Time Series Analysis and Forecasting: Monthly Household Electricity Consumption in Singapore

Si Chen, Yanjin Chen, Joshua Gagnier, Nishat Islam

Executive Summary

Understanding and monitoring household electricity consumption is crucial in large cities for optimizing energy management. Accurate forecasting of energy consumption can significantly benefit both the economy and the environment.This project introduces a seasonal Time Series (TS) model tailored to monthly household electricity consumption data in Singapore, covering January 2005 to December 2019. Another objective of this project is to consolidate our understanding of course materials and be able to apply the knowledge to real-world data.

Our exploratory data analysis utilized time plots and sample Autocorrelation Function (ACF) plots, revealing an upward trend and seasonality in the data. We employed three different differencing methods to stabilize the data, ultimately selecting single seasonal difference and double difference of first and seasonal for the final model specification. We developed various potential models, inspecting each for residual patterns and coefficient validity. Three candidate models were considered to do residual analysis to evaluate their fitness. Lastly, a train-test split approach was applied for forecasting, using historical data to benchmark the performance of these models.

In conclusion, we implement a combination of both the dynamics method and a method outlined in the textbook to fit 3 adequate models, and narrow it down to a complex and a simpler model. We decide to use the simpler model overall, and keep the complex model as a comparison over time in order to ensure it does not eventually substantially outperform the simple model.

Introduction

Household electricity forecasting is vital in the planning process at a city level. This is because electricity cannot be stored in large quantities. Therefore, it is imperative to provide accurate forecasts as they directly translate into the cost of production and maintenance. For instance, underestimated electricity consumption for the summer months leads to the failure to cover the actual demand of residents.

In this project, we will consider the monthly electricity consumption. In general, such time series contain trends (linear or nonlinear), annual seasonality, and random disturbances (such as extreme heat waves). There is a wide range of methods implemented on forecasting electricity consumption, such as linear regression, ARIMA, etc. In particular, the linear regression models consider the linear trend. However, they require extra steps on modeling seasonal cycles.

The dataset we used is from Kaggle. It records monthly average household electricity consumption (in GWh) in Singapore from Jan 2005 to July 2020 across various dwelling types. It is a multivariate series of 10 variables with information on electricity consumption of various housing types (besides date column). Since multivariate TS analysis is out of scope of this project, we decide on using the 'overall' column to make it an univariate TS analysis. In addition, we remove year 2020 data, as it was partial and contains huge outliers for the model. Lastly, we sort the data as the original file is not ordered by time properly.

Our goal of this project is to consolidate and expand our understanding of the concepts studied in this course. In this project, we will recognize the trend and pattern of the overall electricity consumption time series data. Specifically, we will check stationarity and seasonality for the given series. After our analysis, we aim to find the best possible time series model to do forecasting on future electricity consumption. We will implement both the Box-Jenkins and the dynamics approach as our model-building strategies. During our model specification, we will choose the most appropriate values for p, d, q, P, D, Q for the selected model and also check on the appropriateness of the fitted model.

Analysis

Exploratory Data Analysis

First, we plot the time series data of the monthly household electricity consumption with a fitted linear trend. In Figure 1(left), we can easily observe an upward trend, which leads us to specify the model as non-stationary. Since the stationarity of the TS plays a key role in the model building process, we will need to transform the TS into a stationary form before attempting to form a stationary model. Thus, we perform the lag-1 difference to remove the upward trend. From the sample ACF plot on the right, we can prove the upward trend by the fact that the ACF fails to die out rapidly as lag increases. Besides, strong yearly seasonal autocorrelation relationships can be detected, especially in lag 12, 24, 36 and so on.

The monthly plot in Figure 2 (top) shows the trends over years in the individual months. Again, we observe strong upward trends across all months, in particular the summer months. The seasonal plot in Figure 2 (bottom) allows us to observe features of annual cycles. Note that the shapes of the annual cycles show similarities: the electricity consumption in the summer months (Jun and Jul) is greater than the electricity consumption in the winter months (Jan and Feb). This is reasonable as Singapore has a tropical climate.

First difference on monthly electricity consumption data

We apply a non-seasonal first difference to stabilize the data, which means we subtract the current observation from the previous one. Figure 3 shows the change of monthly electricity consumption against time. In the time series plot, we can see that the upward trend disappears, and the seasonality still presents. From the sample ACF plot on the right, we can see that there appears to be some decrease in correlation but still multiple spikes exceed the significance bounds. The fixed interval of significant spikes represents a seasonal pattern in the data.

Seasonal difference on yearly electricity consumption data

Since we already know that there is strong seasonality in the data, we can also apply the seasonal difference with the lag of 12 to see its effect. After applying the seasonal difference, Figure 4 shows that both the upward trend and the seasonal component have been removed. From the sample ACF plot, we can see spikes decrease rapidly and most of the spikes are within the significant bounds which means most of the autocorrelation has been removed. It tells us that the seasonal differencing is effective in stabilizing the mean of series and a seasonal model is appropriate for us to do the forecasting.

Apply both first difference and seasonal difference

Figure 5 appears to show no trend or seasonality, and the ACF plot shows that the autocorrelations are within the significance bounds for most lags, indicating that the differencing has been effective. After taking both the first difference and season difference on electricity consumption, we observe that the time series data also satisfies the assumptions of ARIMA modeling: the data is stationary with mean centered around zero as shown in Figure 5. Note that the spike at Jun 2016 could be random.

In summary, as both the single seasonal differencing and the double differencing remove the trend effectively, we will continue our model specification with the two options.

Model Specification

Box-Jenkins Method

First, we perform the Box-Jenkins method on the differenced data. We start with ACF and PACF testing to determine tentative model parameters. We also consider the results of `Auto.Arima()` function. In addition, the EACF testing failed due to the seasonal effects within the time series object.

