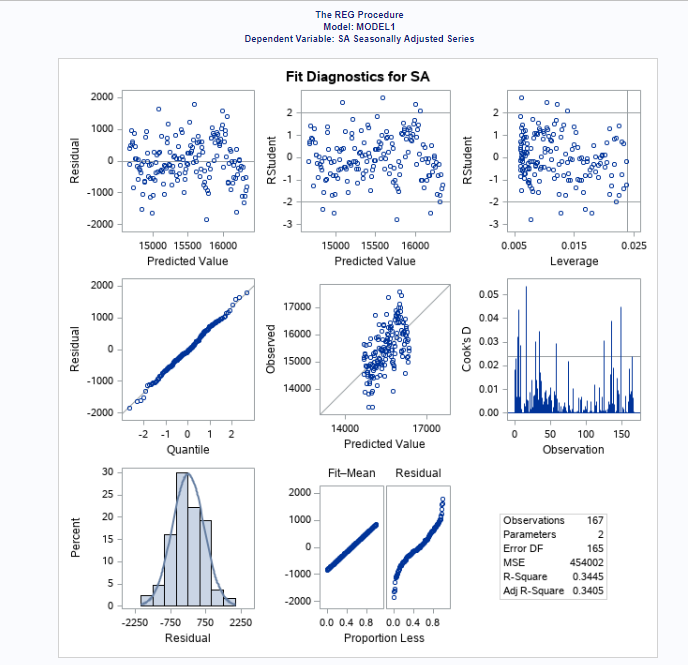




The values obtained above are the values we obtained before we got seasonally adjusted data.

The below figure shows the values obtained after re-seasonalising the data.





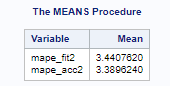
Model Assumptions :

* Relationship between aep and time variable looks linear.
* The histogram is bell shaped symmetric which indicates normal distribution is true for residuals.
* The scatter plot does not show any pattern which indicates residuals have constant variances.
* As per DW test , p-value is less than alpha , this indicates positive auto correlation.We reject the null hypothesis, residuals are independent of each other and assumption is not true.

Model Evaluation:

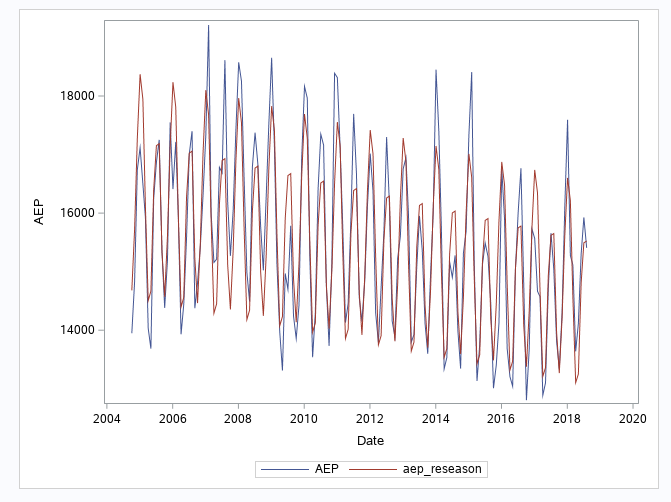
* The slope is declining as the coefficient of slope is negative.
* p-value is less than alpha , indicating slope is statistically significant.
* R2 Coefficient of determination is 34% which shows 34 percent of variation in energy consumption is explained by the variation in time variable . The value of R2 is not good enough , indicating mean monthly energy consumption alone is not sufficient to forecast.

Error Measures maps fit and maps accuracy after re-seasonalizing the data were better .



MAPE accuracy is 3.38% indicates good fit.

