ASSIGNMENT 3

1. PROCESS

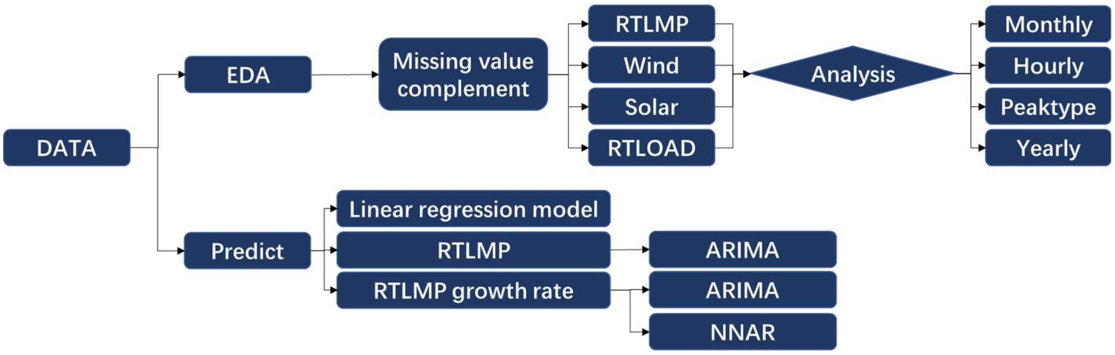


figure 8: flowchart

2. EDA

2.1 Missing Value

After summarizing all the data, we found that there are missing values in ERCOT.(WIND_RTI) and ERCOT.(GENERATION_SOLAR_RT) columns.

Then we plot the time series plot of these two columns.

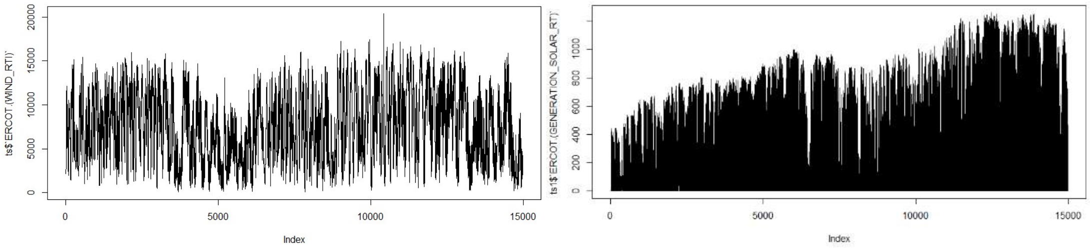


figure 9: Time series plot of Wind and Solar. The left figure is WIND_RTI and the right one is SOLAR_RT.

The WIND_RTI shows a random fluctuation, and the SOLAR_RT shows a rising trend with fluctuation, we can consider using the average of the near value of the missing value to fill them.

After filling missing values, we get the summary of the data as follows:

table 1: summary of data

Data	Min	1st	Median	Mean	3rd	Max
RTLMP	-17.86	18.04	20.06	25.77	25.03	2809.36
WIND	54.44	4135.72	7281.52	7532.73	10852.41	20350.40
SOLAR	0.00	0.00	22.07	291.92	608.58	1257.54

Data	Min	1st	Median	Mean	3rd	Max
RTLOAD	25567	35432	39934	42372	47873	73265

2.2 RTLMP Analysis

i. Monthly

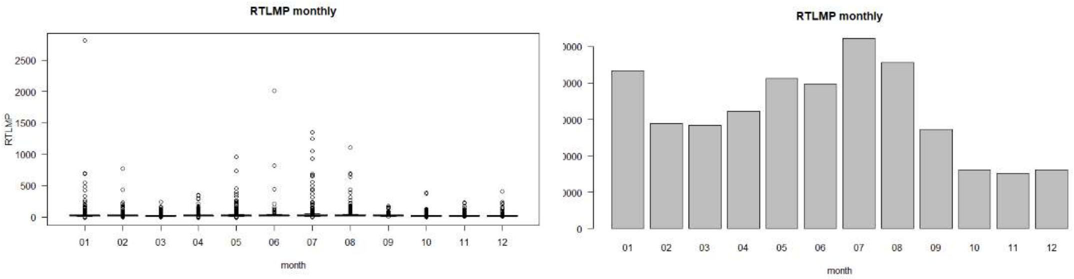


figure 10: boxplot and bar plot of RTLMP monthly

From the boxplot, we found that each month has different degrees of outliers, among which Jan, Jun, Jul outliers deviate greatly. And from the bar plot, RTLMP showed a trend of first decreasing, then increasing and then decreasing throughout the year, and reached its peak in July.

ii. Hourly

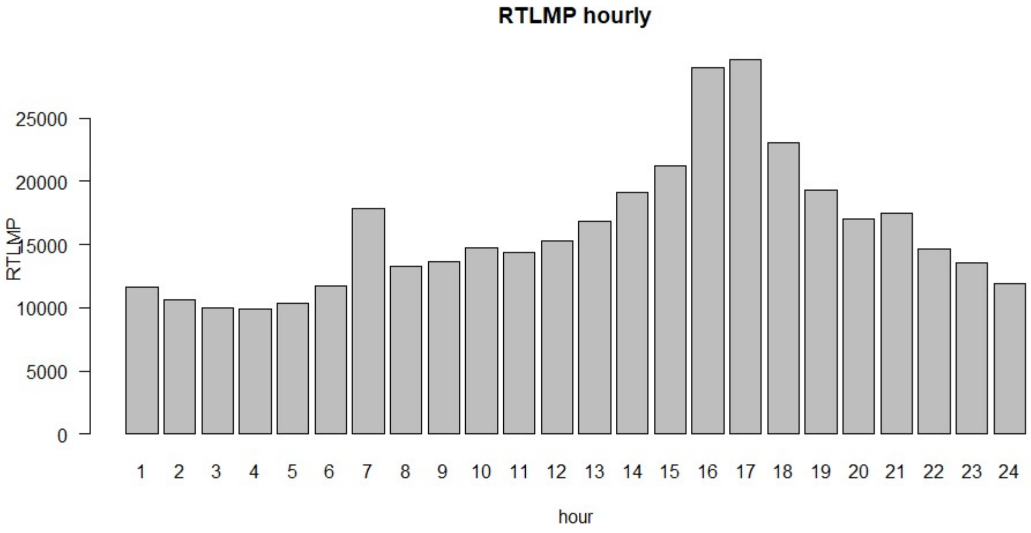


figure 11: RTLMP hourly distribution

The price is the highest in the afternoon (16-17), showing a trend of first rising and then falling as a whole within a day.

iii. Peaktype

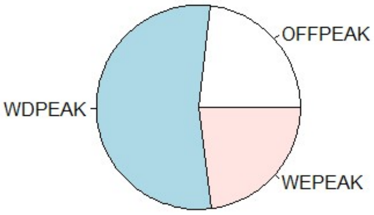


figure 12: pie chart for RTLMP in different peaktype

iv. Yearly

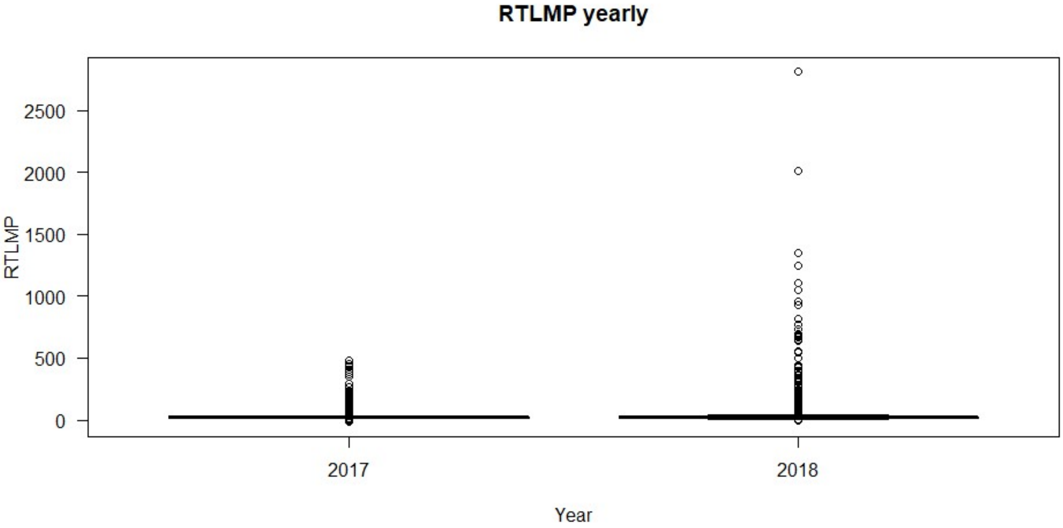


figure 13: boxplot for RTLMP yearly

Outliers are more pronounced during 2018. The average RTLMP of each year looks similar.

