Time Series Analysis for the Electric and Gas utilities in Industrial Production

Shangchen Han

4/25/2020

Abstract

This time series report focuses on the industrial production especially in electric and gas field. All monthly production data from January 2000 to March 2019 which captured from Kaggle website. Total observations are 233, and it was divided into two parts: train (204) and test (29). There are some steps of my project. Convert data into time series form and also check whether missing valuse. Plot the time series data, and see whether it need to transform. Then use appropriate method ARMA or non-ARMA method to fit the data. After fitting models, compare MAE,RMSE,MAPE to choose the best one model. Finally,use the compatible model to forecast future trend. Thus, the resulting model, SARIMA(2,1,2)(0,1,1)[12] because it has the smallest AICc. Finally, the future pattern could be found from forecasting picture.

1 Data and Transformation

From `Figure` 1,it shows that there is obviously seasonal pattern in the time series data, in every year, May has the least amount of production and in terms of Febrary, the production reach at peak. In general, we could see a slightly increasing trend among years. Thus, it means the data is not stationary. In other words, we need to use some appropriate methods to transform it to be stationary.

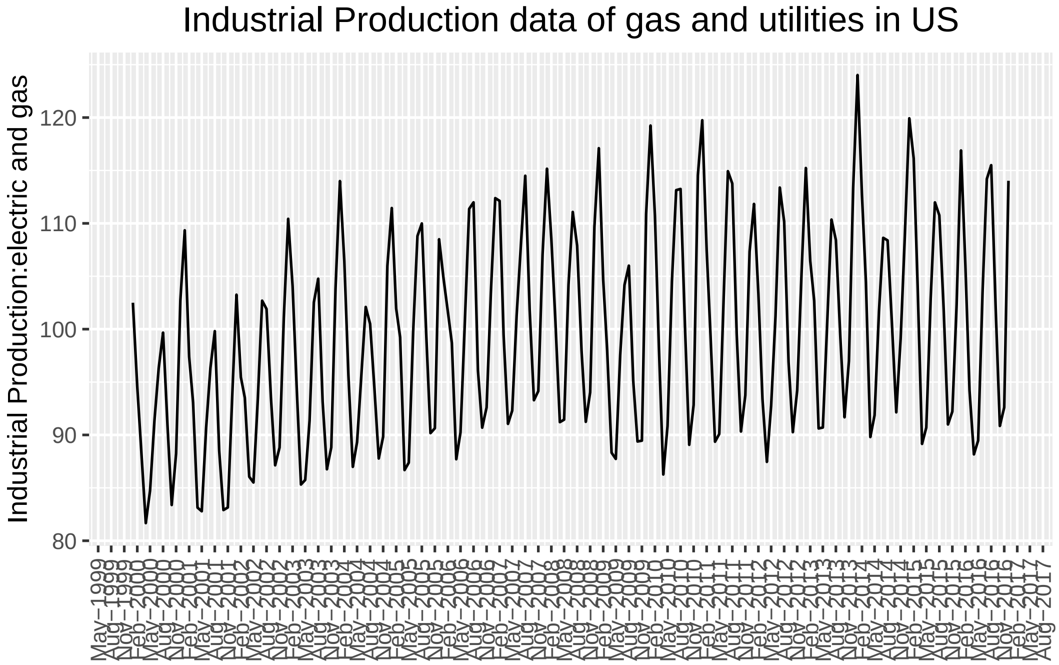

Figure 1: Time series plot

In order to see more clear trend and seasonal patterns, it is better to see the decomposition of the time series data. From `Figure` 2,overall trend is increasing and it has clearly seasonal pattern. Thus, we could use differencing method to remove both trend and seasonal features.