Time series plot of deseasonalized data and actual data:

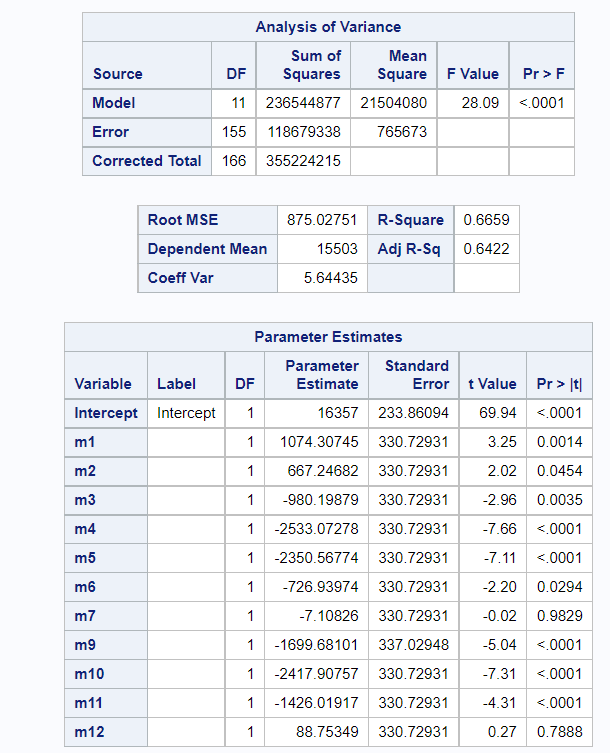


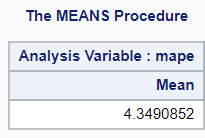
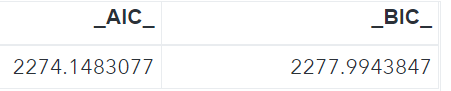
**Deseasonalizing using Dummy variables:**

The other method to account for seasonality in a regression model is by using dummy variables. The categorical variable date in our time series has 12 levels representing monthly data. Hence, we created 11 dummy variables for each month with August month as the reference period.

Model evaluation:

The model is logical as the sign of the slope coefficients of independent variables makes sense. The slope of all the months except July and December are statistically significant. The p-value of the model from the Analysis of Variance shows that the model is statistically significant. The adjusted R-squared is 64.22 percent which indicates that the seasons explains 64.22 percent of the variations in the energy consumption. This indicates that this model is not a very good fit but a decent fit though. The Variance Inflation Factor for all the dummy variables are less than 10. This shows that there is no multicollinearity between them.





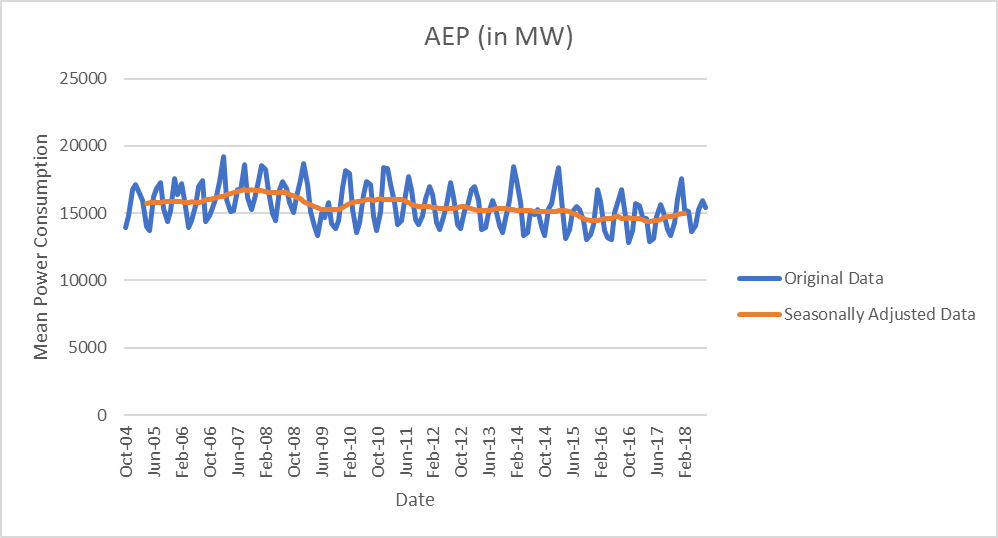
The MAPE is 4.34% which indicates a good fit.

As we go with a simpler model and to follow the principle of parsimony we ignored this model as we had to create 11 dummy variables .