2.3 WIND Analysis

i. Monthly

There are no outliers, and the mean WIND_RTI for each month shows a certain periodicity. January to June is an upward arc, July-December is a downward arc.

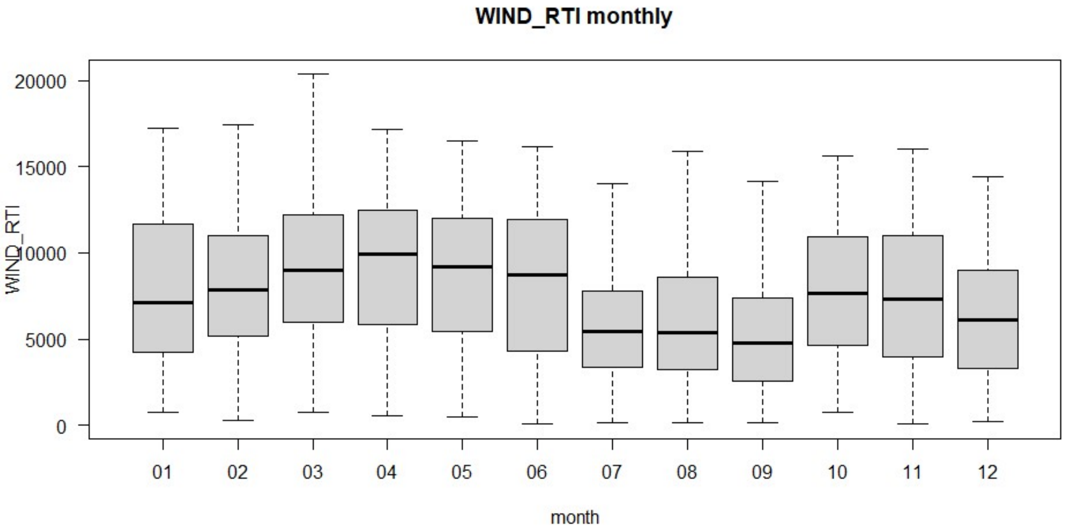


figure 14: boxplot for WIND_RTI monthly

ii. Hourly

Wind generation are higher in the morning and at night, with lower wind generation at noon.

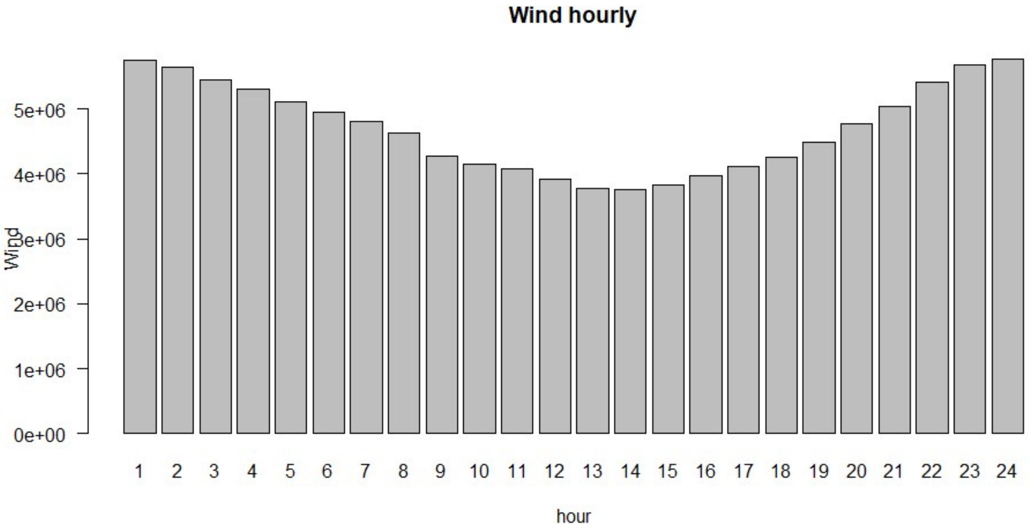


figure 15: bar plot for WIND_RTI houly

iii. Peaktype

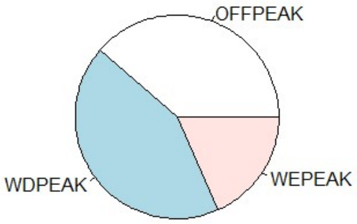


figure 16: pie plot for WIND_RTI in different peaktype

iv. Yearly

The mean of wind generation in 2018 is higher than that of 2017.

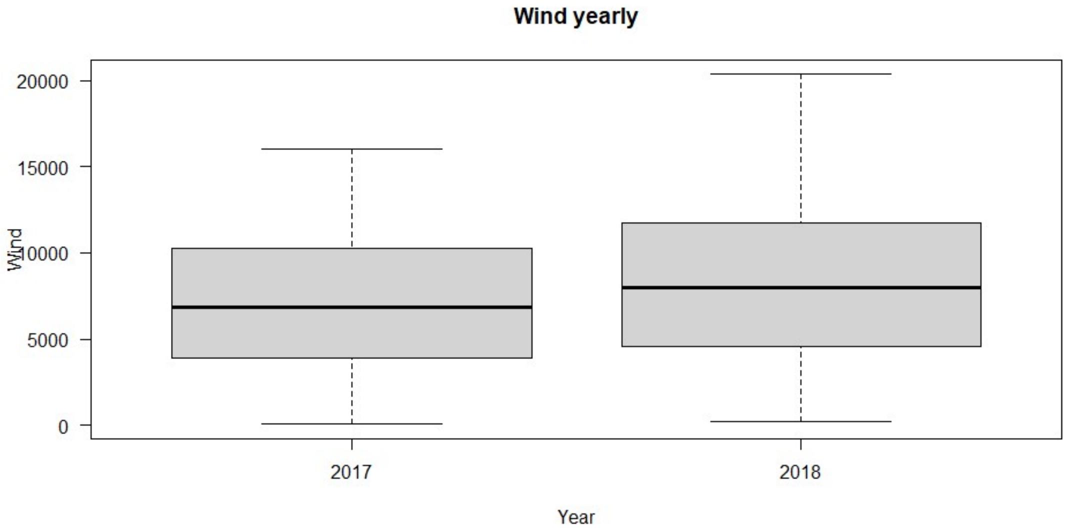


figure 17: boxplot for WIND_RTI yearly

2.4 SOLAR Analysis

i. Monthly

Outliers are more pronounced in January and December. April to September is the month with strong solar generation, and the mean of solar generation is almost 0 in other months.