The ACF plot in Figure 6 shows a decay at lag 1 which is the characteristic of the AR component, and the PACF plot has a spike cut-off at lag 1, which suggests a possible AR(1). For seasonal parts of the SARMA model, if we only look at the lag 12, 24, 36 and so on, we will see a slow decay of correlation, which is the indication of a possible AR component.

In the PACF plot of Figure 7, there is a tailing off which suggests a possible MA component. Therefore, we will consider an MA(2) by the significant spike at lag 2 in the ACF plot. For seasonal parts of the SARMA model, we again only look at lags that are multiples of 12, suggesting a possible MA(1).

In summary, the tentative models suggested by the Box-Jenkins method are $\text{ARIMA}(1,0,0)(1,1,0)_{12}$ and $\text{ARIMA}(0,1,2)(0,1,1)_{12}$. However, the orders of AR and MA components are just suggestions that we will conduct model diagnostics to test their appropriateness later.

Dynamics Method

The second method we follow is the dynamics method. First, we start fitting a $(1,0,0)(1,1,0)_{12}$ model, the least polynomially complex model that takes into account seasonality (Figure 8):

```
Coefficients:
          ar1      sar1   constant
       0.6716   -0.5146     0.7634
s.e.   0.0568    0.0657     0.2734
```

Looking at this model in Figure 8, we can see that there are still ACF's that are not explained sufficiently, and the lag 2 correlation is substantially different from the norm, because of this we will increase P, and fit the ARIMA$(2,0,1)(1,1,0)_{12}$ as follows:

```
Coefficients:
          ar1     ar2     ma1      sar1   constant
       0.2532  0.1646  0.6043   -0.5081     0.7615
s.e.   0.1683  0.1379  0.1428    0.0656     0.2406
```

Fitting this model in 9, we can see that there is now much less ACF that is not explained in the first few lags. However, the significance of both the AR1 and AR2 components are not substantially different than 0, which implies we do not need the AR2 component, and by the dynamics method, should attempt to fit a $(1,0,1)(1,1,0)$ next.

```
Coefficients:
       ar1     ma1       sar1   constant
       0.4300  0.4465   -0.5036     0.7600
s.e.   0.0985  0.1026    0.0657     0.2237
```

As shown in Figure 10, this model seems to fit fairly well. Following the dynamics method further, we fit both a (2,1,1) and a (1,1,0) to the seasonal trend without changing the nonseasonal form and found that both AR components were not significantly different from 0 (not shown in figures). Because of this, the final model we find that fits adequately is a $(1,0,1)(0,1,1)_{12}$ as shown in Figure 11.

```
Coefficients:
          ar1     ma1      sma1   constant
       0.3688  0.4649   -0.9998     0.8110
s.e.   0.1025  0.1002    0.1603     0.0543
```

Textbook Method

The next method we use is a method in the textbook (Time Series Analysis With Applications in R) on page 235. This method relies on performing differencing on both the seasonal, and the nonseasonal components, and then fitting MA's to the remaining non-noise terms. To start with, we fit a $(0,1,1)(0,1,1)$.

```
Coefficients:
            ma1      sma1
        -0.2471   -0.9999
```

```
s.e.    0.2418   0.1239
```

Looking at this model as shown in Figure 12, it seems to do a good job for every lag except 2, which has a huge outlier ACF. It performs better on the residual Q-Q plot than any model we have fit so far. However, we are concerned that it is overly complex, based on the fact that it is doing two instances of differencing, and a large polynomial complexity. Because of that huge lag 2 correlation, we extend the model to a $(0,1,2)(0,1,1)_{12}$.

```
Coefficients:
           ma1       ma2      sma1
        −0.198   −0.4900   −1.0000
s.e.     0.066    0.0647    0.1706
```

This model (Figure 13) fits extremely well to the data and has no ACF's outside of the standard range for the first 20 lags. It still maintains a good normal QQ plot and doesn't seem to have much trend. However, it may overfit as it has a large polynomial complexity compared to the dynamics method model. We will see how it performs comparatively in our assessment stage.

Assessing the fitted models

Model 1: $(1,0,1)(1,1,0)_{12}$: As shown in Figure 14, we see that: Normality assumption met , histogram mode at 0, normally distributed , residual plot shows no trend , most of the observation lies at 45 degree line at QQ plot. Residual scatter plot shows strong positive correlation.

Model 2: $(1,0,1)(0,1,1)_{12}$: As shown in Figure 15, we see that: Normality assumption met , histogram mode at 0, normally distributed , residual plot shows no trend , most of the observation lies at 45 degree line at QQ plot. Residual scatter plot shows strong positive correlation.

Model 3: $(0,1,2)(0,1,1)_{12}$: As shown in Figure 16, we see that: The histogram mode at 0, normally distributed , residual plot shows no trend ;striking deviation of observations at the tail but most of the observation lies at 45 degree line of QQ plot. Residual scatter plot shows weak positive correlation.

Predicting using the fitted models

We split the electricity consumption data into train and test sets using a 2:1 ratio. We first fit the three selected models using the training data set and perform 5 years forecasting. Then we add the actual testing data to do the comparison. Figure 17 shows the three models applied to the electricity consumption data from Jan 2005 to Dec 2014, with the forecasts compared against actual values in the next 5 years (60 months). We note that model 3 ($ARIMA(0,1,2),(0,1,1)_{12}$) forecasts are close to the observed values and also capture the overall upward trend.