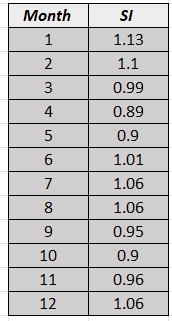
**X11 and Classical Decomposition:**

**Classical Decomposition:**

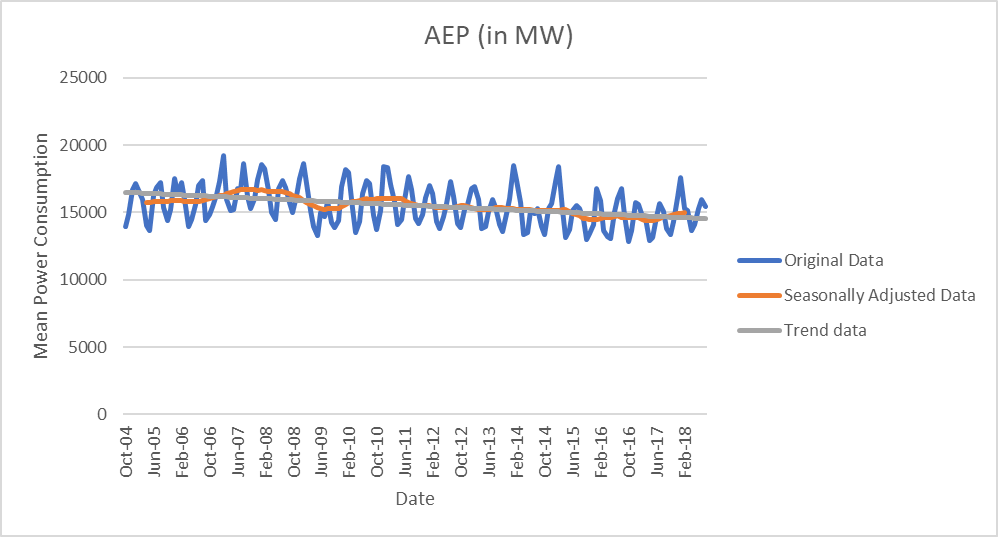
The classical decomposition method allows us to examine the time series components individually by breaking the series and then combine them to produce a better forecast. The first step in this method is to remove short term fluctuations, that are the seasonality and irregular variants, in order to clearly identify the long term trend and cyclic component. However, there is no long term trend in our data,and we could observe the potential cycles peak from October 2007 to June 2011. This process is accomplished by using a 12 period moving average(MA). Since we have used an even number of periods in calculating the moving average, we calculate a centered moving average(CMA) by performing a 2 period moving average of the moving averages calculated in the previous step. This Centered Moving Average CMA is also known as Deseasonalized data or Seasonally adjusted data. In order to measure the degree of seasonality, that is the seasonal factor (SF), we get the ratio of actual value to the seasonally adjusted value. We then take the average of SF’s for every season to obtain the seasonal Index(SI).



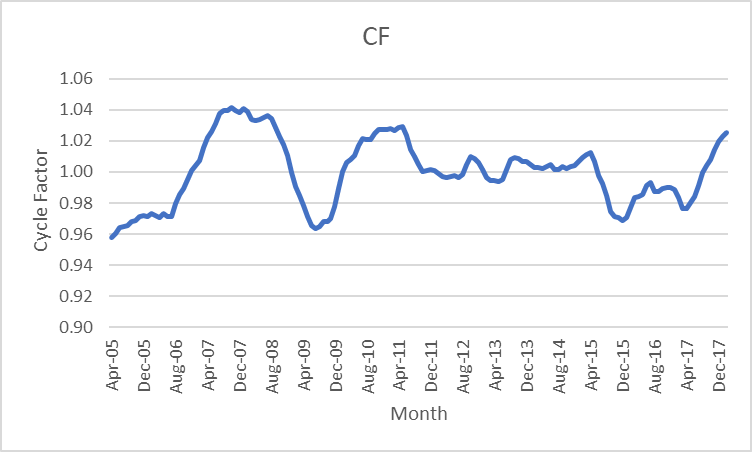
The seasonal Index of the month January is the highest with 1.13. This indicates that the mean monthly consumption of energy in January is 13% higher than the average month. The seasonal index of the month April is the lowest with 0.89. This indicates that the mean monthly consumption of energy in April is 11% lower than the average month.