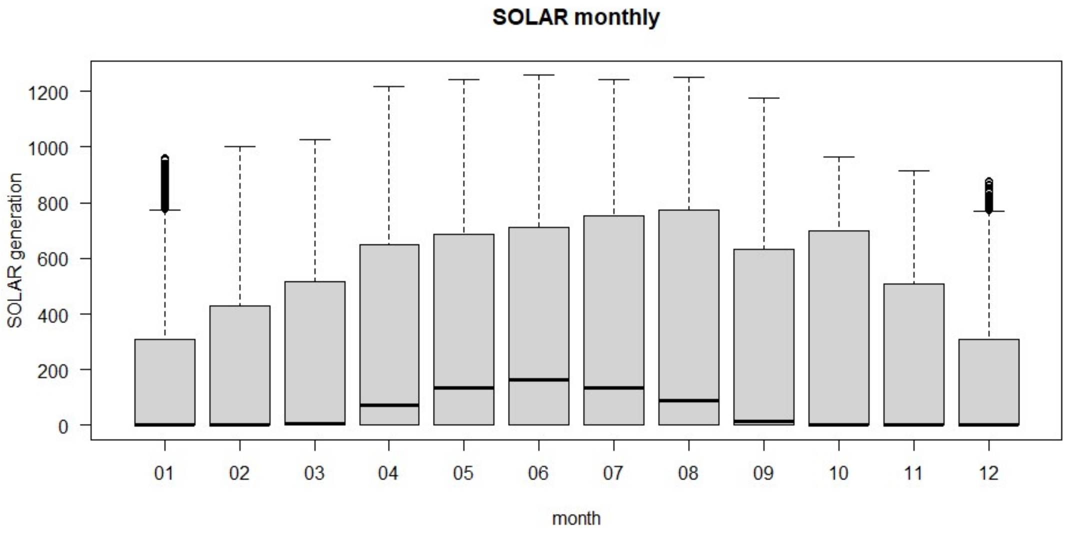


figure 18: boxplot for SOLAR_RT monthly

ii. Hourly

8am to 9pm is the period of solar generation. During this period, the solar generation shows a trend of rising first, reaching the highest peak at 2pm, and then decreasing. It is consistent with the intensity of the sun we perceive in our daily life.

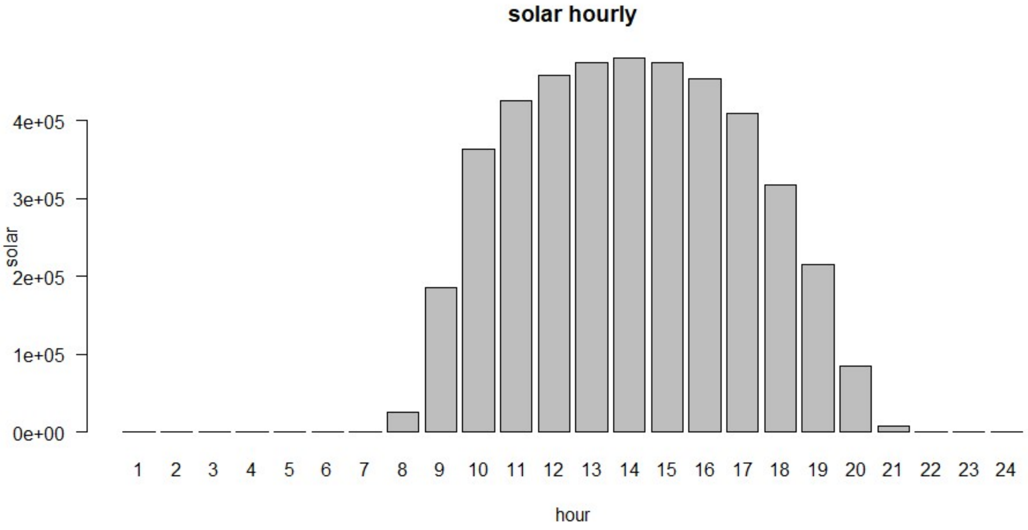


figure 19: bar plot for SOLAR_RT houly

iii. Peaktype

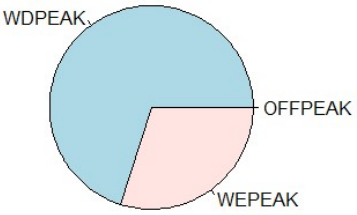


figure 20: pie plot for SOLAR_RT in different peaktype

iv. Yearly

The solar generation in 2018 is higher than that of 2017.

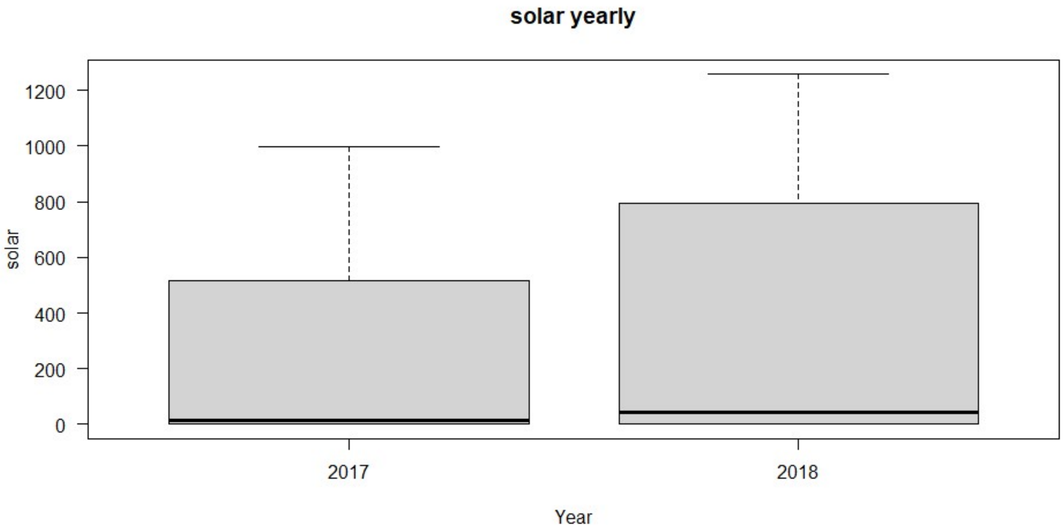


figure 21: box plot for SOLAR_RT yearly

2.5 RTLOAD Analysis

i. Monthly

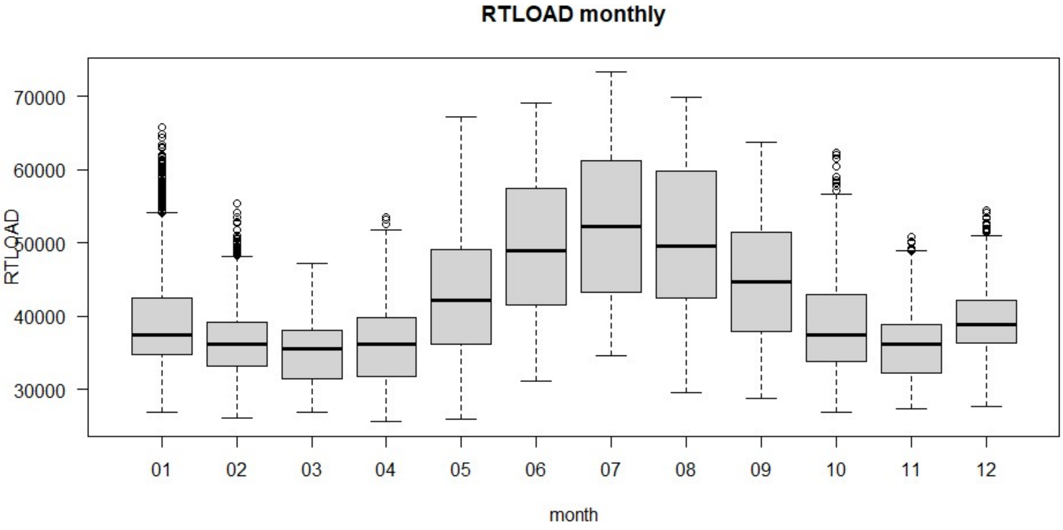


figure 22: boxplot for RTLOAD monthly

RTLOAD has a clear upward arch during May to September, but there are no outliers during this period, and the other months are in a lower position, but there are more outliers. This phenomenon is caused by the generation of a lot of solar energy in May-September.

ii. Hourly

RTLOAD has a clear upward arch during afternoon. I think it is also caused by more solar energy during this period.

