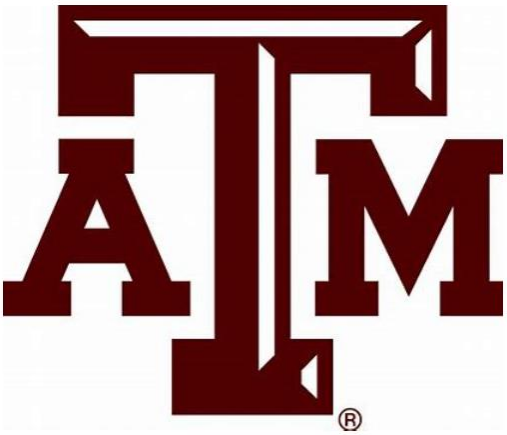
# TIME SERIES ANALYSIS

## ELECTRICITY DEMAND FOR THE STATE OF TEXAS



ANALYSIS & FINAL REPORT BY

AMRENDRA KUMAR – (amrendra93@tamu.edu)
ASHWIN PRASANNA – (aprasanna@tamu.edu)
CHRISTOPHER HAN – (christopherhan@tamu.edu)
JAMES CADENA – (jcadena@tamu.edu)
NARAYANAN VAIDHYANATHAN – (narayananv@tamu.edu)

## OBJECTIVE

The objective of the project is to analyze the electricity demand for the state of Texas during the year 2015 to 2020. As more people and businesses choose Texas as their home, it will be important to accurately forecast energy demands so that generation can be scheduled to meet peak demands while dispatching the cheapest available resources, to relieve grid congestion. With the trove of wind energy usage and other flexible fuel and resource availability, Texas is uniquely positioned to safely and reliably deliver power to its inhabitants. Hence, the aim is to accurately forecast the future electricity demand for the state of Texas. We plan to fit several time series models to the data and come up with the best model that has the highest accuracy in predicting future demand.

## DATA EXPLORATION

The data is taken from the U.S. Energy Information Administration and contains two variables. First is time, which consists of the date and hour at which the energy consumption is measured. Second is the energy consumption in Megawatt Hours (MWh) for the given hour on a date. The timeline for the analysis is July 2015 to the end of May 2020. Initially the aim of the team was to carry out the analysis from an hourly perspective. However, given the enormous nature of the hourly data for 4 years, the time series plot appeared to be too noisy.

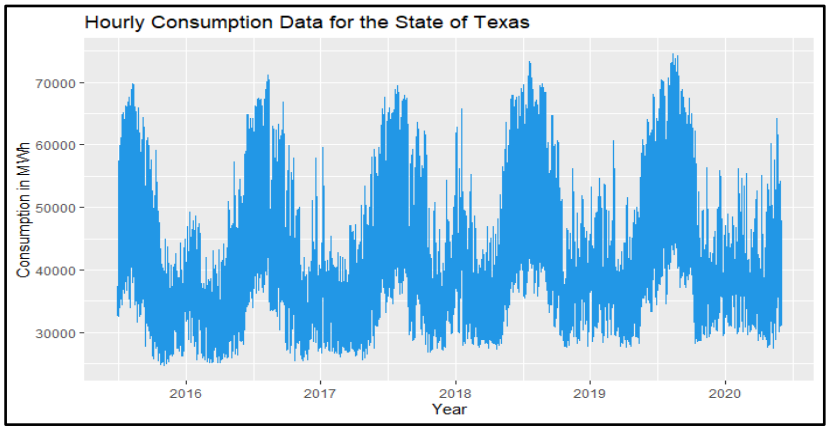



*Figure 1: Plot of Hourly Consumption Over the years*

From the above hourly plot, we can infer that the hourly consumption over the years is quite noisy and extraction of information from the plot can be difficult. However, we can infer that there is a periodicity of 1 year which can be observed. Apart from the seasonality, the evidence of trend is not quite clear from the plot.

Hence, the team decided to use the average monthly demand to conduct the time series study. This in effect would reduce the noise in the data since there is one point for the entire month as opposed to 24 points for an entire day multiplied by 30 for the entire month. The team also noticed that random time points are missing in the raw data file; the averaging effect helps ease this issue since individual hourly data do not hold weight over the average monthly data. Ultimately, we hope to achieve more accurate forecasting of future data because of better data quality and reduction in outliers or missing data that could possibly thwart modeling and parameter estimation.

For averaging the data over a month, the hourly data is averaged over the day to find out the daily average and then the daily average is averaged over the month to obtain the monthly average equating to getting a single estimate of the average hourly consumption for the month. Hence, all the data

analysis and future prediction would be done for average hourly consumption of electricity in each month in the State of Texas during the 2015 to 2020 timeframe.