Conclusion

After we fit the 3 models and generated the predictions, we can see that the three models all have very competitive predictive power. Model 1 (Differencing in only Seasonality) and Model 3 (Differencing in Both Seasonal and Nonseasonal) are the two models we would consider most effective, with model 1 performing worse overall, but having considerably fewer components. While model 3 has less correlation in the residuals, implying that there is more information being caught by the more complex model. In conclusion, we decide to follow the principle of parsimony and use model 1 $(1, 0, 1)(1, 1, 0)_{12}$ for prediction by default, and keep model 3 $(0, 1, 2)(0, 1, 1)_{12}$ as a comparison.

Appendix

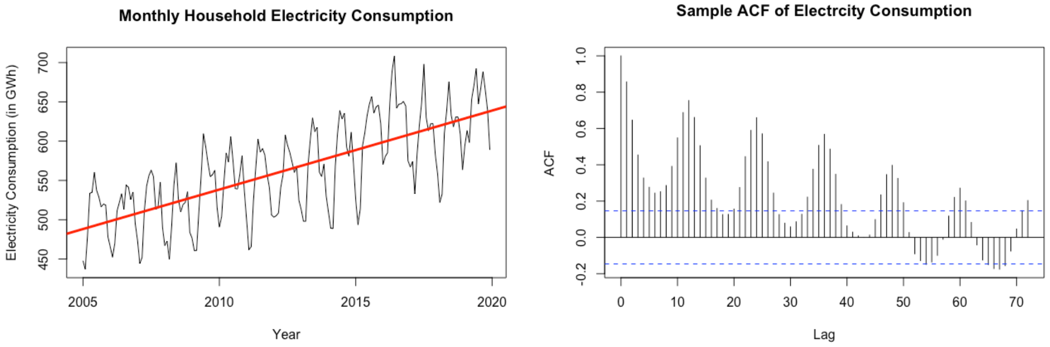

Figure 1: Monthly electricity consumption time series (left) and sample ACF (right)

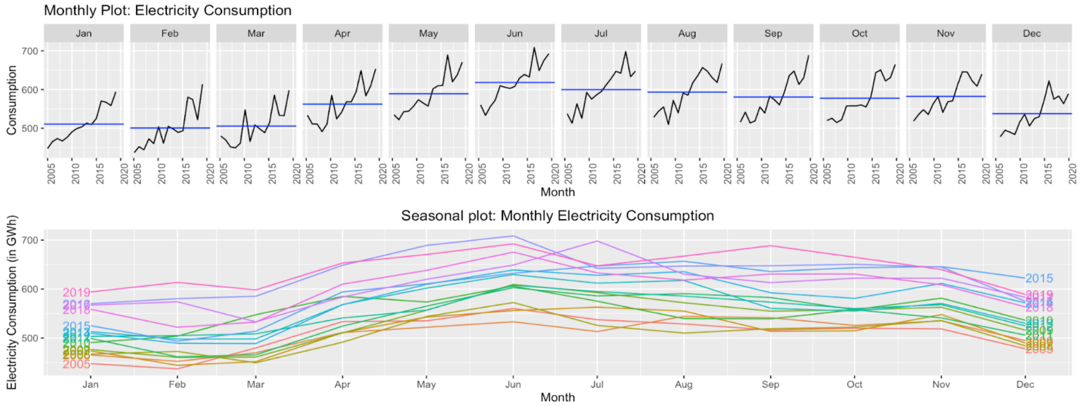

Figure 2: Monthly electricity consumption plot (top) and seasonal plot (bottom)

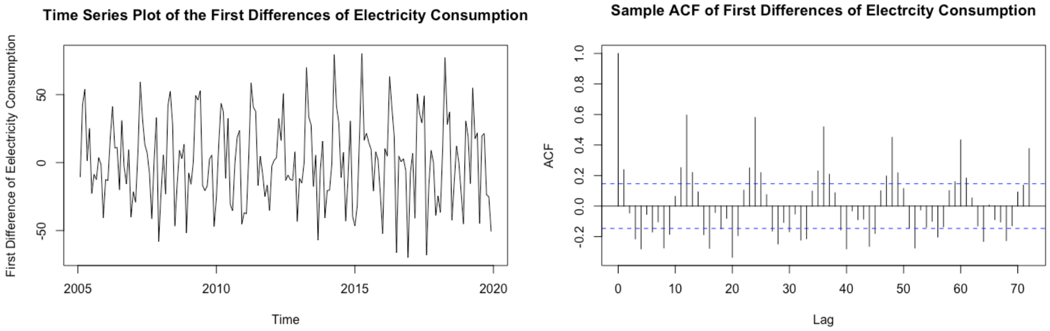

Figure 3: First difference monthly electricity consumption data (left) and sample ACF (right)

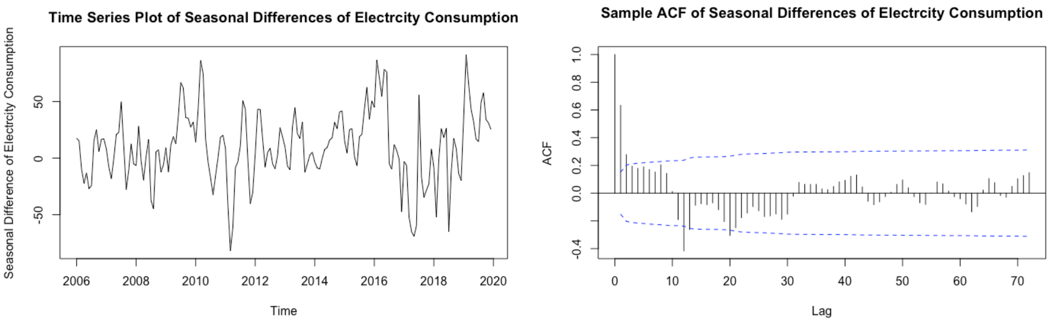

Figure 4: Seasonal difference monthly electricity consumption data (left) and sample ACF (right)