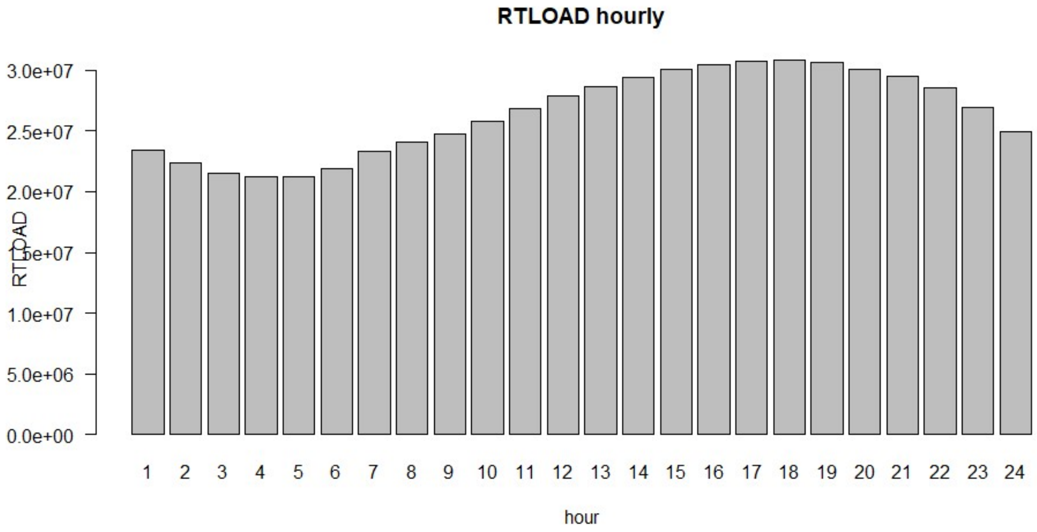


figure 23: bar plot for RTLOAD hourly

iii. Peaktype

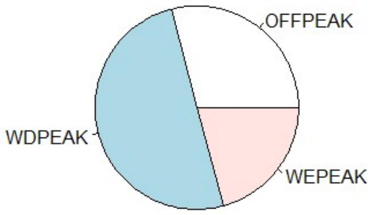


figure 24: pie plot for RTLOAD in different peaktype

iv. Yearly

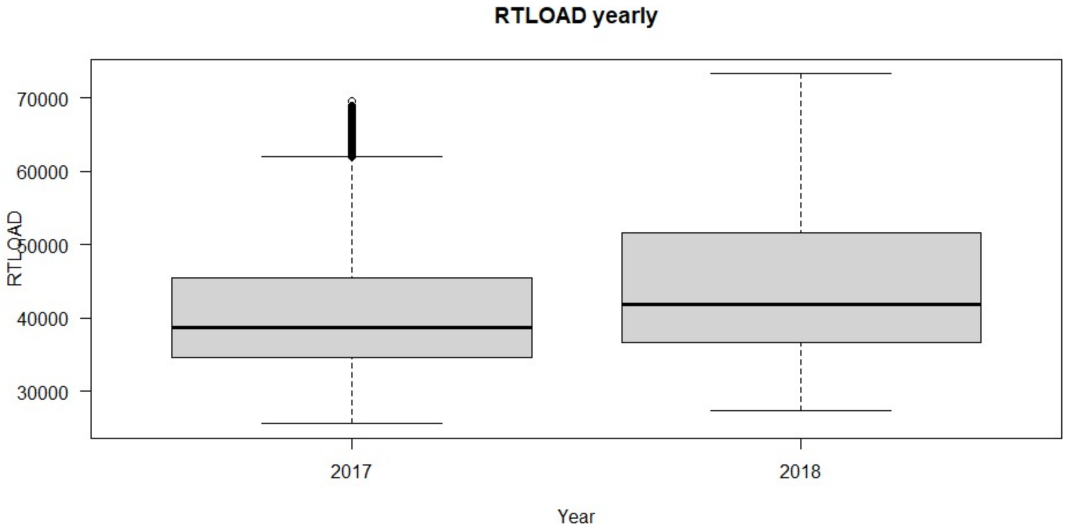


figure 25: boxplot for RTLOAD yearly

The mean of RTLOAD in 2018 is higher than that of 2017. And 2017 contains more outliers. Our statistical dataset found that there were 8760 data in 2017 and 6227 data in 2018. In 2018, the statistics are only up to September, and the entire energy production will decrease after September, which will lower the average value. Therefore, the lack of statistics for the whole year is also the reason for the higher average in 2018.

2.6 Correlation

We examine correlations between data.

table 2: correlation table

	RTLMP	WIND	SOLAR	RTLOAD	HOUR	Peaktype
RTLMP	1	-0.1512	0.1512	0.2385	0.0081	0.0050
WIND	-0.1512	1	-0.2355	-0.1671	-0.360	-0.1958
SOLAR	0.1512	-0.2355	1	0.4664	0.1791	0.0571
RTLOAD	0.2385	-0.1671	0.4664	1	0.4245	0.2252
HOUR	0.0881	-0.0360	0.1791	0.4246	1	-0.0010
Peaktype	0.0050	-0.1958	0.0571	0.2252	-0.0010	1

SOLAR and RTLOAD, SOLAR and class of peaktype are weakly related. The correlation between other data is not obvious.

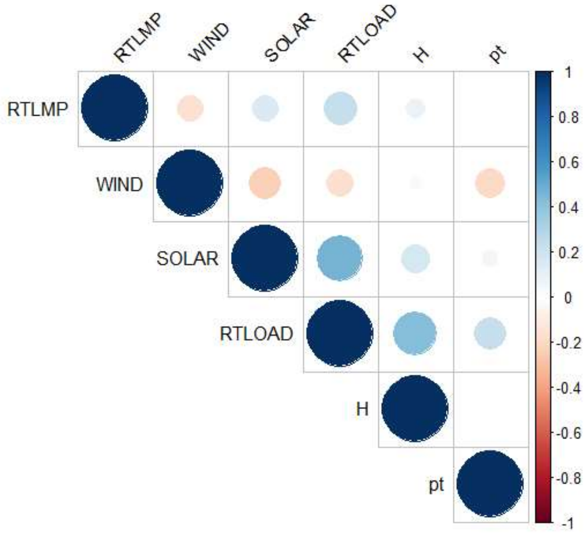
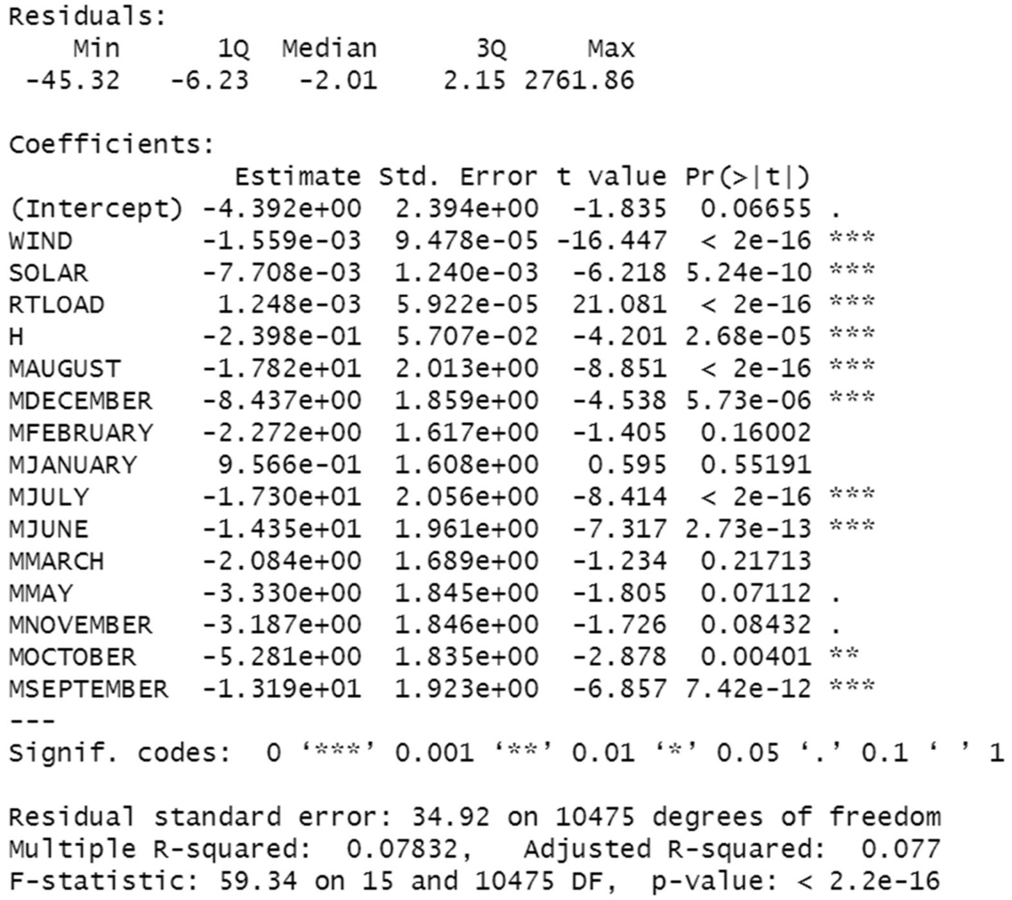