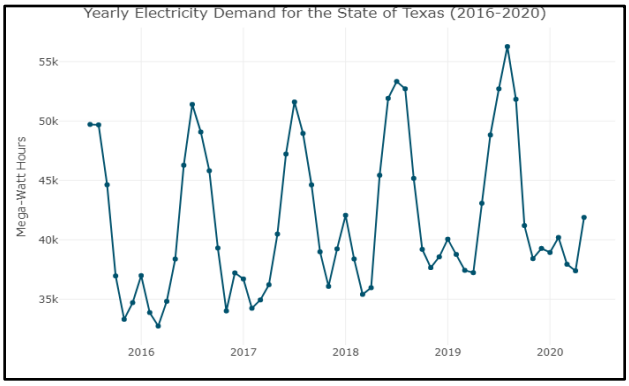



*Figure 3: Seasonality Plot of Aggregated Data*

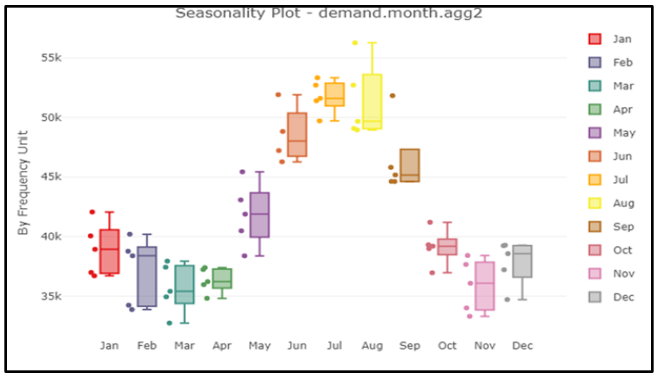



*Figure 2: Hourly Consumption Average Over Month*

The above plots show the hourly consumption data averaged for the given month and plotted over the 5 years. From the plot, we can infer that there is a clear visibility of seasonality in the data. Also, there is evidence of a steady increasing trend in power consumption over the years. From the box plots, we can infer that the peak consumption of electricity occurs during the month of July and August, whereas the lowest consumption of electricity occurs during March and November.

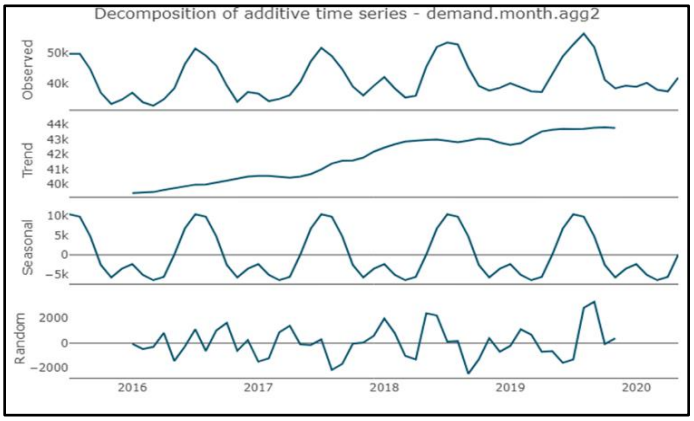



*Figure 4: Decomposition of Time Series*

To get a clearer picture of the components in the time series, the data was decomposed to identify the trend, seasonality and the randomness in the data. From the decomposition, we found out that there is a significant positive trend in the electricity consumption indication accompanied with seasonality and randomness.

## ANALYSIS

As a part of analysis, the objective of the team is to transform the data to a stationary series and estimate the Autoregressive and Moving Average Components. The technique used for data transformation is seasonal differencing of order 1.