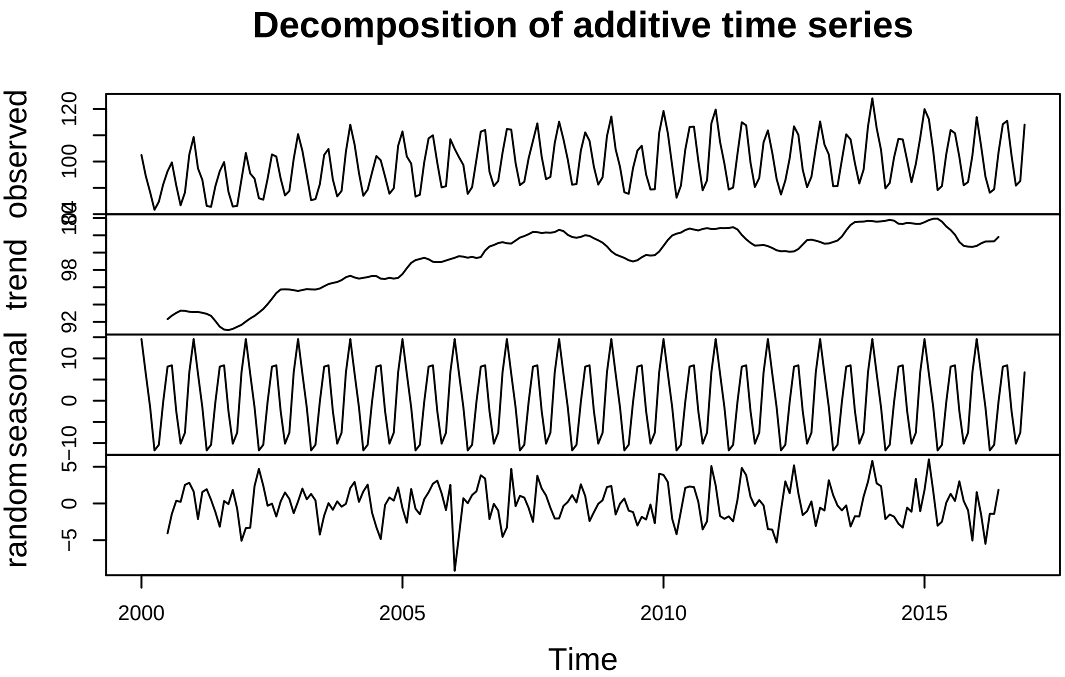

Figure 2: Data Decomposition

According to Figure 3,after using differencing method,the new data seems like stationary. In order to be accuracy, the p-value of Dickey-Fuller test is significant smaller than 0.05, which means we could reject the null hypothesis. Thus, the transformed data becomes stationary.

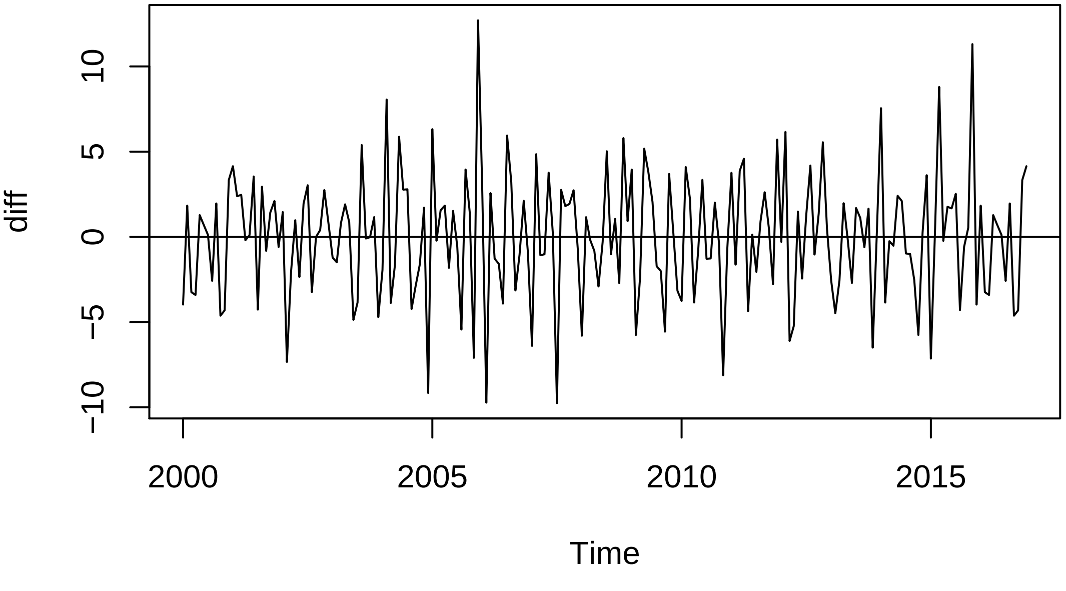

Figure 3: Transformed (differencing) data

2 Modelling

2.1 ARIMA Models

In terms of Figure 4,it is helpful for checking both ACF and PACF graphs. AR(2) model is a good choice because PACF plot cuts-off after lag 2 and ACF plot exponential decays to zero. Also MA(2) model could be useful because ACF plot cuts-off after lag 2, and PACF plot exponential decays to zero. In general, ARMA(2,2) is compatible because both ACF and PACF exponential decay to zero.