Figure 26: correlation plot of data

3. PREDICT

3.1 Linear Regression Model

Firstly, we used the first 70% of the data as the training set and the last 30% as the test set. We then perform a stepwise regression of RTLMP with all the data, and finally we choose the model with the lowest AIC value. The model results are as follows:



```
Residuals:
    Min      1Q  Median      3Q     Max 
 -45.32   -6.23   -2.01    2.15 2761.86 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.392e+00  2.394e+00  -1.835  0.06655 .  
WIND        -1.559e-03  9.478e-05 -16.447  < 2e-16 ***
SOLAR       -7.708e-03  1.240e-03  -6.218 5.24e-10 ***
RTLOAD       1.248e-03  5.922e-05  21.081  < 2e-16 ***
H           -2.398e-01  5.707e-02  -4.201 2.68e-05 ***
MAUGUST     -1.782e+01  2.013e+00  -8.851  < 2e-16 ***
MDECEMBER   -8.437e+00  1.859e+00  -4.538 5.73e-06 ***
MFEBRUARY   -2.272e+00  1.617e+00  -1.405  0.16002    
MJANUARY     9.566e-01  1.608e+00   0.595  0.55191    
MJULY       -1.730e+01  2.056e+00  -8.414  < 2e-16 ***
MJUNE       -1.435e+01  1.961e+00  -7.317 2.73e-13 ***
MMARCH      -2.084e+00  1.689e+00  -1.234  0.21713    
MMAY        -3.330e+00  1.845e+00  -1.805  0.07112 .  
MNOVEMBER   -3.187e+00  1.846e+00  -1.726  0.08432 .  
MOCTOBER    -5.281e+00  1.835e+00  -2.878  0.00401 ** 
MSEPTEMBER  -1.319e+01  1.923e+00  -6.857 7.42e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.92 on 10475 degrees of freedom
Multiple R-squared:  0.07832,	Adjusted R-squared:  0.077 
F-statistic: 59.34 on 15 and 10475 DF,  p-value: < 2.2e-16
```

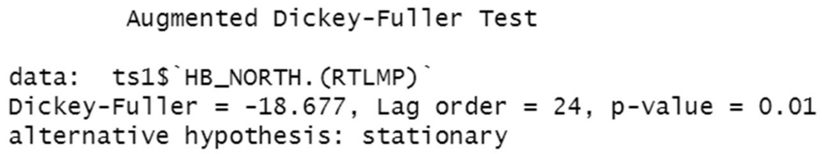
figure 27: linear regression model results

Even though the p-value is less than 0.05 which indicate that the model is significant, it can be seen that the linear regression model performs poorly, because its R-squared is 7.832%.

3.2 RTLMP Time Series

i. Stationary

The Augmented Dickey–Fuller test (ADF) is used to test whether the sample data in time sequence analysis has a single root. The null hypothesis is that a unit root is present in a time series sample. And, the augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.



```
        Augmented Dickey-Fuller Test

data:  ts1$`HB_NORTH.(RTLMP)`
Dickey-Fuller = -18.677, Lag order = 24, p-value = 0.01
alternative hypothesis: stationary
```

figure 28: ADF test results for RTLMP

The p-value of ADF test is less than 0.01, so we can conclude that the RTLMP time series data is stationary.

ii. Outliers

In the EDA analysis, we found that there are many outliers in RTLMP, so we deal with the outliers.

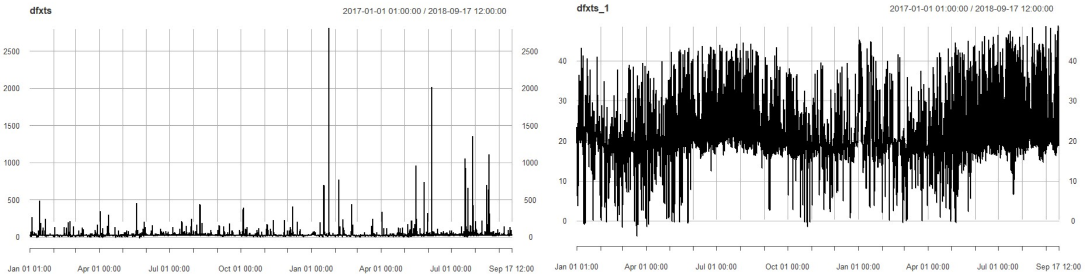


figure 29: The left picture is the RTLMP time series plot without processing outliers, and the right picture is after processing outliers.

iii. ARIMA Model with first-order difference

After dealing with outliers we plot the time series autocorrelation plot.

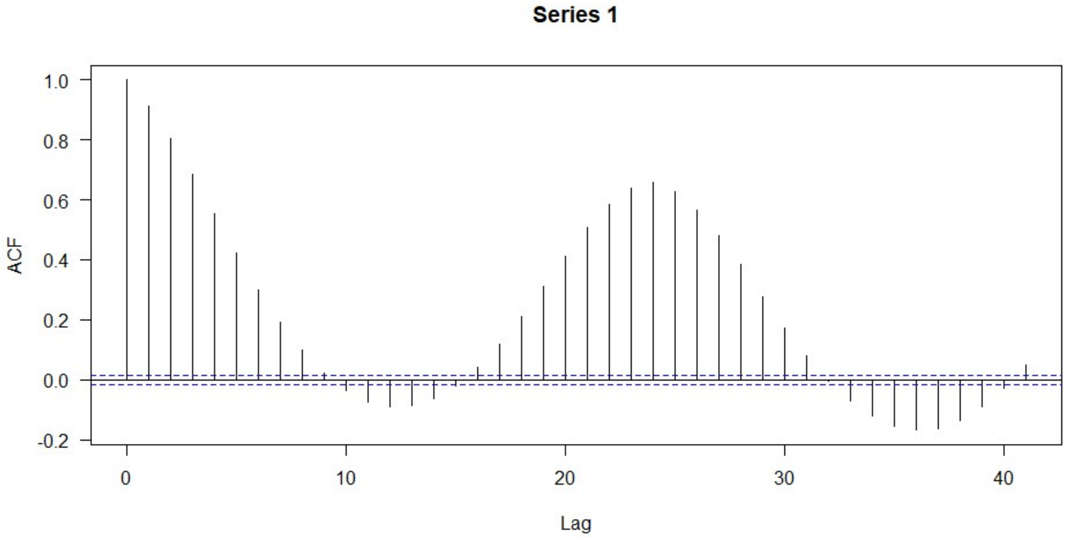


figure 30: ACF plot of RTLMP

Although the time series is stable, there is a certain periodicity. So, we perform first-order differential on the data, and get the diagram.