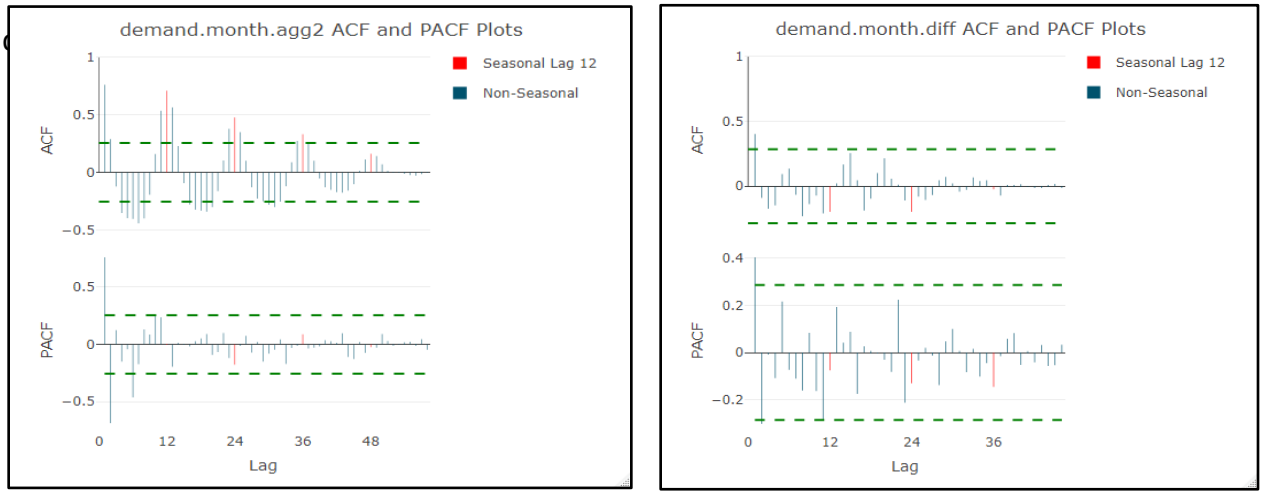
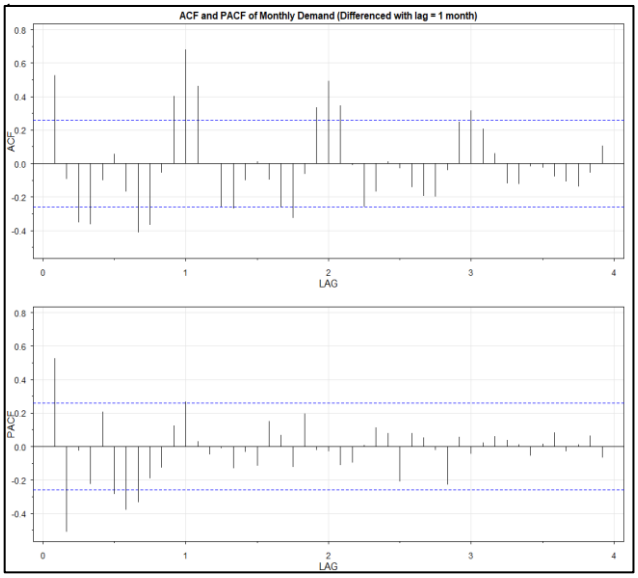
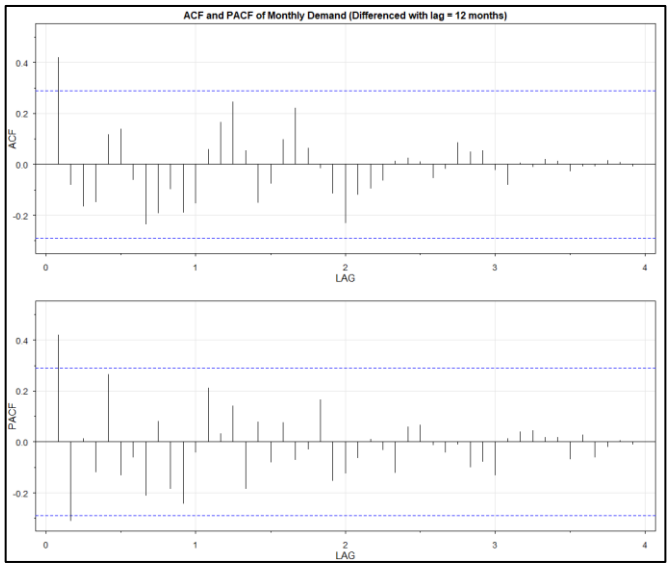



*Figure 5: Comparison of ACF And PACF for Raw and Differenced Data*

The left plot shows the ACF and PACF for the raw monthly aggregated data. There is pronounced seasonality that can be observed from the raw data. The plot on the right shows that ACF and PACF of the differenced data, that appears to exhibit mostly white noise like behavior. Also, we wanted to confirm the impact of using lag 1 and lag 12 differencing of the time series data.




From the above plots we can see that the ACF and PACF are higher for increasing lag for lag 1 differencing (above left plot) as compared to lag 12 differencing (above right plot). For lag 12 differencing, the PACF function after lag 2 behaves as white noise as can be seen from the above plots.

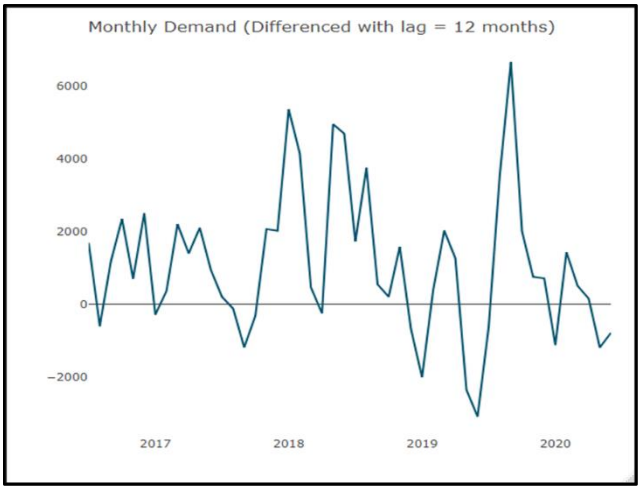



*Figure 7: Plot of Lag 12 Differenced Monthly Data*

**PERFORMING ADF TEST TO UNDERSTAND STATIONARITY:**

p-value smaller than printed p-value

```
        Augmented Dickey-Fuller Test

data:  diff_12

Dickey-Fuller = -4.1876, Lag order = 0, p-value
= 0.01

alternative hypothesis: stationary
```