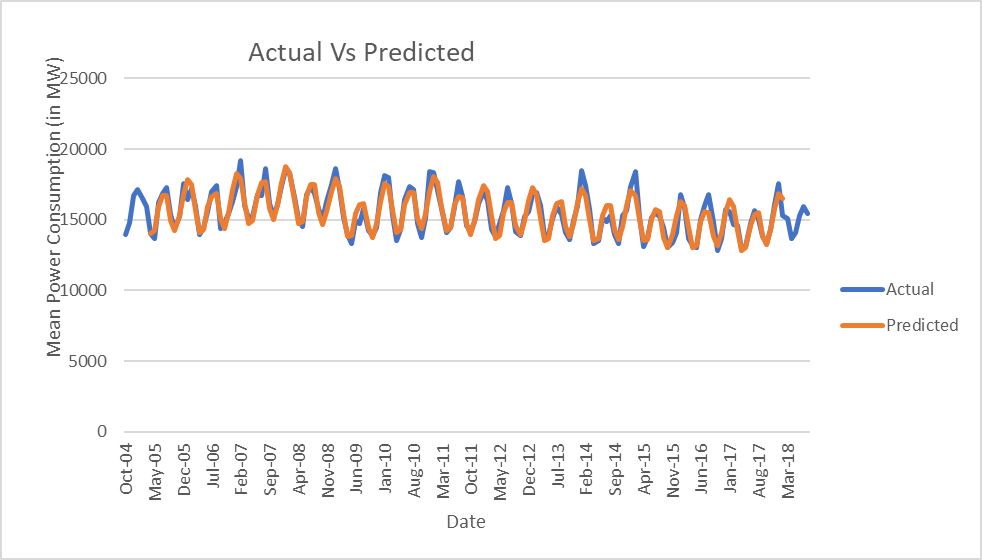
The second step in this method is to forecast the trend. This is done by applying regression on the seasonally adjusted data to obtain a linear equation with time variable. The CMA Trend(CMAT) series is obtained by substituting the time index in the trend equation.



The third step is to forecast the cyclical component which is most difficult to analyze and predict. This is done by obtaining the cycle factor(CF) which is the ratio of CMA and CMAT. From the below figure, we can see that anytime CF is greater than 1, the power consumption is in expansion phase and anytime CF is less than 1, the power consumption is in contraction phase.



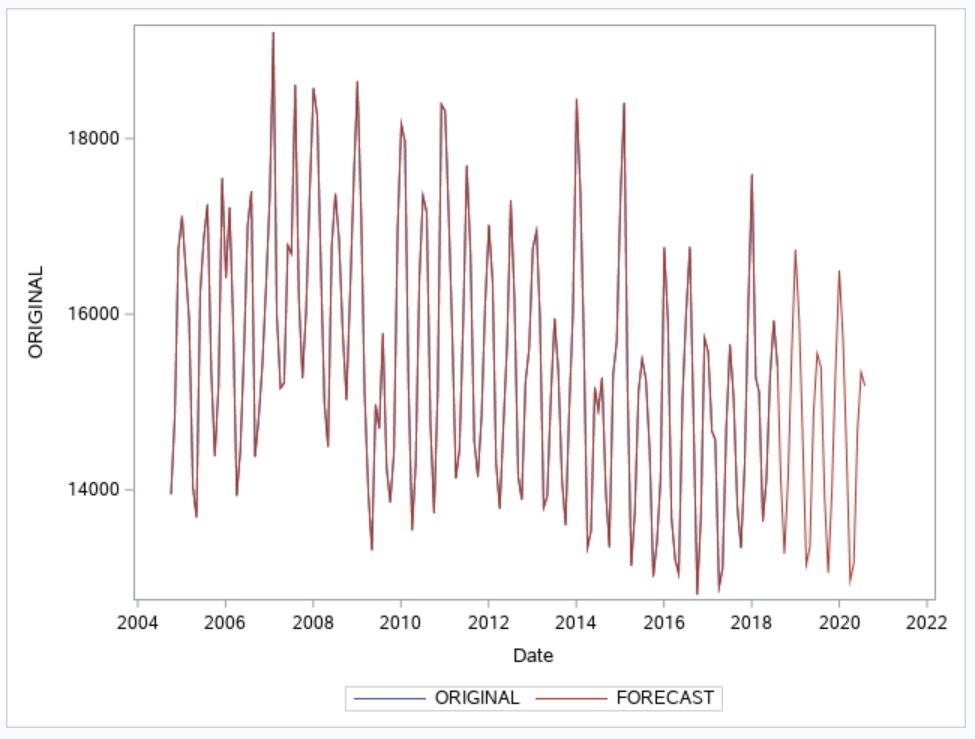
The final step is to combine the forecasted components by taking the product of SI,CF and CMAT. MAPE of the model fit is 2.67% which is less compared to other models. MAPE of the model accuracy is 3.21%. Based on these error measurements, this model is a good fit for the data.