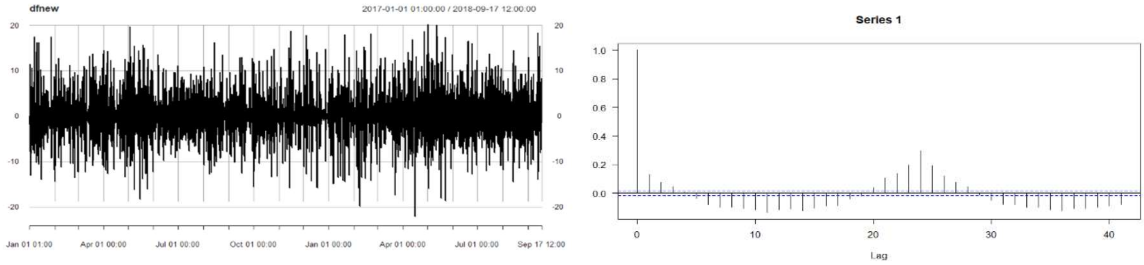


figure 31: time series and ACF plot of data after differential

Then, using the training set data, we build the ARIMA model for the time sequence of RTLMP. The formula of $ARIMA(p,q,d)$ shows as follows:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=(1+\sum_{i=1}^{q}\theta_i\, L^i)\varepsilon_t$$

Based on Hyndman-Khandakar Automatic Arima modeling algorithm, we find the Arima model parameter. Step1: Determine the differential number $d\ (0\le d\le 2)$ by repeating the KPSS test; Step2: After data difference, the optimal (p,q) are selected by minimizing AICC. The algorithm used stepwise search rather than traversing all possible to find (p,q).

We also checked the accuracy of the model by calculating the RMSE of the model training set and testing set. Finally, a randomness test (Ljung-Box test) is performed on the residuals of the model to test whether the residuals are white noise.

The best model is ARIMA (1,0,1) with zero mean, and the model’ s results as follow:

table 3: summary of ARIMA model result

DATA	ME	RMSE	MAE	MASE
Train set	-0.0005	2.8082	1.6233	0.7951
Test set	0.0033	3.2134	2.010	0.9580

Then the randomness test (Ljung-Box test) is performed on the residuals of the model. P-value of the test is less than 0.05, which indicates that the residual of the model is not a white noise sequence. So, this ARIMA model not fit the RTLMP data well.

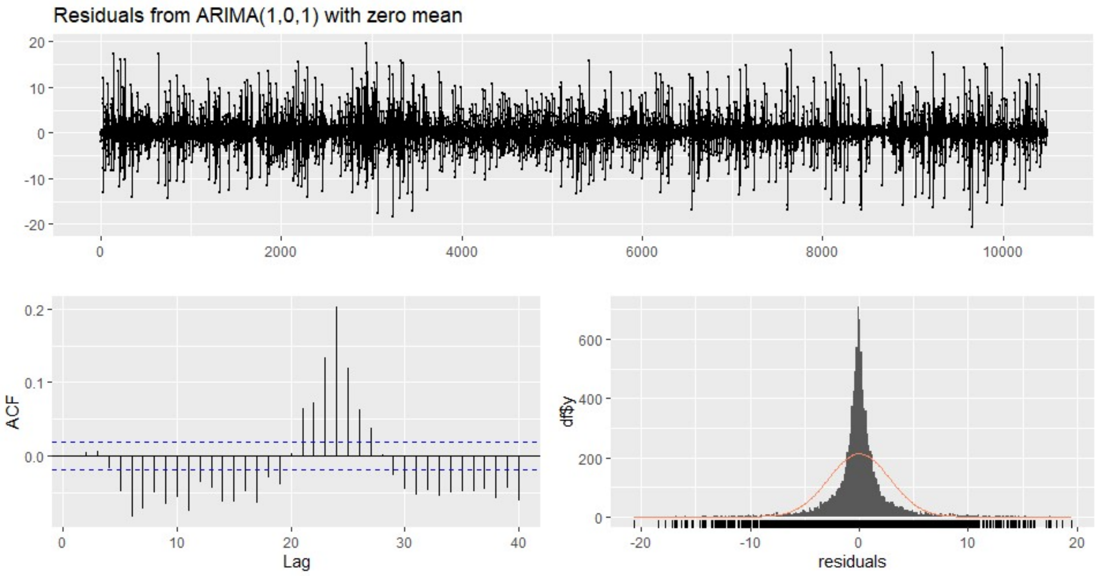


Figure 32: Residual analysis diagram of ARIMA model

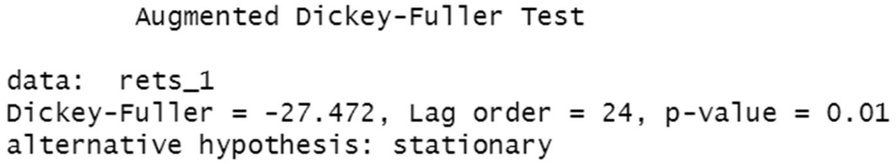
3.3 RTLMP Growth Rate Time Series

i. Stationary

We calculated the growth rate of RTLMP under the same hour as a new time series $\{Y_t\}$.

$$Y_t = \frac{RTLMP_t - RTLMP_{t-24}}{RTLMP_{t-24}}$$

The p-value of ADF test for $\{Y_t\}$ is less than 0.01, so we can conclude that the $\{Y_t\}$ time series data is stationary.



```
            Augmented Dickey-Fuller Test

data:  rets_1
Dickey-Fuller = -27.472, Lag order = 24, p-value = 0.01
alternative hypothesis: stationary
```

figure 33: ADF test results for $\{Y_t\}$

ii. Outliers

As Before, we handled the outliers.

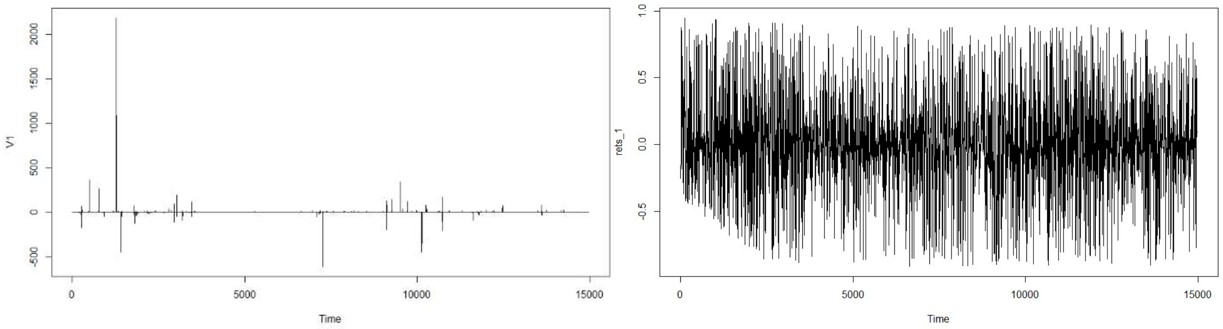


figure 34: The left picture is the $\{Y_t\}$ time series plot without processing outliers, and the right picture is after processing outliers.

iii. ARIMA Model

Now the autocorrelation diagram of the sequence shows the characteristics of tailing, and there is no obvious periodicity. We directly construct the Arima model for the sequence.