From the above plot, we can confirm that the Lag 12 differenced Monthly data is stationary. This can also be verified by the Augmented Dickey Fuller (ADF) test, the result of which is shown above. The null hypothesis of the ADF test is that the series has a unit root, whereas the alternate hypothesis is one of

stationarity. Since the p-value = 0.01 < .05, we can reject the null hypothesis and concluded that there is significant evidence that the series exhibits stationarity.

## COVARIATES

The team decided to explore possible covariates for the monthly energy demand. A significant predictor of demand would naturally be the temperature during the day. However, this data is not readily available and on further exploration the team decided to explore the Cooling Degree Days (CDD) and Heating Degree Days (HDD) as the possible covariates. As defined by the U.S. Energy Information Administration (EIA) a degree day measures how cold or warm a location is. A *degree day* compares the mean (the average of the high and low) outdoor temperatures recorded for a location to a standard temperature, usually 65° Fahrenheit (F) in the United States. The more extreme the outside temperature, the higher the number of degree days. A high number of degree days generally results in higher levels of energy use for space heating or cooling [1].

***Heating degree days*** (HDD) are a measure of how cold the temperature was on a given day or during a period of days. For example, a day with a mean temperature of 40°F has 25 HDD. Two such cold days in a row have a total of 50 HDD for the two-day period.

***Cooling degree days*** (CDD) are a measure of how hot the temperature was on a given day or during a period of days. A day with a mean temperature of 80°F has 15 CDD. If the next day has a mean temperature of 83°F, it has 18 CDD. The total CDD for the two days is 33 CDD [1].

It is noted that many of the missing values in the HDD and CDD dataset for the months of April, May, June 2020, were imputed using the historical mean for that specific month in the previous 4 years (2015-2019).

Shown the model diagnosis here.

```
Call:
lm(formula = demand.monthly.agg2 ~ hdd$Value + cdd$Value)

Residuals:
     Min       1Q   Median       3Q      Max
-13202.7  -1222.2    499.6   1641.9   4126.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30864.477    976.144  31.619  < 2e-16 ***
hdd$Value      14.976      2.718   5.509 9.02e-07 ***
cdd$Value      35.265      2.345  15.039  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2472 on 57 degrees of freedom
Multiple R-squared:  0.8495,    Adjusted R-squared:  0.8442
F-statistic: 160.9 on 2 and 57 DF,  p-value: < 2.2e-16
```