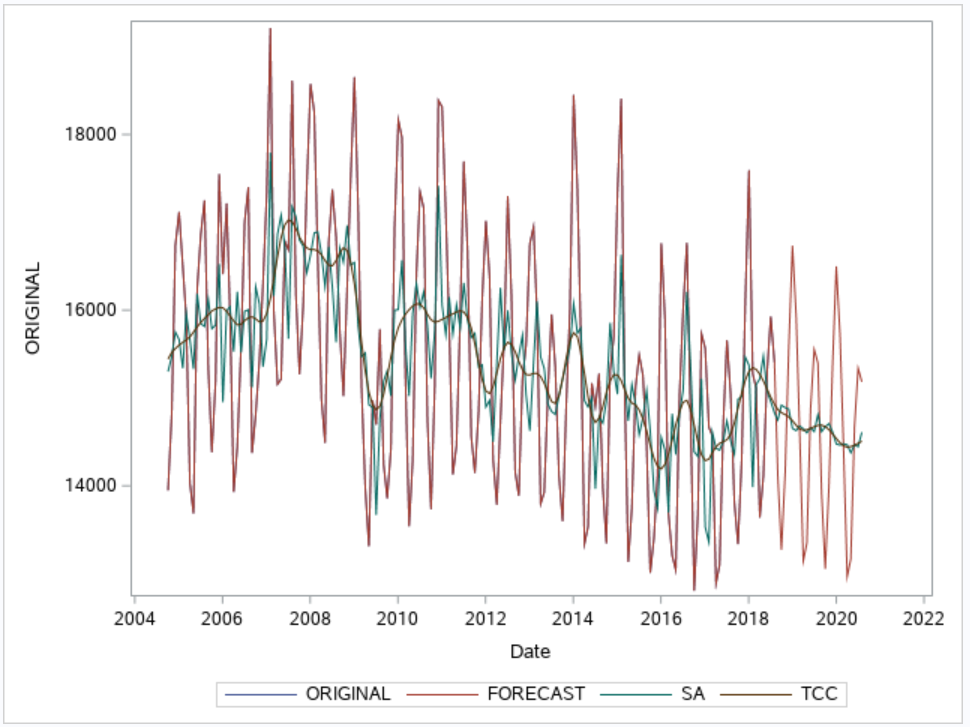


**X11 Decomposition:**

Looking at the graph with the original and forecasted plots of the X11 model, the plots line up perfectly. We can see a clear seasonality component from 2004 to 2019 and into the forecasted years of 2020 and 2021. The seasonally adjusted data and trend cycle component were also plotted and we see that there is less noise in the seasonally adjusted data, but the data still captures the seasonality of the original data. The trend cycle component has even less noise than the seasonally adjusted data but has multiple spikes in the data to be able to determine if there is a cyclical component. A potential trough is identified in 2009 to another potential trough in 2017. We can see that the energy consumption demand is at its lowest point in 2005 and then again in 2009 and one last time in 2017. We see a peak in 2007 when the demand for energy was at its highest and then again in 2014-15 as well as 2018.