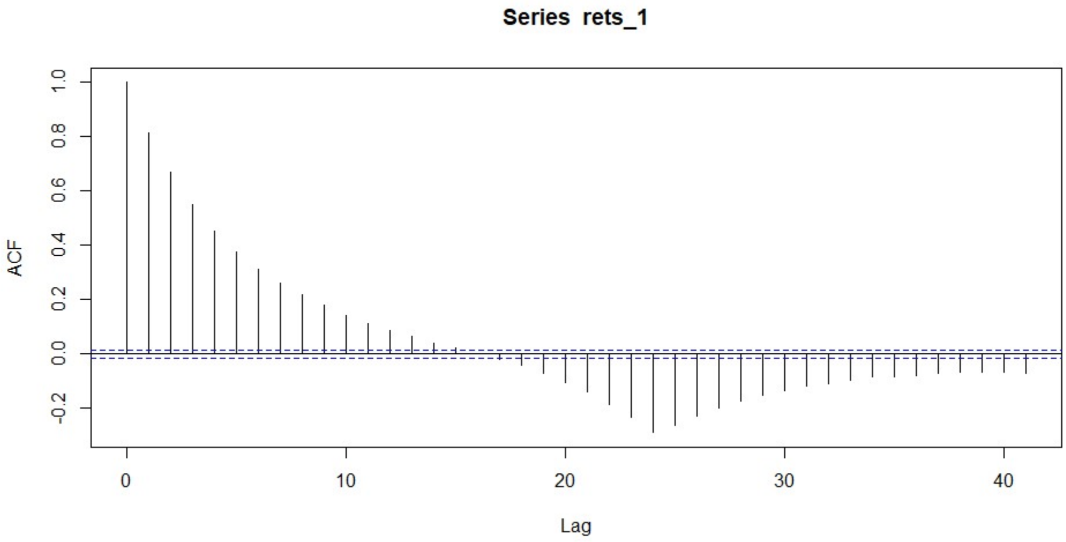


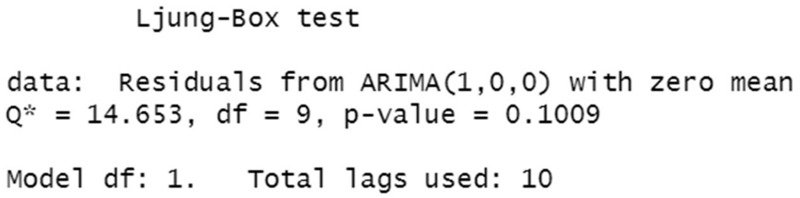
figure 35: ACF plot for $\{Y_t\}$

The best model is ARIMA (1,0,0) with zero mean which equals to AR (1) model, and the model’ s results as follow. This model’ s accuracy is much better than ARIMA model build on RTLMP series.

table 4: summary of ARIMA model result

DATA	ME	RMSE	MAE	MASE
Train set	0.0014	0.1572	0.0982	0.9830
Test set	0.0013	0.1666	0.1047	0.9617

Then the randomness test (Ljung-Box test) is performed on the residuals of the model. P-value is greater than 0.05, we can conclude that this model’ s residual is a white noise.



```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,0) with zero mean
Q* = 14.653, df = 9, p-value = 0.1009

Model df: 1.   Total lags used: 10
```

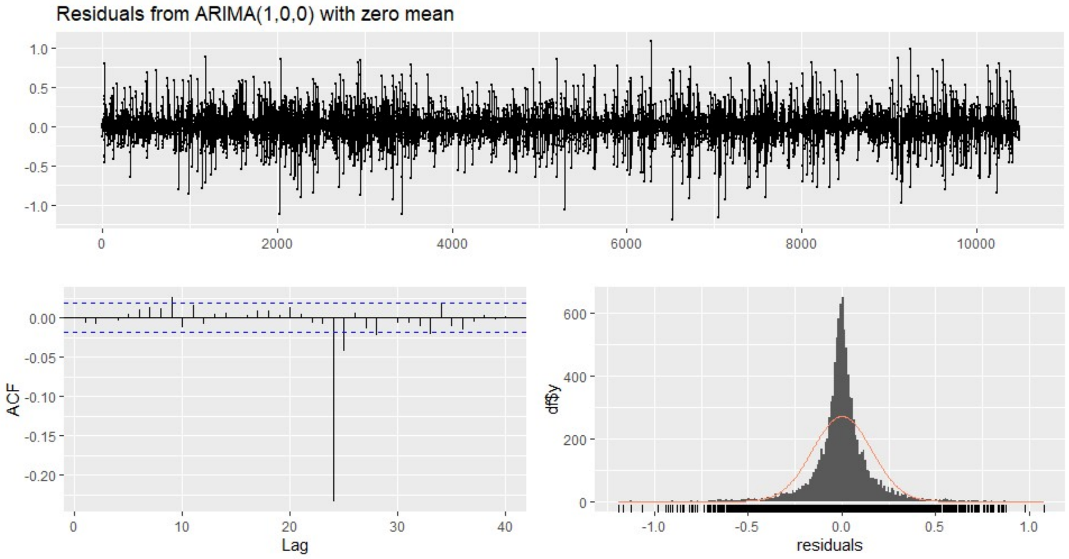


figure 36: Residual analysis diagram

We used this model to make predictions, predicting the distribution of Y_t for the next 24 hours. You can see that the model predicts that the future will remain near the mean after volatility. Although the prediction results show that the prediction accuracy is high, the regression to the mean does not match the actual situation of RTLMP. Because RTLMP is always in a fluctuating state.

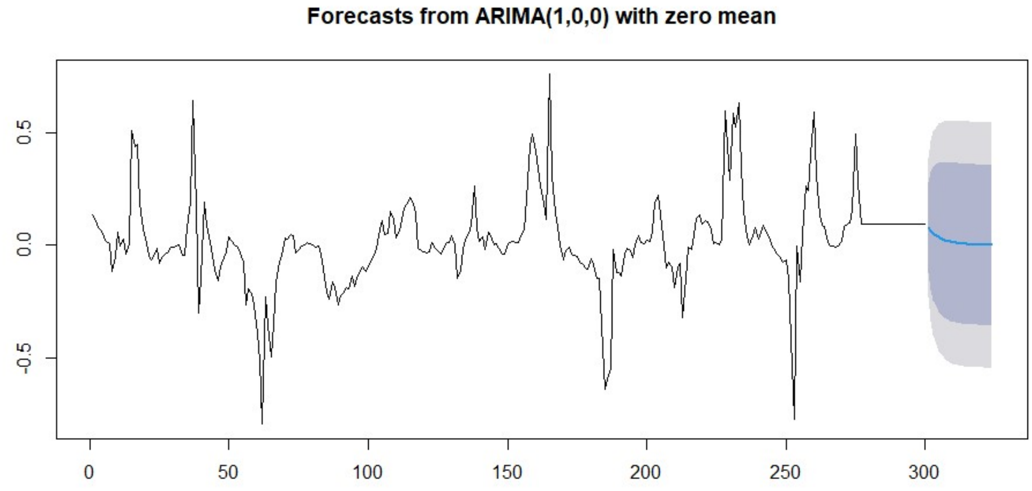


figure 37: Prediction based on ARIMA (1,0,0)

Calculated the predicted RTLMP, we can see that the model predicts that the fluctuation of the RTLMP in the next 24 hours is almost the same as the previous 24 hours.