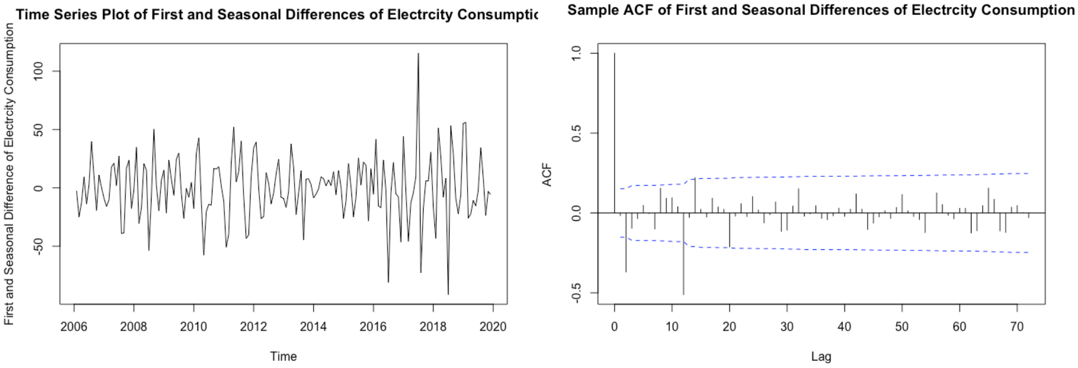

Figure 5: First and seasonal difference monthly electricity consumption data (left) and sample ACF (right)

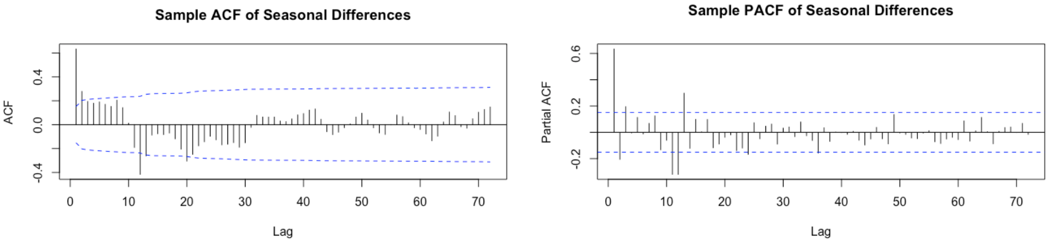

Figure 6: Sample ACF and sample PACF of seasonal seasonal difference monthly electricity consumption data

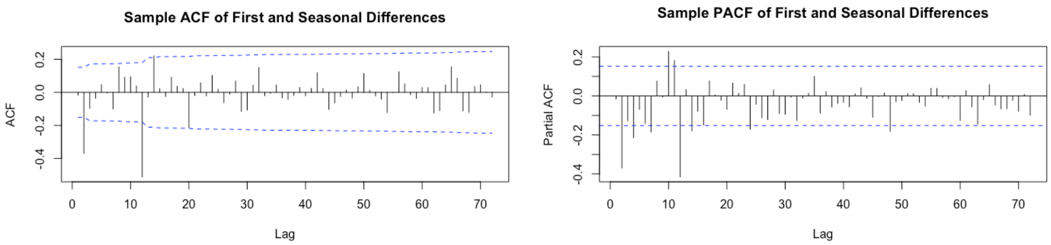

Figure 7: Sample ACF and sample PACF of first and seasonal difference monthly electricity consumption

