Time Series Analysis and Forecasting of Average Monthly Electricity Price among cities in U.S.

Shaoxuan Chen

Abstract

The aim of this project is to fit a valid time series model to predict the future average monthly electricity price in the United States. The data of this project is about 10 years average monthly electricity price per kWh in the United States from 01/2010 to 03/2020, which was scratched from the U.S. Bureau of Labor Statistics website. The last 15 out of 123 observations were set as the testing data, and the others were set as the training data. Multiple ARIMA and SARIMA models were built in model identification procedure based on training data, and adequate models with roughly lowest AICc were chosen to be compared with Holt-Winters Algorithm in forecasting procedure. Forecast accuracy were evaluated by MAE, RMSE and MAPE on testing data, suggesting SARIMA models have much better performance than Holt-Winters Algorithm. Based on these results, conclude that SARIMA$(0,1,1)\times(0,1,1)_{12}$ is the best model for this data. And the explicit form of the model for the stationary data is $X_t = e_t - 0.2669e_{t-1} - 0.6367e_{t-12} + 0.17e_{t-13}$.

1 Introduction

As the demand for electricity has been increasing a lot in the past few decades, especially with the rapid progress of hi-tech industrialization, the prices of electricity was driven up quickly. Thus, it is an interesting topic to gain insights on the growth trend of the electricity price in time, or even fitting the time series model to perpetuate the exactly electricity price at certain time in the future. In this project, we are interested in the average monthly electricity price per kilowatt-hour (kWh) in U.S. region. The primary objectives for this project is to fit a valid time series model to predict the average price of electricity in the future.

2 Data description & Prepossessing

The data of this project is the average monthly electricity price per kilowatt-hour (kWh) in the United States, which is provided by the United States Department of Labor.[1]. Ten years of data was scratched from the website, from January 2010 to March 2020, which means 123 monthly observations were included. The data was transformed into monthly time series data in the prepossessing step, which makes it convenient to identify the corresponding month/year of the data in the later steps. And there is no missing value in this data set. The time series plot of the data is shown in Figure 1. It is clear that the data is not stationary, and there is a clear trend and seasonal pattern. Also, the variance of the data increase little overtime, so the data also subjects to non-constant variance. To further identify the period of the data, the auto-correlation function(ACF) of the data, up to lag 100, was shown in Figure 2. It is clear that the ACF behaves like a sinusoid, and it takes 12 points to complete one cycle. So we can conclude that the period of the data is 12.

After identifying the period, we can decompose the data, as shown in Figure 2. we can see that the trend component shows some fluctuations from in 2016 to 2020, but overall it has an increasing pattern. The seasonal plot also affirm the period what has been identified above.

[1]Website: U.S. Bureau Of Labor Statistics `https://data.bls.gov/timeseries/APU000072610?amp%253bdata_tool=XGtable&outp`