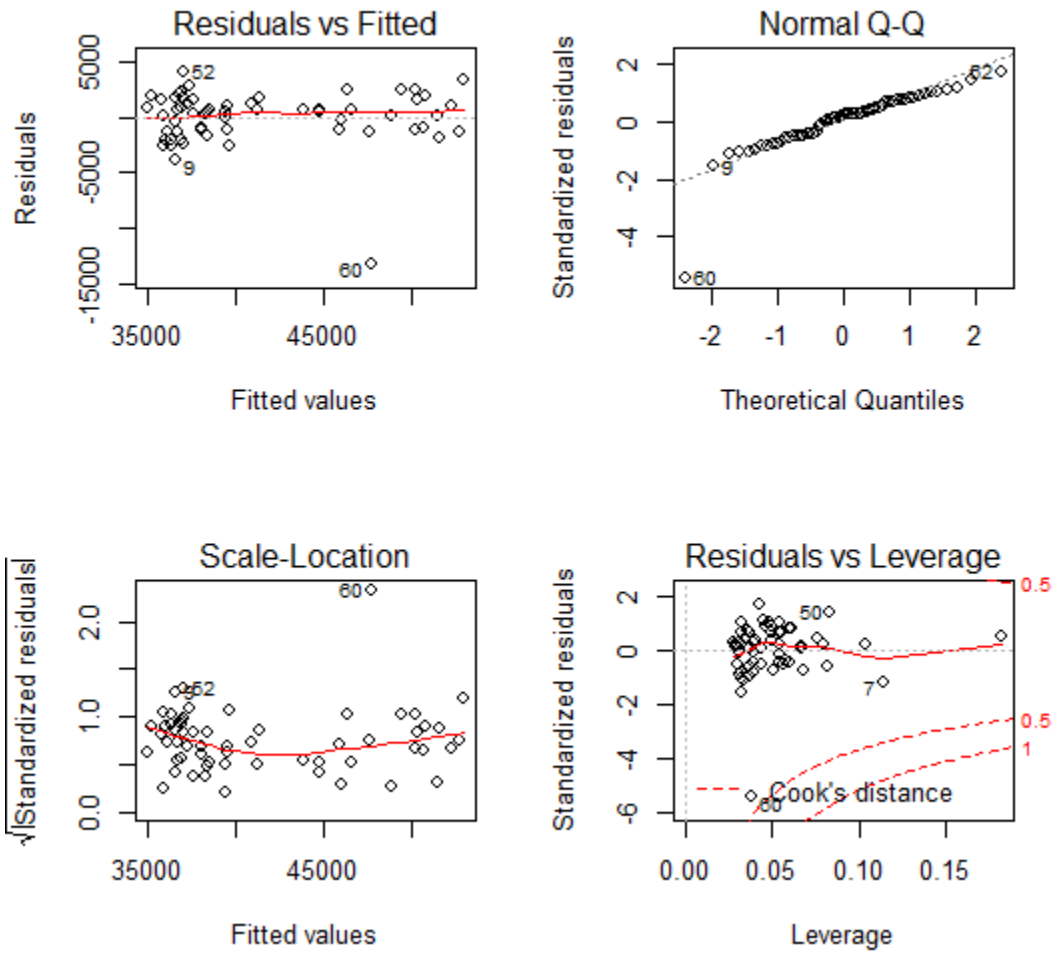



*Figure 8: Model Diagnostics for Linear Model using CDD, HDD as covariates*

The Wald statistics for both the predictors appear to be significant at the 5% level. The adjusted R^2 values of 0.8442 suggest that most of the variation in our model is explained by the independent variables of HDD and CDD. The plots in Figure 8, also show that the residuals are fairly normally distributed with constant variance and no significant outliers that are influential (except point #60). We also explored the possibility of both the independent variables being highly correlated to one another, a phenomenon commonly called Multicollinearity. We quantified this possibility by calculating the VIF (Variance Inflation Factor), commonly use to explain how well one variable is explained by other independent variables. The calculated VIF for the HDD and CDD covariate is 2.573, does indicate moderate correlation. In general, a VIF above 10 indicates high correlation and is a cause of concern [2].

## TRAIN/TESTING DATASETS AND MODELS

The entire dataset was split into training data for model fitting and estimation and testing data set for forecasting and evaluating model performance. The training dataset encompasses the time frame, July 2015 to May 2019. The testing/forecasting dataset is for the timeframe of June 2019 to May 2020.

ARIMA model here is defined as ARIMA($p,d,q$)($P,D,Q$)[12], where the upper case P, D, Q are the seasonal autoregressive, differencing, and moving average orders and their lowercase counterparts are the non-seasonal components. . The following seasonal ARIMA models were considered for review.

1. (2,0,0)(0,1,1)[12]
2. (0,0,1)(0,1,1)[12]
3. (0,1,2)(1,1,0)[12]
4. (2,0,0)(0,1,1)[12] with covariates
5. (1,0,0)(0,1,1)[12] with covariates
6. (1,1,0)(1,0,0)[12] with covariates

## MODELING

The potential models chosen for analysis were based on the lowest AIC (Akaike Information Criterion) values and percentage error using the MAPE (Mean Absolute Percentage Error) statistic during forecast. The MAPE is unitless and used frequently to compare forecast performances between data sets. The table below outlines specifics of the model performance.

| MODEL | COVARIATES | AIC | BIC | MAPE |
|---|---|---|---|---|
| 1. (2,0,0)(0,1,1)12 | - | 13.61799 | 13.78645 | 3.921629 |
| 2. (0,0,1)(0,1,1)12 | - | 13.52122 | 13.65598 | 3.76173 |
| 3. (0,1,2)(1,1,0)12 | - | 13.57621 | 13.71146 | 4.253591 |
| 4. (2,0,0)(0,1,1)12 | CDD, HDD | 12.41394 | 12.61609 | 1.845633 |
| 5. (1,0,0)(0,1,1)12 | CDD, HDD | 12.53126 | 12.69972 | 2.250063 |
| 6. (1,1,0)(1,0,0)[12] | CDD, HDD | 16.43198 | 16.62881 | 2.188582 |

Based on the above table, the 4$^{th}$ model: ARIMA (2,0,0)(0,1,1)[12] with the CDD and HDD values as covariates, appears to be the best if selected based on AIC and MAPE values. The AIC value of 12.4 as well as lowest BIC value of 12.6, with the lowest MAPE of 1.84. However, residual analysis (detailed in the next section) does show deviations from normality for the residuals as well as significant autocorrelation of lags.