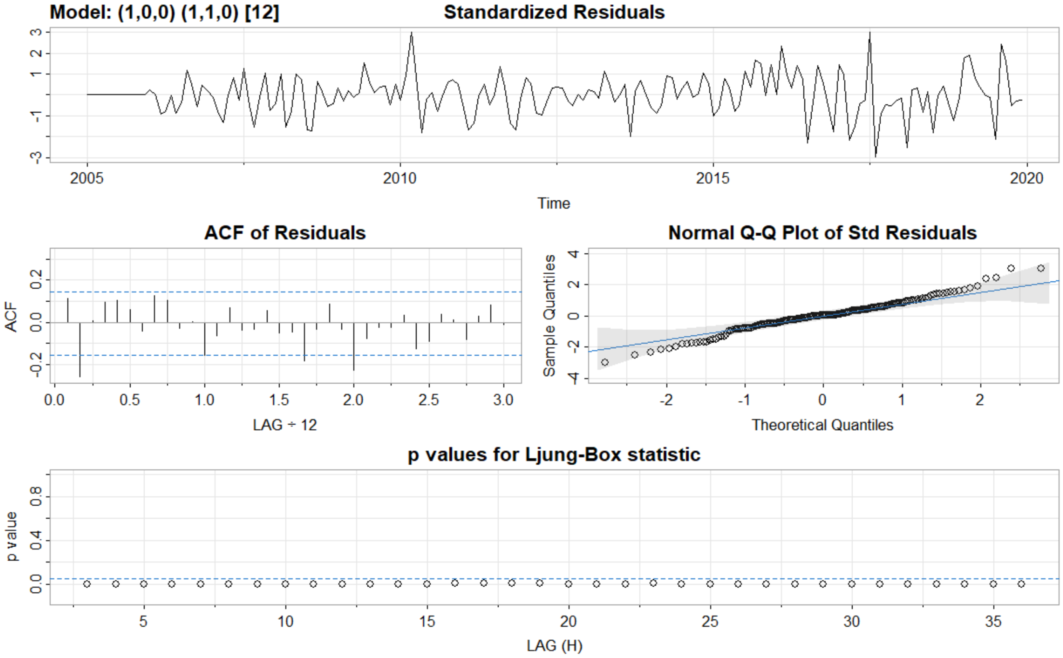

Figure 8: Dynamics Method Model 1

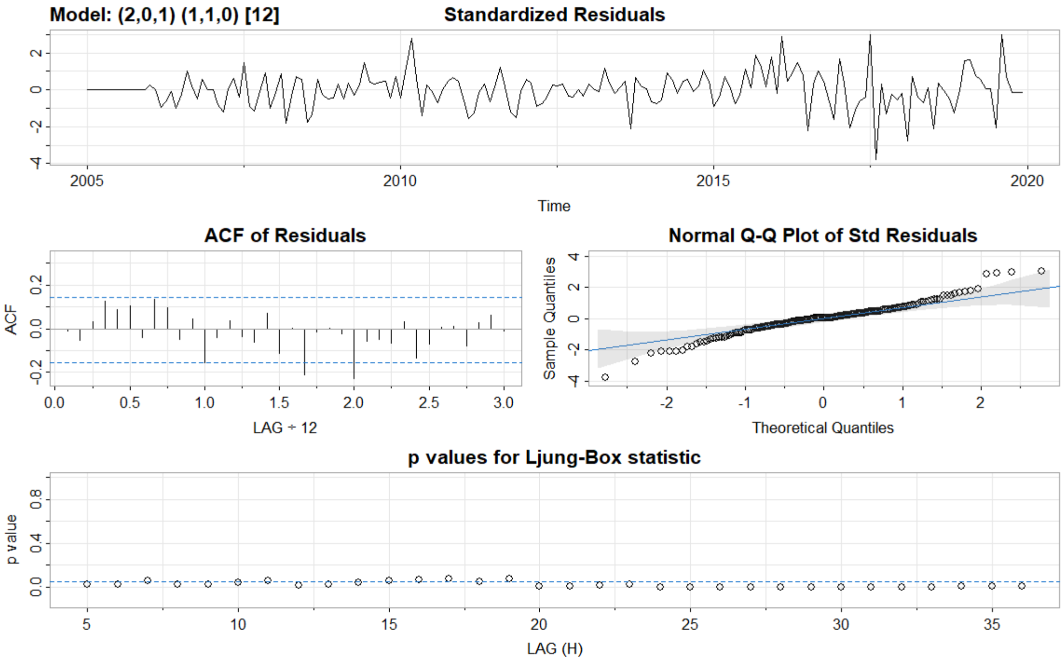

Figure 9: Dynamics Method Model 2

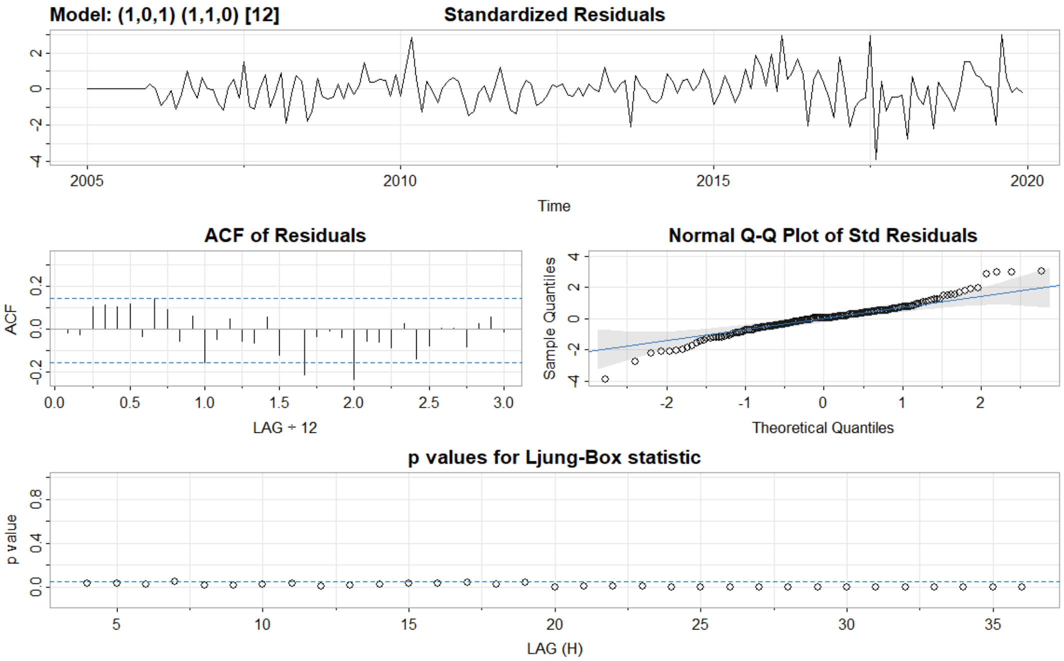