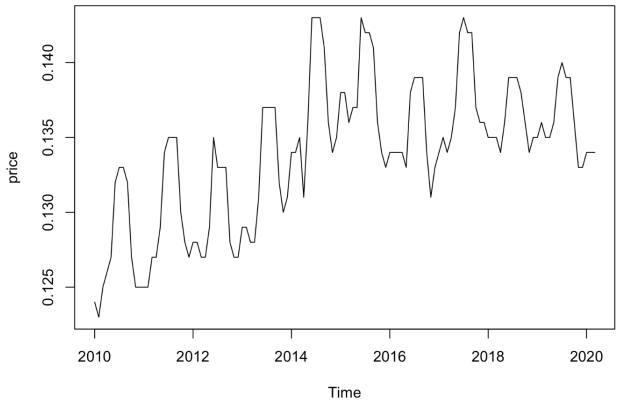

Figure 1: Time Series Plot of the data

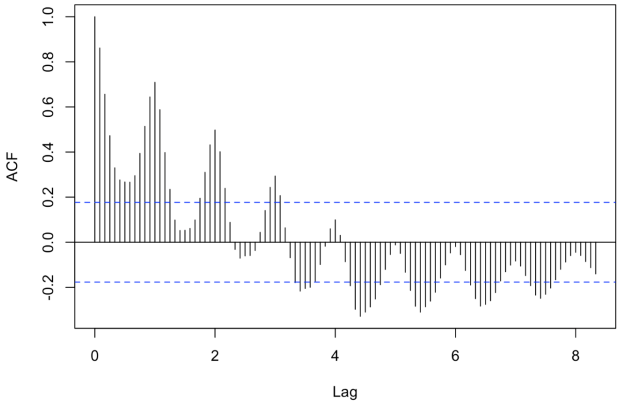

Figure 2: ACF of the original data

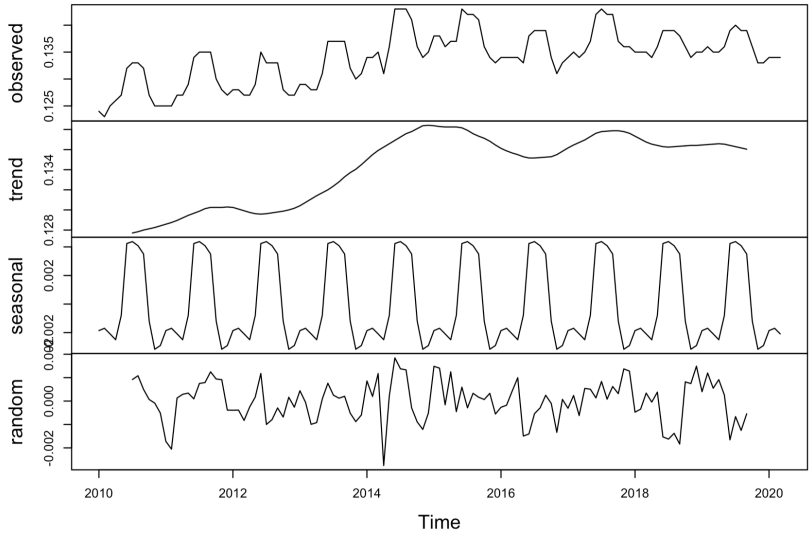

Figure 3: Decomposition of the data

3 Model Identification

In this section, models were built based on four different approaches and evaluated by Akaike information criterion with correction(AICc).

3.1 Lower order ARMA model

To deal with none stationary. At first, the Box-Cox transformation was performed to take care of the non-constant variance. After justification, the log transformation was chosen. Moreover, the trend component and seasonal component were eliminated via differencing. Then the data becomes stationary as shown in Figure 4, the ACF & PACF of the data were also shown in Figure 5.

From Figure 5, it is obvious that both ACF and PACF decay to zero like damped since waves, so lower order ARMA model is possible. Then different combination of lower order ARMA model(up to order 2) were fitted and the result is shown in Table 1. Since ARMA(2,2) has the lowest AICc, so for lower order ARMA model, we should choose model ARMA(2,2).

Table 1: AICc of Lower Order ARMA model

Model	**AICc**	**Model**	**AICc**
MA(1)	-646.74	ARMA(1,2)	-655.8
AR(1)	-644.89	ARMA(2,1)	-660.4
ARMA(1,1)	-647.19	ARMA(2,2)	-680.74

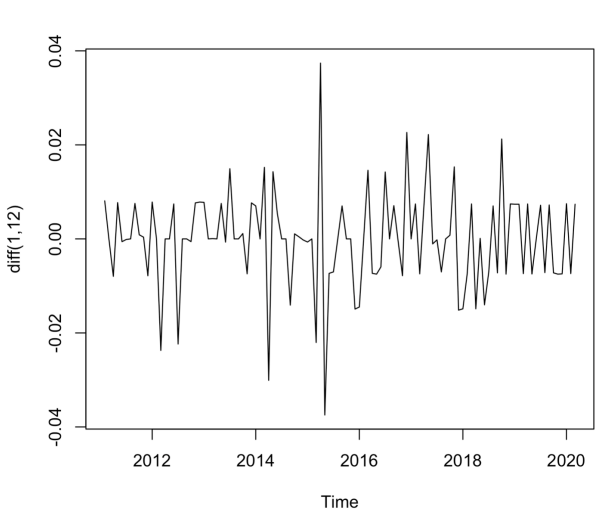

Figure 4: stationary data, Diff(1,12)log(Price)