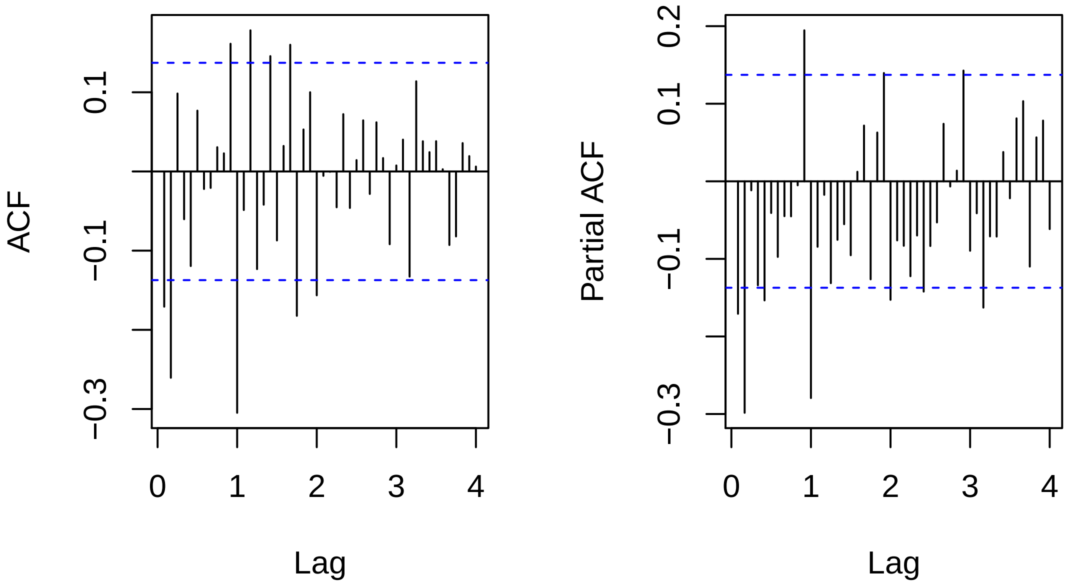

Figure 4: ACF and PACF Plots for Training Data

From `Figure` 5,the principal is to choose one model which has the smallest and clearest figure. Although the first one has the smallest BIC, the second one is clearer than the first one. Thus,the second one would be the optimal one to fit.

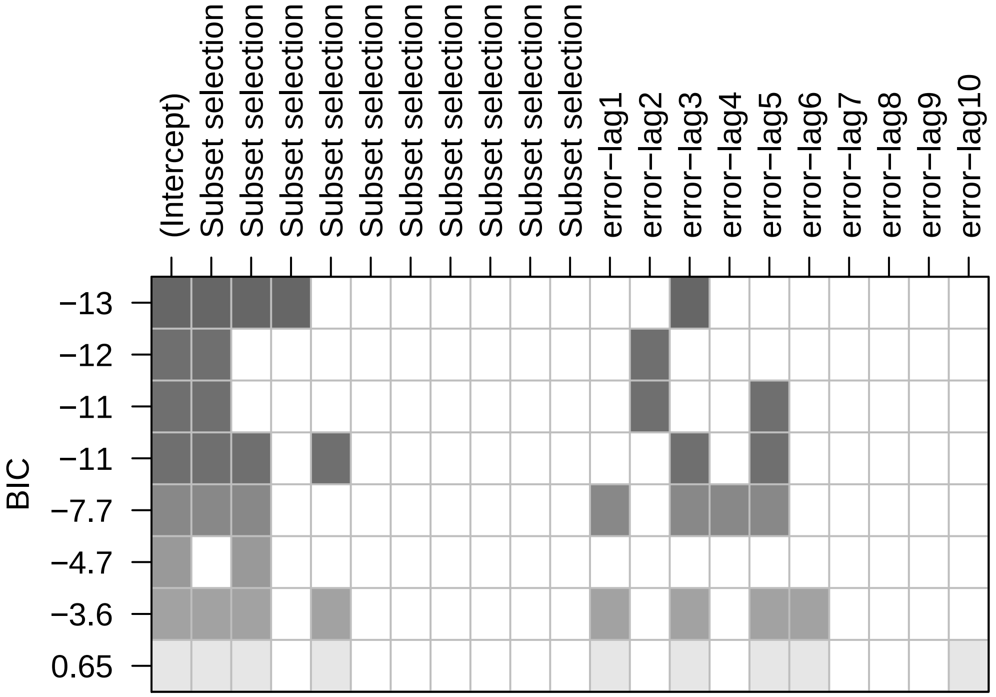

Figure 5: BIC plot of subset model selection

Since right now we have four ARIMA models, it is better to compare the AICc,and choose the model with the smallest AICc.