Figure 10: Dynamics Method Model 3

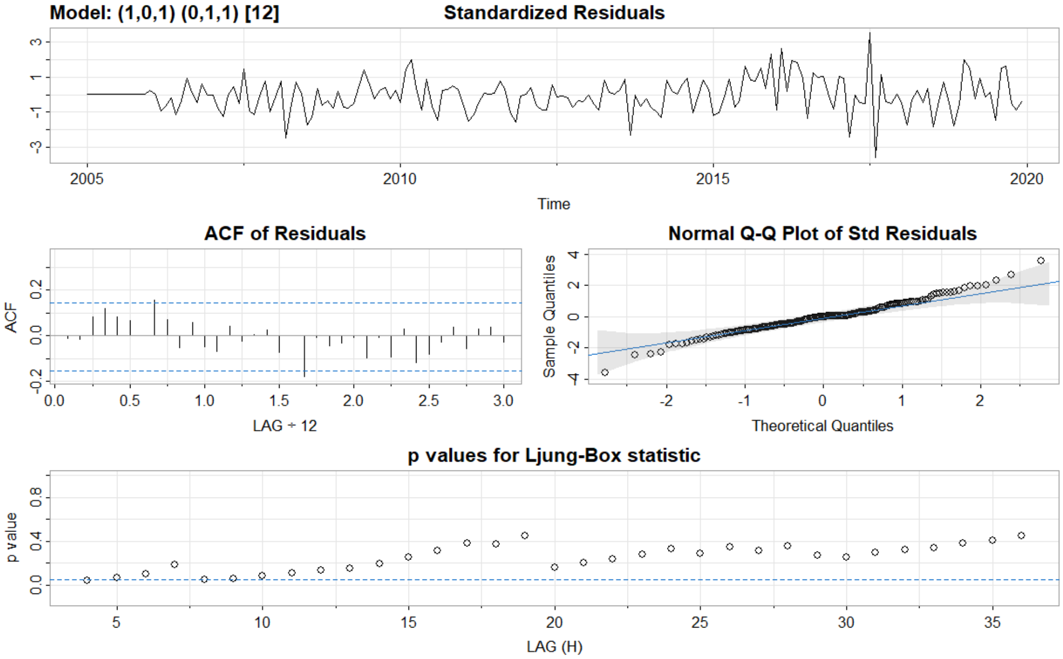

Figure 11: Dynamics Method Model 4

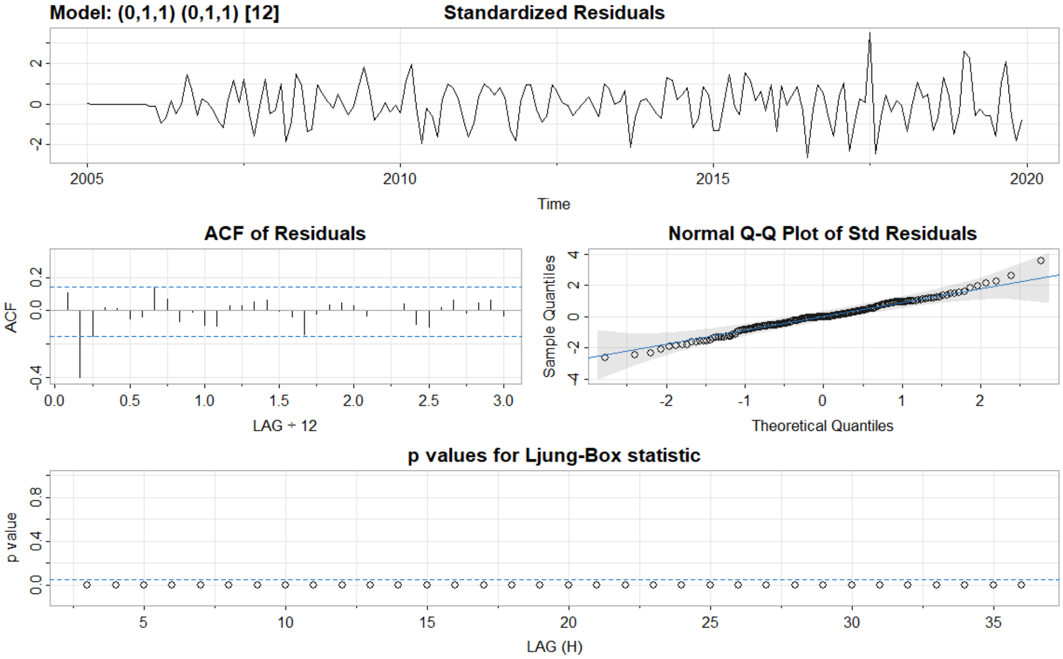

Figure 12: Textbook Method Model 1