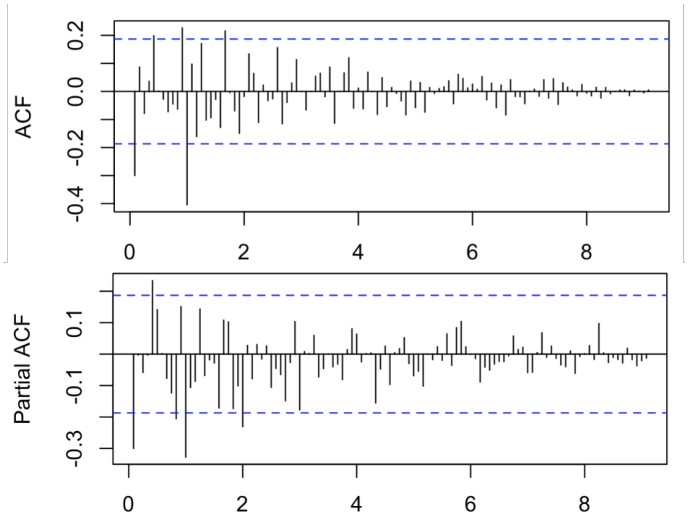

Figure 5: ACF and PACF of stationary data

3.2 Auto ARIMA Algorithm

The *auto.arima* function in R was also performed as a method of model selection. And the lag of both AR and MA parts was chosen to be 15 to make sure the lag is large enough. And the model derived from this approach is SARIMA$(0,1,1) \times (0,1,2)_{12}$ with AICc -720.27.

3.3 SARIMA model based on ACF & PACF

As for seasonal lags in Figure 5, it is obvious that the ACF cuts off at lag *1S* and the PACF at seasonal lags decays to zero exponentially, hence a seasonal MA(1) model is appropriate. But for the sake of assurance, different combination of lower order SARIMA models were fitted and the result is shown in Table 2.

Table 2: AICc of Lower Order SARIMA model

Model	**AICc**	**Model**	**AICc**
$(0,1,1) \times (0,1,1)_{12}$	-720.65	$(0,1,1) \times (1,1,0)_{12}$	-711.43
$(1,1,0) \times (0,1,1)_{12}$	-719.81	$(1,1,0) \times (1,1,0)_{12}$	-711.06
$(1,1,1) \times (0,1,1)_{12}$	-718.71	$(1,1,1) \times (1,1,0)_{12}$	-709.28

From Table 2, it is clear that seasonal MA(1) models have better performance, and SARIMA$(0,1,1)\times (0,1,1)_{12}$ model has the lowest AICc among all of the seasonal MA(1) models.

3.4 SARIMA model based on Subset Selection method

From Figure 6, we can see the best model chosen by subset selection method is SARIMA$(6,1,0) \times (0,1,1)_{12}$, with the AR coefficients at lag2-lag4 constrained to be 0. However, when the corresponded model is fitted in R, the coefficients of ar6 is 0.0636 and the standard error is 0.1013, which implies that coefficient ar6 might not be significant. So the best two models were fitted for further selection and the result is shown in Table 3.

Table 3: AICc of SARIMA model chosen by Subset Selection

Model	**AICc**
$(6,1,0) \times (0,1,1)_{12}$	-720.61
$(5,1,0) \times (0,1,1)_{12}$	-722.41

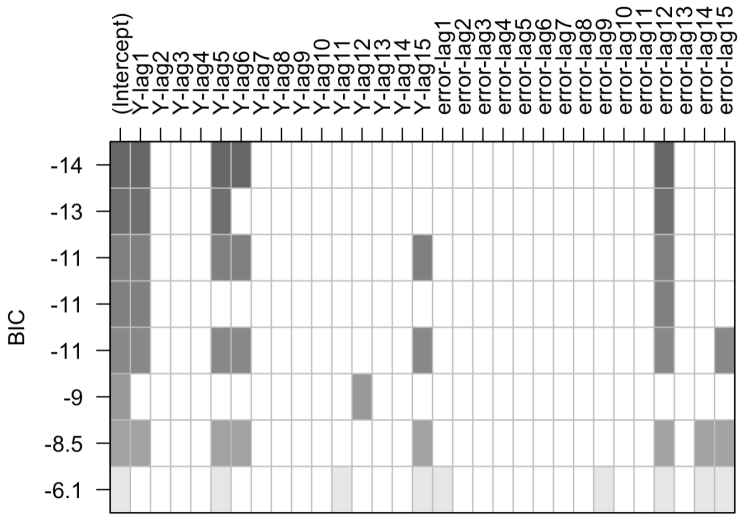

Figure 6: Subset SARIMA Model selection based on BIC

From Table 3, combined with all the results shown in Table 1 and Table 2, it is clear that in terms of AICc criteria, SARIMA$(5,1,0) \times (0,1,1)_{12}$ with the AR coefficients at lag2-lag4 constrained to be 0 is the best model, and the AICc of this model is -722.41.