The MAPE of the model fit and model accuracy was calculated for the X11 decomposition. The MAPE for the model fit was 0% and the MAPE of the model accuracy was 4.17%. The error measurements on this model are very impressive and therefore indicate that the X11 decomposition model is a good fit for forecasting the demand for energy consumption.





**ARIMA:**

The first step in creating the ARIMA models is to identify if the overall data is white noise. After checking the data, we see that all the p-values are less than alpha at the .05 significance level. During this first check, we also see that the data is not stationary. So we have to apply first order seasonal differencing. Once this first order seasonal differencing is applied, we see that the data data is still not stationary, so we proceed with non seasonal differencing to check if the data became stationary. After applying regular differencing we could observe some stationarity in the data .

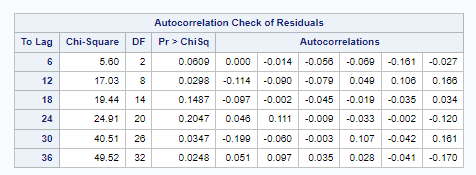
The next step is to identify the potential model and proceed with order of the seasonal and nonseasonal MA and AR terms. Since the data is monthly, we are looking for significant lags at every 12th lag in the ACF and PACF plots. In the ACF plot, we see a significant spike at the 12th lag. In the PACF plot, we do not see a significant spike at the 12th lag, the 24th lag, or the 36th lag. Now onto the nonseasonal terms. Looking at the PACF plot, there is no pattern and a potential pattern in the ACF. So we can test two ARMA models, one with an AR approach and the other with an ARIMA approach.

If we were to continue with an ARIMA model that has an AR approach, we would then look to the PACF plot for any more significant spikes other than the 12th lag. There are three significant spikes in the PACF plot. If we were to continue with an ARMA approach where there are no patterns in either the PACF or ACF plot, we find two significant spikes in the PACF and one significant spike in the ACF plot. So, the two ARIMA models that we tested were

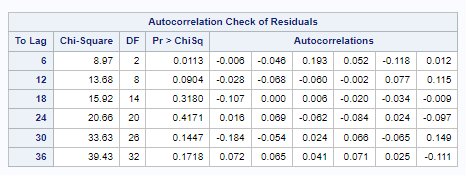
1. ARIMA (3,1,0)(0,1,1)
2. ARIMA (2,1,2)(0,1,1)

But, within the estimate step, we see that not all of the p-values for those residuals are above a .05 significance level for either of the two models, so we cannot continue with either ARIMA model to forecast the demand for energy consumption.

ARIMA (3,1,0)(0,1,1):



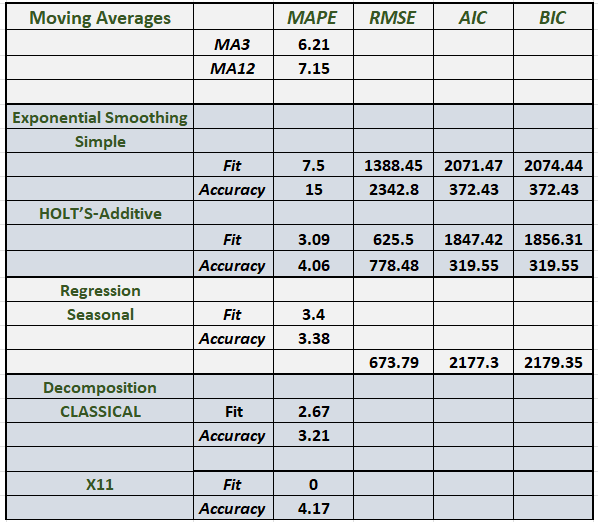
ARIMA (2,1,2)(0,1,1):



**Comparing the Models:**

On comparing all the methods tried above, based on the actual vs forecasted plots, the X11 decomposition method is the best forecasting method that would fit the dataset well. The forecasted values from X11 decomposition perfectly fits/overlaps with the actual data.