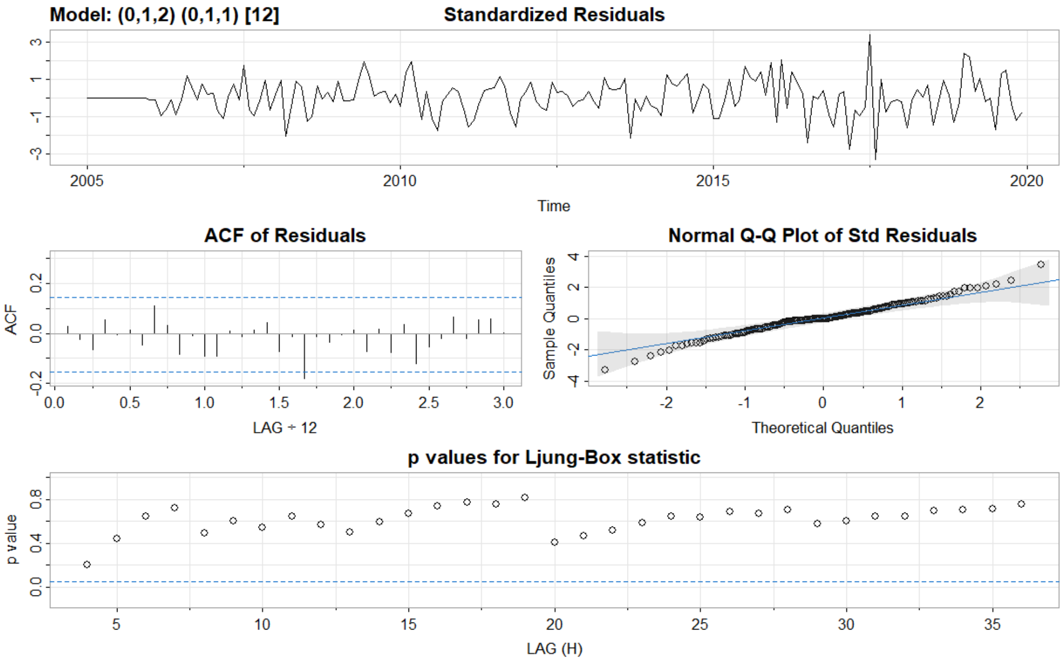

Figure 13: Textbook Method Model 2

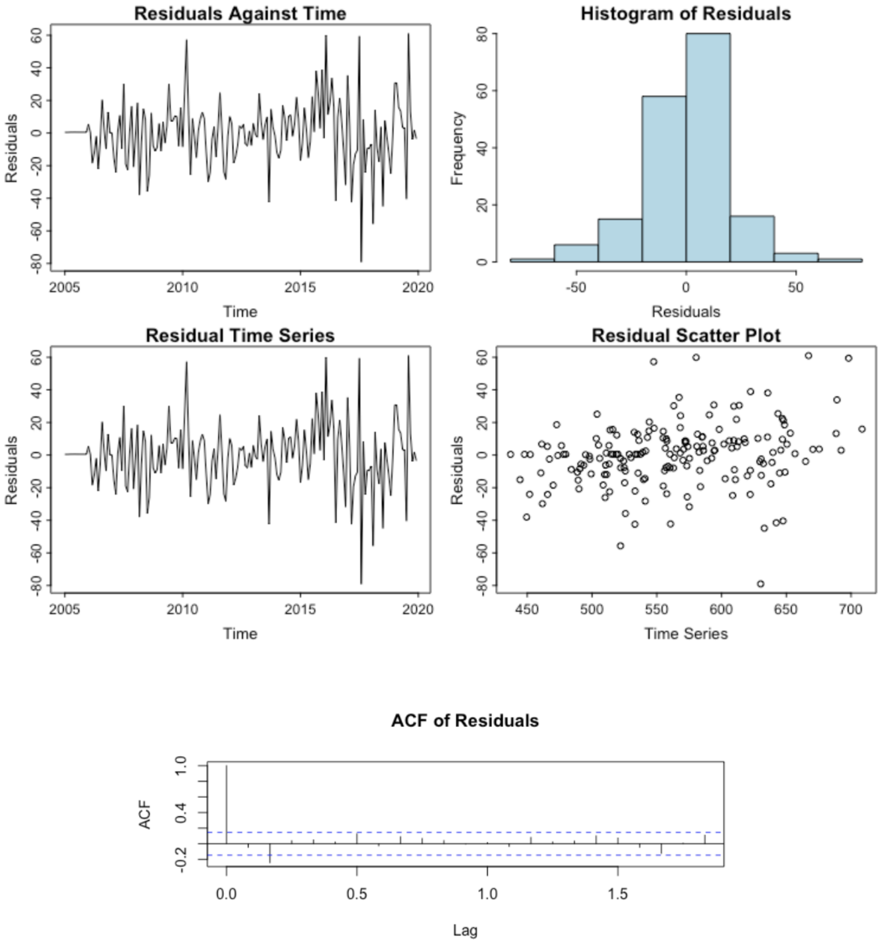

Figure 14: Model 1 Diagnostic plots

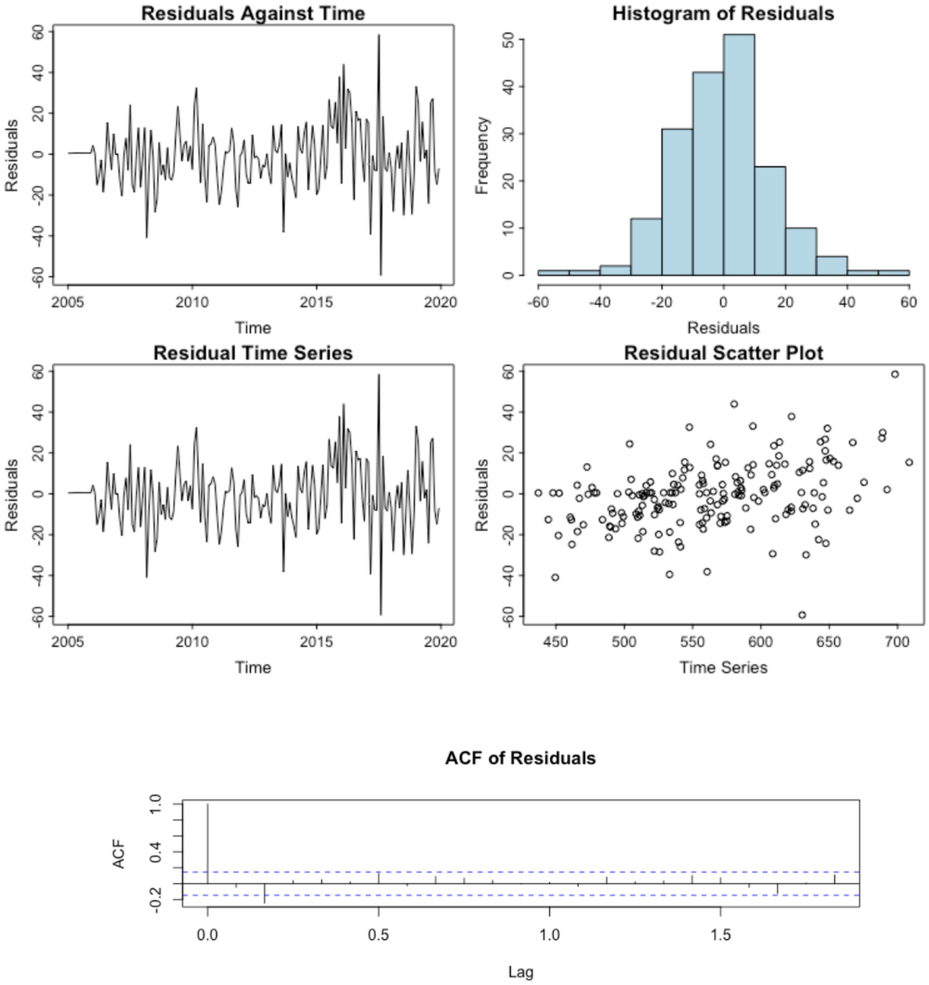