4 Model Diagnostics

In this section, at first, the adequacy of model SARIMA$(5,1,0) \times (0,1,1)_{12}$ is checked and the results were shown in Figure 7 & Figure 8. In the second plot of Figure 7, it is clear that all of the sample ACF of of residuals fall within interval $[-\frac{1.96}{\sqrt{n}}, +\frac{1.96}{\sqrt{n}}]$. Also, the p values for Ljung-Box statistic are all greater than 0.05. Thus, we can draw the conclusion that the residual process behaves like White Noise process. As for the normality checking in Figure 8, the points roughly follow the reference line, which suggests that the residuals are approximately normal distributed. Overall, we can conclude that model SARIMA$(5,1,0) \times (0,1,1)_{12}$ is adequate.

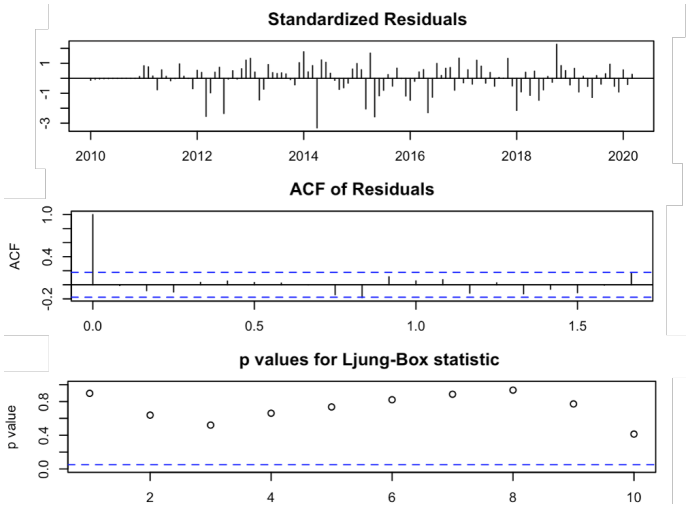

Figure 7: Residual Diagnostics

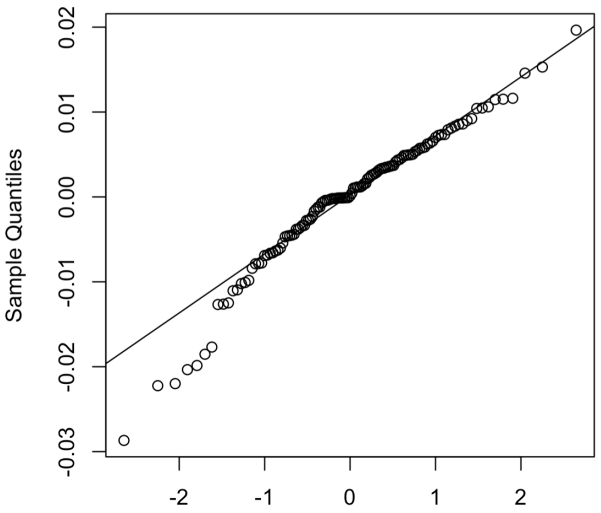

Figure 8: Normal Q-Q Plot

However, since the AICc of several models are roughly comparable as shown in Section 3, so it is possible that models with relatively larger AICc might have better performance in terms of forecasting. The adequacy of those models were also checked, and the adequate models were also selected as candidate models in the later forecasting section.

5 Forecasting Comparison

In this section, based on MAE, RMSE and MAPE, the forecast accuracy were compared among the adequate SARIMA models derived above and the Holt-Winters forecasting method. Note that the last 15 out of 123 observations were set as the testing data. Since it contains a whole period as well as the last three observations. The 15 data points correspond to the monthly electricity price from 01/2019-03/2020. And the other 108 data points were treated as training data. The results were shown in Table 4.

Table 4: Forecast Accuracy of different approaches

SARIMA forecasting Model	**MAE**	**RMSE**	**MAPE**
$(0,1,1)\times(0,1,1)_{12}$	0.00182	0.00208	1.34304
$(1,1,0)\times(0,1,1)_{12}$	0.00197	0.00224	1.45111
$(1,1,1)\times(0,1,1)_{12}$	0.00174	0.00199	1.27571
$(0,1,1)\times(0,1,2)_{12}$	0.00183	0.00209	1.34438
$(5,1,0)\times(0,1,1)_{12}$	0.00229	0.00255	1.68795
$(6,1,0)\times(0,1,1)_{12}$	0.00230	0.00255	1.69039
non-SARIMA forecasting Method	**MAE**	**RMSE**	**MAPE**
Holt-Winters Seasonal(additive)	0.00276	0.00317	2.02817