Considering the error measures, the Classical decomposition method appears to be the best forecasting method as the MAPE of the model accuracy is lowest compared to the other methods.



**Conclusion:**

From the comparison above, we can conclude that the X11 decomposition method is best suited for our dataset. Based on the forecasted values, the demand for energy decreased in 2020 and 2021. We know this is due to the repercussions of the pandemic. Since everyone was sent into quarantine, the use of energy was not nearly as demanded as the previous years due to the fact that people were staying and working from home instead of traveling. Oil is a huge factor in energy consumption, which took a severe hit once everyone was stationary. Some businesses used solar panels as energy to run their businesses. The demand for solar energy decreased as much of the population that traveled into the offices for work was no longer able to. Therefore, companies aren’t going to spend money on electricity, gas, and water bills on keeping a business or building running while no one is occupying it.

References:

1. <https://www.kaggle.com/code/robinteuwens/forecasting-energy-demand/notebook>
2. <https://aep.com/about/businesses/opcos>

**SAS Code:**

proc import out=aep datafile="/home/u62176551/sasuser.v94/aep\_monthly data.xlsx"

dbms=xlsx replace;

run;

proc sgplot data=aep;

series x=date y=aep;

title "Monthly Mean Energy Consumption(in MW)";

xaxis label="Months";

yaxis label="Mean Energy Consumption(in MW)";

run;/\*possible trend,seasonal\*/

proc timeseries data=aep plots=(acf pacf) out=\_null\_;

var aep;

corr acf/nlag=36;

run;/\*confirms seasonal,no trend\*/

/\*WINTERS\*/

/\*Multiplicative winters\*/

proc esm data=aep lead=24 back=24 outfor=outaep plot=forecasts out=\_null\_ print=all;

id date interval=month;

forecast aep/model=winters;

run;

proc sgplot data=outaep;

series x=date y=Actual;

series x=date y=Predict;

title "plot of actual versus forecasted";

run;

/\*Additive winters\*/

proc esm data=aep lead=24 back=24 outfor=outaep1 plot=forecasts out=\_null\_ print=all;

id date interval=month;

forecast aep/model=addwinters;

run;

proc sgplot data=outaep1;

series x=date y=Actual;

series x=date y=Predict;

title "Plot of actual versus forecasted";

run;

/\* deseasonalising data\*/

data aep1;

set aep;

t=\_n\_;

aep\_new=aep;

if t>143 then aep\_new =.;

run;

proc reg data=aep1 outest=aep11new;

model aep\_new=t/AIC BIC;

output out=AEP11\_out r=r\_AEP11 p=p\_AEP11;

run;

data AEP11\_out;

set AEP11\_out;

mape\_fit=(abs(r\_AEP11)/aep\_new)\*100;

if t>143 then mape\_acc=(abs(aep-p\_AEP11)/aep)\*100;

run;

proc means data=AEP11\_out mean;

var mape\_fit mape\_acc;

run;

/\*mape\_fit 7.5691252 and mape\_acc 6.5155743\*/

proc sgplot data=AEP11\_out;

series x=date y=aep;

series x=date y=p\_AEP11;

run;

/\* Obtaining seasonally adjusted data \*/

proc timeseries data=aep1 outdecomp=sa\_aep out=\_null\_;

decomp sa;

id date interval=MONTH;

var aep;

run;

data aep\_combined;

merge sa\_AEP aep1;

si=aep/sa;

run;

proc sgplot data=aep\_combined;

series x=t y=aep;

series x=t y=sa;

run;

proc reg data=aep\_combined outest=aep2new;

model sa=t/aic bic dwprob;

output out=sa\_aepout r=r\_sa p=p\_sa;

run;

data sa\_aepout;

set sa\_aepout;

mape\_fit2=(abs(r\_sa)/sa)\*100;

if t>143 then mape\_acc2=(abs(sa-p\_sa)/sa)\*100;

aep\_reseason=si\*p\_sa;

run;

proc means data=sa\_aepout mean;

var mape\_fit2 mape\_acc2;

run;