Figure 15: Model 2 Diagnostic plots

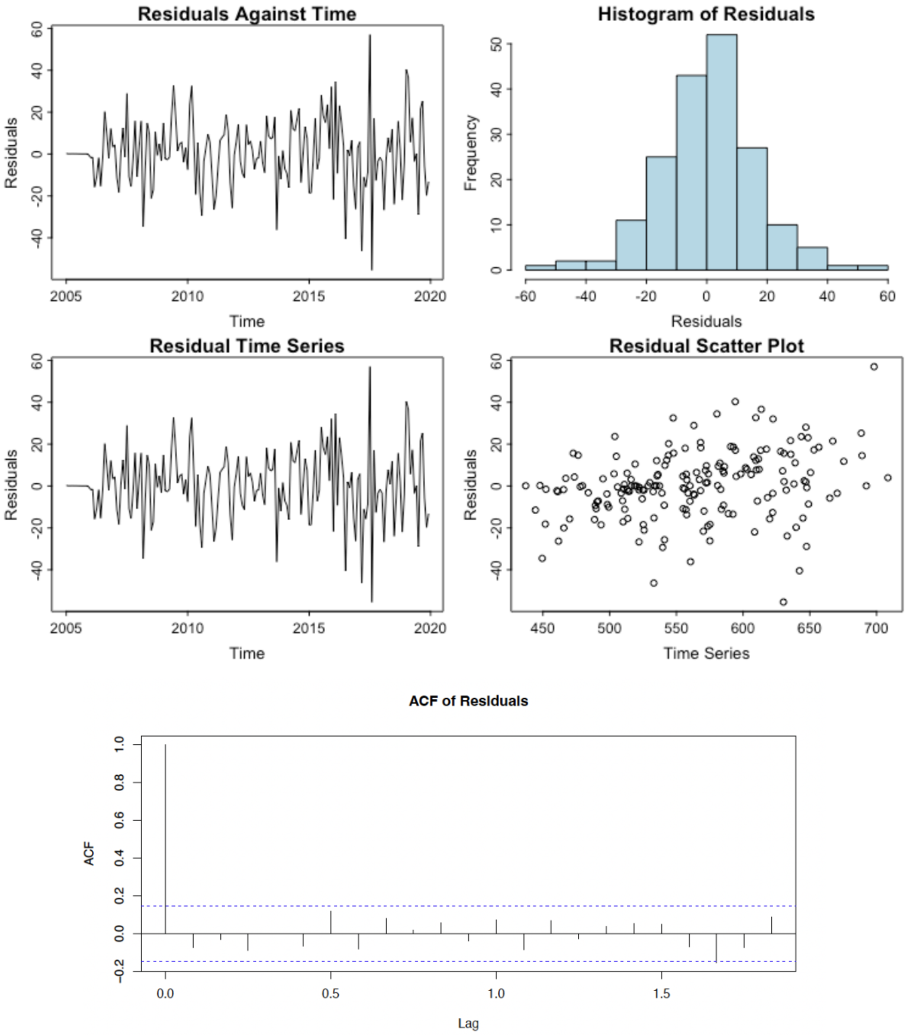

Figure 16: Model 3 Diagnostic plots

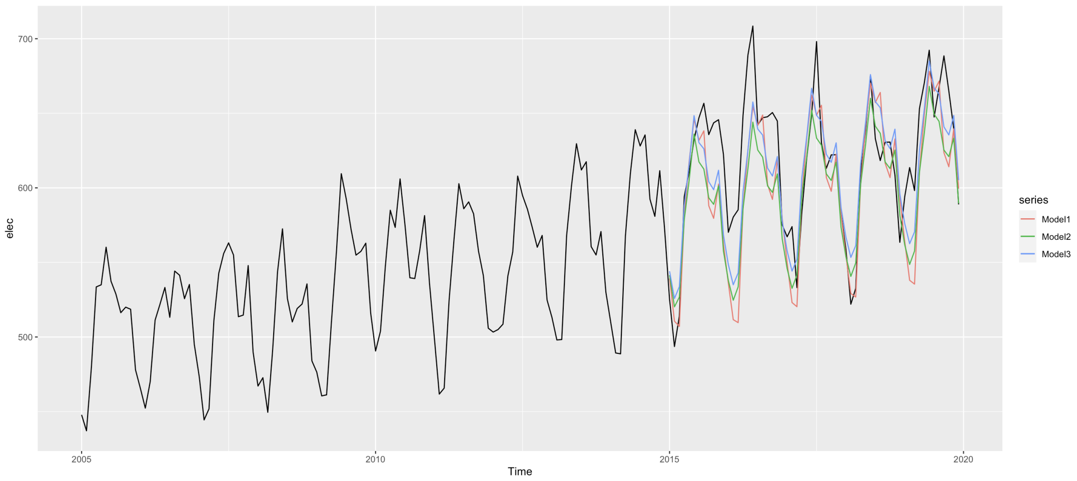

Figure 17: Three models on prediction the last five years