From above, it is clear that the SARIMA models have much better performance than Holt-Winters Seasonal method in forecasting. And SARIMA$(1,1,1)\times(0,1,1)_{12}$ and SARIMA$(0,1,1)\times(0,1,1)_{12}$ are the best two models. Since the AICc of these two models are relatively the same, in consideration of model simplicity, we can conclude that SARIMA$(0,1,1)\times(0,1,1)_{12}$ is the best model. After parameter estimation, for the stationary data after transformation, the specific from of the final model is:

$$X_t = e_t - 0.2669e_{t-1} - 0.6367e_{t-12} + 0.17e_{t-13}$$

Moreover, for comparison, the plot of forecasts along with forecast interval from model SARIMA$(0,1,1)\times(0,1,1)_{12}$ and Holt-Winters Seasonal method were shown in Figure 9.

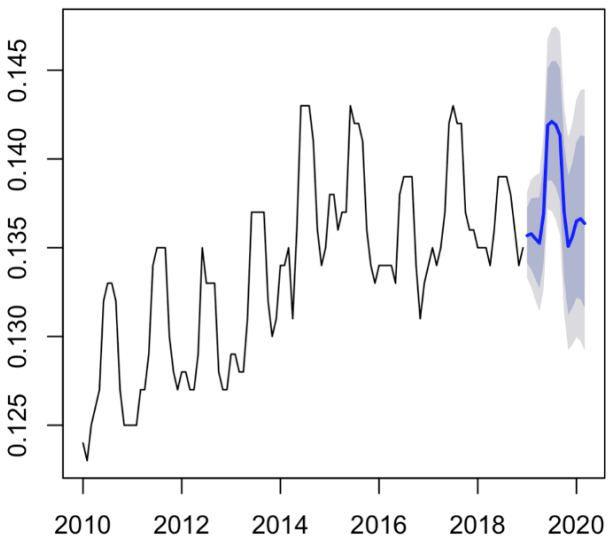

Figure 9: SARIMA$(0,1,1)\times(0,1,1)_{12}$ Forecasts

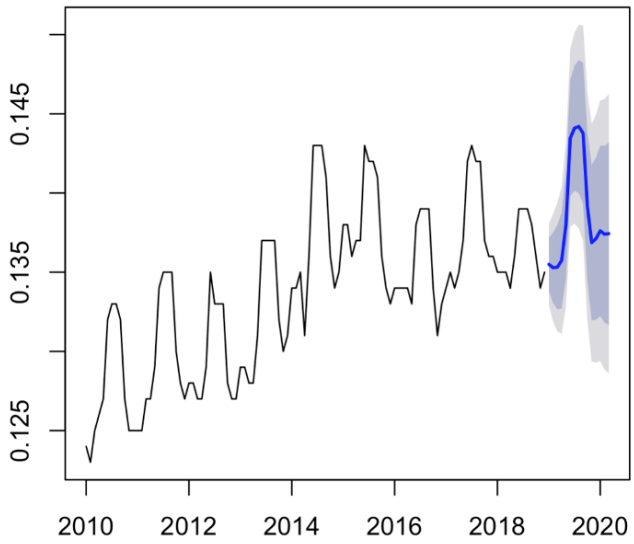

Figure 10: Holt-Winters Forecasts

6 Discussion

As for the monthly average price of electricity among cities in U.S, we can see it is highly correlated with seasons. The price is higher in the summer and lower in the winter. Overall the price has an increasing pattern and the final model fit the data well.

As for model fitting, the final model chosen to predict average monthly electricity price per kWh in U.S. region is the SARIMA$(0,1,1)\times(0,1,1)_{12}$ model. For stationary data after transformation, the specific form of the model is $X_t = e_t - 0.2669e_{t-1} - 0.6367e_{t-12} + 0.17e_{t-13}$. In model identification procedure, we can see that compared with SARIMA models, lower order ARMA models do not have good performance as shown by the much larger AICc values. So we can speculate lower order ARMA model might not be a good choice in dealing with seasonal data. Moreover, even the subset selection method can help us identify the higher order significant lags in the model, the model with the lowest AICc does not necessary has the best performance in forecasting. As for forecasting, the SARIMA$(0,1,1)\times(0,1,1)_{12}$ model performs much better than Holt-Winters Algorithm, which verified that Holt-Winter Algorithm does not handle the seasonal component well.