/\*mape\_fit2 3.4407620 and mape\_acc2 3.3896240\*/

proc sgplot data=sa\_aepout;

series x=date y=aep;

series x=date y=aep\_reseason;

run;

/\*X11 decomposition\*/

data aep\_x11;

set aep;

t=\_n\_;

aep\_new=aep;

if t>143 then aep\_new =.;

run;

proc x11 data=aep\_x11 noprint outextrap;

monthly date=date;

var aep\_new;

arima forecast=4;

output out=aep\_out a1=original d10=sf d11=sa d12=tcc a15=forecast;

run;

data aep\_x11\_merge;

merge aep\_x11 aep\_out;

run;

data aep\_x11\_merge;

set aep\_x11\_merge;

t=\_n\_;

mape\_fit=(abs(original-forecast)/original)\*100;

if t>143 then mape\_acc=(abs(aep-forecast)/aep)\*100;

run;

proc means data=aep\_x11\_merge mean;

var mape\_fit mape\_acc;

run;

proc sgplot data=aep\_out;

series x=date y=original;

series x=date y=forecast;

series x=date y=sa;

series x=date y=tcc;

run;

proc sgplot data=aep\_out;

series x=date y=original;

series x=date y=forecast;

run;

/\*Regression using dummy variables\*/

data aep1;

set aep;

month=month(date);

if month=1 then m1=1; else m1=0;

if month=2 then m2=1; else m2=0;

if month=3 then m3=1; else m3=0;

if month=4 then m4=1; else m4=0;

if month=5 then m5=1; else m5=0;

if month=6 then m6=1; else m6=0;

if month=7 then m7=1; else m7=0;

if month=9 then m9=1; else m9=0;

if month=10 then m10=1; else m10=0;

if month=11 then m11=1; else m11=0;

if month=12 then m12=1;else m12=0;

run;

proc reg data=aep1 outest=outaepreg;

model aep=m1 m2 m3 m4 m5 m6 m7 m9 m10 m11 m12/aic bic adjrsq vif corrb;

output out=aep1out p=aep\_pred r=aep\_res;

run;

data aep1out;

set aep1out;

mape=(abs(aep\_res)/aep)\*100;

run;

proc means data=aep1out mean;

var mape;

run;

/\*Classical Decomposition Method\*/

proc timeseries data=aep plots=(tc sa cc tcc) outdecomp=aep1 printdetails;

id date interval=month;

var aep;

decomp orig tc sc sa cc ic/mode=additive;

run;

/\*Moving averages\*/

proc expand data=aep out=aep\_ma\_output1;

id date;

convert aep=moving\_average12/transout=(movave 12);

convert aep=moving\_average3/transout=(movave 3);

run;

proc sgplot data=aep\_ma\_output1;

series x=date y=aep;

series x=date y=moving\_average3;

label moving\_average3="MA3";

title "Moving Average 3";

run;

proc sgplot data=aep\_ma\_output1;

series x=date y=aep;

series x=date y=moving\_average12;

label moving\_average12="MA12";

title "Moving Average 12";

run;

data aep\_ma\_output1;

set aep\_ma\_output1;

mape\_MA3=(abs(aep-moving\_average3)/aep)\*100;

mape\_MA12=(abs(aep-moving\_average12)/aep)\*100;

run;

proc means data=aep\_ma\_output1 mean;

var mape\_MA3 mape\_MA12;

run;/\*MAPE 3 is 6.2185027 and MAPE 12 is 7.1552610\*/

/\*Simple exponential smoothing\*/

proc esm data=aep print=all outfor=outaep back=24 lead=24 out=\_null\_ plot=forecasts;

forecast aep/model=simple;

run;

/\*Trying ARIMA with less order\*/

proc arima data=aep;

identify var=aep(1,12) nlag=36 whitenoise=ignoremiss;

estimate p=3 q=(12) whitenoise=ignoremiss;/\*ARIMA (3,1,0)(0,1,1)\*/

estimate p=2 q=(2)(12) whitenoise=ignoremiss;/\*ARIMA (2,1,2)(0,1,1)\*/

\*forecast id=date interval=month lead=24 out=out1;

run;