**RESIDUAL ANALYSIS & FORECAST PERFORMANCE**

ARIMA (2,0,0)(0,1,1)12 with covariates

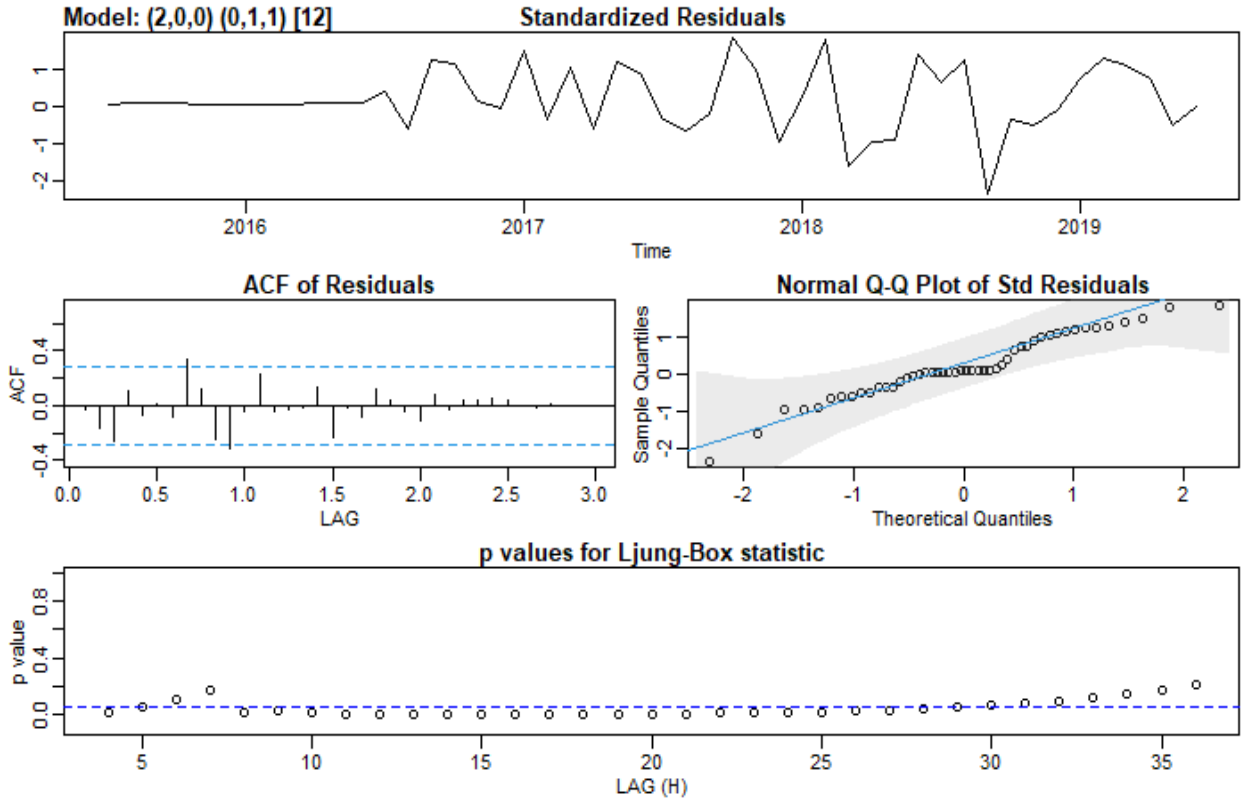



*Figure 9: Residual Analysis for **ARIMA (2,0,0)(0,1,1)12***

An alternate model (model #5) with slightly higher AIC values (12.5) and MAPE (2.25) is the ARIMA (1,0,0)(0,1,1)[12] model also shows significant deviations from normality as well as auto-correlated lags. A better model to consider would be last one (model #6, ARIMA (1,1,0)(1,0,0)[12]). This model has the highest AIC value of 16.4, however, the MAPE is controlled fairly well at 2.19%. Lewis classified models as "best" if the MAPE less than 10%, "good" with MAPE range between 10% - 20%, "acceptable" for MAPE 20% - 50% and "false" if MAPE more than 50% [3].

The plot of the residuals (Figure 10) indicate that there are no significant lags and the residuals follow normality assumptions the best. The slightly higher AIC value could be attributed to addition of the covariates, that add a penalty to the model complexity. An added benefit of this model is that the parameter estimates are all significant at the 5% level using the Wald test statistic (as seen below). Most of the p-values for the Q-statistic is above the .05 level, effectively failing to reject the null hypothesis (null hypothesis states that the autocorrelations up to lag k equal 0. i.e. the data values are random and independent up to a certain number of lags or white noise) as seen in Figure 10.

```
Coefficients:
          ar1    sar1  cdd_train  hdd_train
      -0.4323  0.7525    36.2717     17.762
s.e.   0.1338  0.1598     1.5657      2.277

sigma^2 estimated as 521088:  log likelihood = -381.15,  aic = 772.3

$degrees_of_freedom
[1] 43

$ttable
          Estimate     SE t.value p.value
ar1        -0.4323 0.1338 -3.2305  0.0024
sar1        0.7525 0.1598  4.7094  0.0000
cdd_train  36.2717 1.5657 23.1660  0.0000
hdd_train  17.7620 2.2770  7.8008  0.0000
```