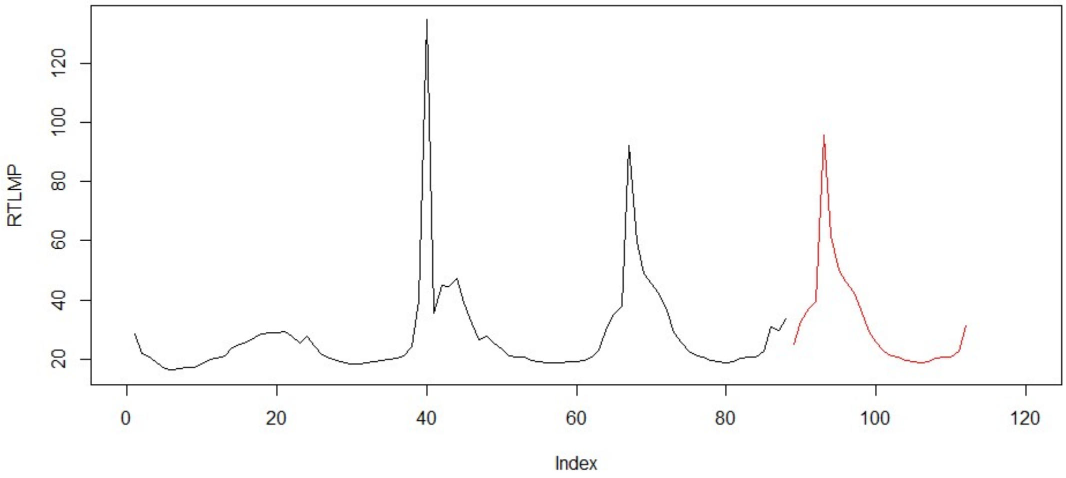


figure 38: prediction for RTLMP based on ARIMA

This ARIMA model formula can be expressed as:

$$Y_t = 0.8261 * Y_{t-1} + \varepsilon_t, \ \ \varepsilon_t \sim WN(0,0.0247)$$

$$Y_t = \frac{RTLMP_t - RTLMP_{t-24}}{RTLMP_{t-24}}$$

iv. NNAR Model

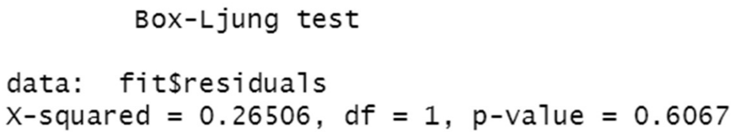
For time series data, the lag values of the time series can also be used as input to the neural network. We call this a neural network autoregressive or NNAR model. In here, we only consider the case of a feedforward network with one hidden layer.

The NNAR model’ s results shown as follows. The accuracy results looks better than ARIMA model.

table 5: summary of NNAR model result

DATA	ME	RMSE	MAE	MASE
Train set	0.0004	0.1400	0.0881	0.8826
Test set	0.0010	0.1566	0.0988	0.9076

Then we checked the residuals. P-value is much greater than 0.05, we can conclude that this model’ s residuals is a white noise.



```
                Box-Ljung test

data:  fit$residuals
X-squared = 0.26506, df = 1, p-value = 0.6067
```

Figure 39: residuals analysis for NNAR

The model is NNAR (26,14) means that there are 26-period lag inputs $(y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-26})$ and 14 nodes in the hidden layer. From the model prediction chart, the growth rate of RTLMP will fluctuate upward in the future.

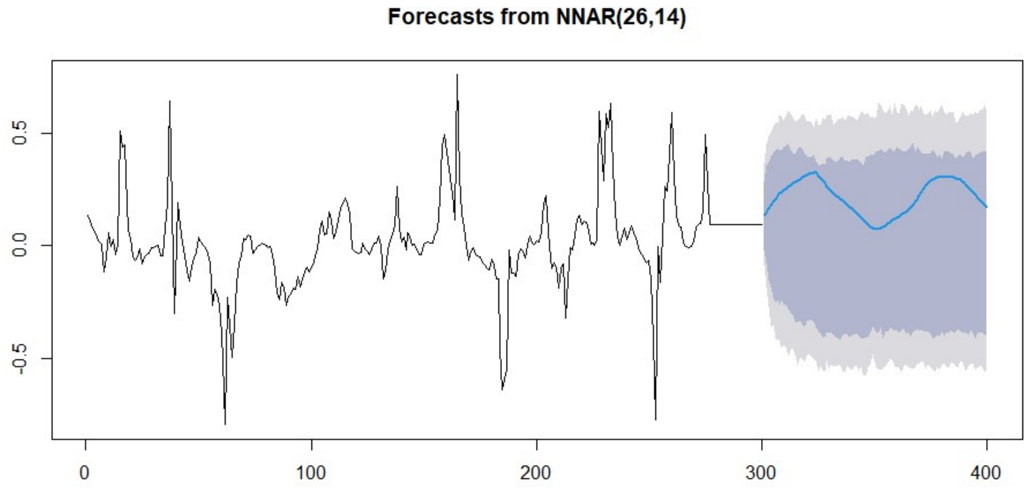


figure 40: Prediction based on NNAR (26,14)

We calculated the corresponding RTLMP based on the predicted growth rate. The model predicts that the RTLMP fluctuation in the next 24 hours will be slightly smaller than the previous period. But there will also exists an outlier with a high level RTLMP.

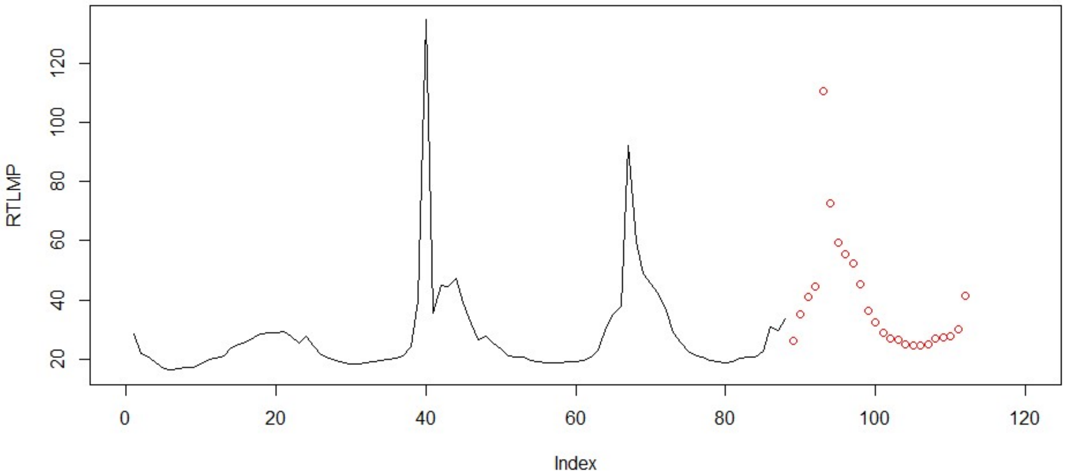


figure 41: prediction RTLMP based on NNAR (26,14)

v. Results

table 6: prediction result of RTLMP

DATETIME	RTLMP	DATETIME	RTLMP
2018-09-17 13:00:00	26.35394	2018-09-18 01:00:00	28.84694
2018-09-17 14:00:00	35.28562	2018-09-18 02:00:00	26.90673
2018-09-17 15:00:00	41.12293	2018-09-18 03:00:00	26.46749
2018-09-17 16:00:00	44.56075	2018-09-18 04:00:00	25.19957
2018-09-17 17:00:00	110.42225	2018-09-18 05:00:00	24.70022
2018-09-17 18:00:00	72.46480	2018-09-18 06:00:00	24.51763
2018-09-17 19:00:00	59.50896	2018-09-18 07:00:00	25.15938
2018-09-17 20:00:00	55.64502	2018-09-18 08:00:00	26.82788
2018-09-17 21:00:00	52.16115	2018-09-18 09:00:00	27.54336
2018-09-17 22:00:00	45.39449	2018-09-18 10:00:00	27.61241
2018-09-17 23:00:00	36.35040	2018-09-18 11:00:00	30.17044
2018-09-18 00:00:00	32.35413	2018-09-18 12:00:00	41.24694