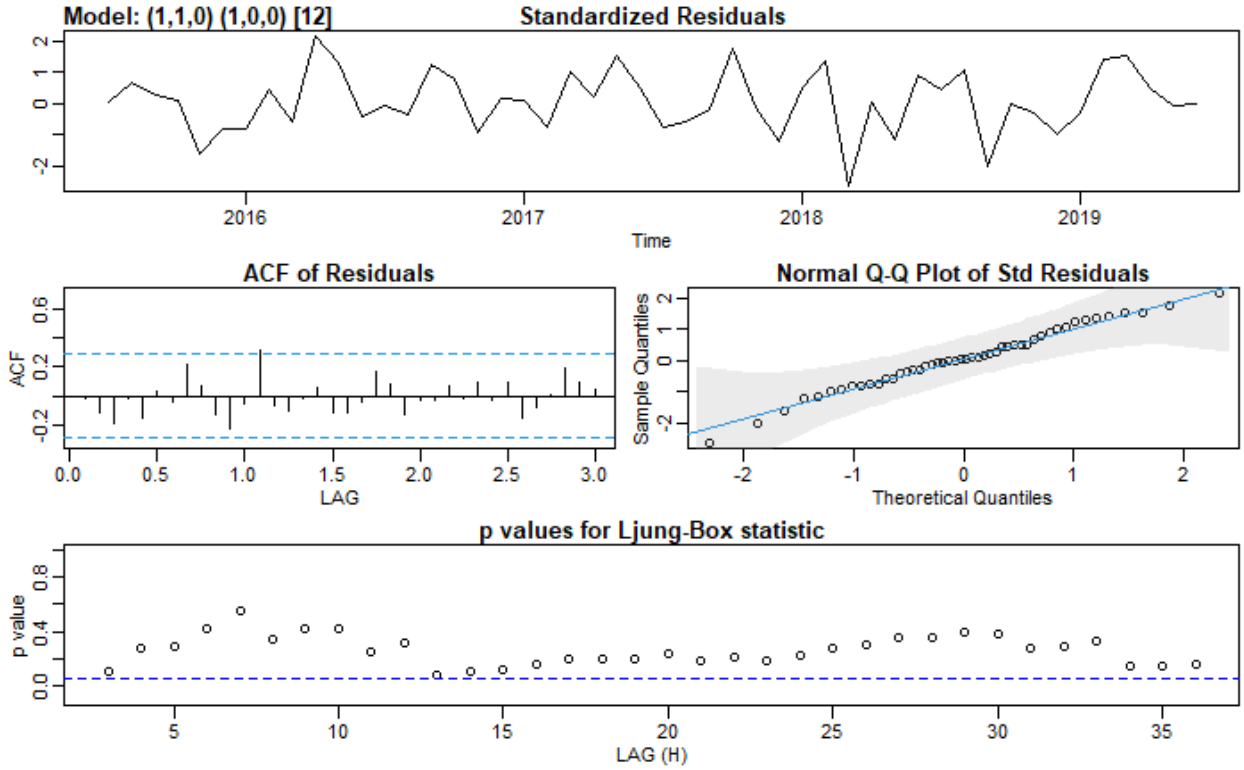



*Figure 10: Residual Analysis selected model: ARIMA (1,1,0)(1,0,0)[12]*

For a model without the covariates included, the ARIMA (2,0,0)(0,1,1)[12] model (1) without covariates. The AIC and MAPE values for this model are 13.6 and 3.92 respectively. The residuals of the chosen model are fairly normal with a slightly heavier tail than the model picked earlier. However, none of the parameter estimates are significant at the 5% level and the MAPE is the 2nd highest among all the models.

Model #2 ARIMA (0,0,1)(0,1,1)12, is slightly better than the prior and with residuals adhering to normality and without the presence of autocorrelation lags. However, the seasonal MA estimate is not significant at the 5% level, leading us to keep this model in our periphery but cautiously be more supportive of model #6. The final model (model #3) is very similar to model #2 but with slightly worse MAPE at 4.25% and residuals showing heavier tails.

The forecasted demand values along with the MAPE from the R output for chosen model (model #6):

```
$pred
          Jan       Feb       Mar       Apr       May       Jun       Jul       Aug
Sep       Oct       Nov       Dec
2019                                                      52294.79 55769.08
51147.51 40159.77 38661.72 39286.40
2020 40135.38 40283.51 39206.88 38595.44 43938.78 50977.40

$se
           Jan       Feb       Mar       Apr       May       Jun       Jul       Aug
Sep       Oct       Nov       Dec
2019                                                        721.8679  830.0831
992.8481 1105.5889 1218.2222 1317.1073
2020 1410.7509 1497.8685 1580.4712 1658.8503 1733.7375 1805.5013

[1] "The MAPE IS"
[1] 2.188582
```

As seen in the below chart (Figure 11) the forecast data points are quite close to the actual values for the timeframe in both the training and testing horizon.

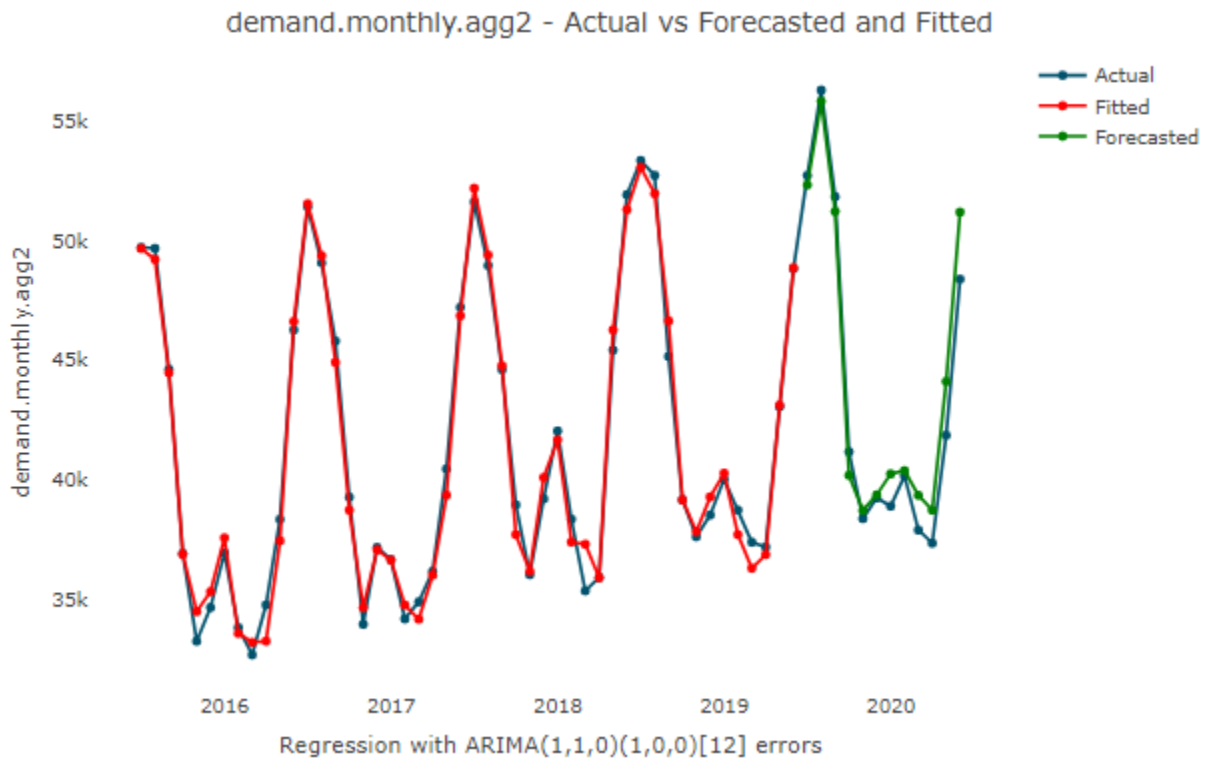



*Figure 11: Model Performance _ ARIMA (1,1,0)(1,0,0)[12]*

## CONCLUSION

In this project we have explored various methods of fitting a time series model to data that exhibit a seasonal behavior over time. We were able to successfully analyze the electricity demand for the state of Texas during the year 2015 to 2019 and forecast demand for the 2020 horizon. We have also learned various methods, tools, and statistical tests that can be used to verify assumptions and prediction accuracy. As the population of the state of Texas grows by leaps and bounds, it will be utmost necessary to forecast demand to plan for projects that address electric grid congestion and develop new generation to meet the demands of the residents. While our modeling and analysis does account for covariates and seasonality, there is room for improvement. For example, regarding the COVID-19 pandemic, most businesses that employ technical workers remain closed while employees work from home. Undoubtedly, this causes a shift in demand for energy and supply as well. Our modelling does not take the pandemic response into consideration and is an area that we could explore to more accurately forecast values for events that are unexpected and extraordinary.

**REFERENCES**

[1] Degree-days - U.S. Energy Information Administration (EIA). (2020). Retrieved 26 July 2020, from https://www.eia.gov/energyexplained/units-and-calculators/degree-days.php#:~:text=Cooling%20degree%20days%20(CDD)%20are,F%2C%20it%20has%2018%20CDD.

[2] Kutner, M. H.; Nachtsheim, C. J.; Neter, J. (2004). Applied Linear Regression Models (4th ed.). McGraw-Hill Irwin.

[3] C.D. Lewis. Industrial and Business Forecasting Methods. Butterworths. 2,194-196, (1982).