Table 1: Comparison of ARIMA models

Model	**Full Model**	**AICc**
ARIMA(2,1,0)	$X_t - 0.6574X_{t-1} + 0.7035X_{t-2} = e_t$	1306.1600154
ARIMA(0,1,2)	$X_t = e_t - 0.1790e_{t-1} - 0.7577e_{t-2}$	1347.7919325
ARIMA(2,1,2)	$X_t - 0.9676X_{t-1} + 0.8952X_{t-2} = e_t - 1.0854e_{t-1} + 0.2180e_{t-2}$	1170.92833
ARIMA(10,1,2)	$X_t + 0.2359X_{t-1} = e_t - 0.6283e_{t-2}$	1363.210731

2.1.1 Diagnostics of ARIMA(2,1,2)

In terms of diagnostics graph of ARIMA(2,1,2) model`Figure` 6, the residual distribution seems like randomly, and mostly lag of ACF residuals are not significant, which means the model is adequate enough to fit. But for the p-value for Ljung-Box statistic, most of points are below 0.05, which means it is not very good for out model fitting.

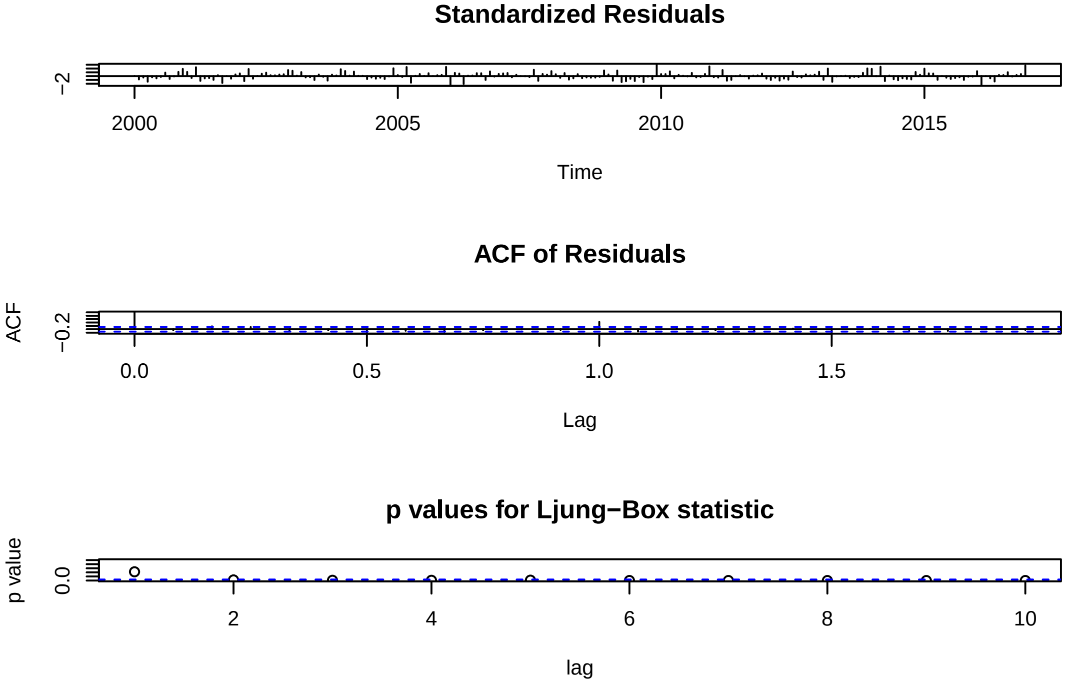

Figure 6: Diagnostics of ARIMA(2,1,2)

2.2 SARIMA models

For the SARIMA model part, in terms of ARIMA model should be same as before. Since ACF cuts off to zero after the first seasonal lag, while PACF at seasonal lags decay to zero. Thus, MA(1) model should be suitable for seasonal part.

Compare three SARIMA models, but ignore the parameters of these models because there are lots of parameters. From the AICc table, it is obviously that SARIMA(2,1,2)x(0,1,1)[12] is the best choise because of the smallest AICc.

Table 2: Comparison of SARIMA Models

Model	**AICc**
$\text{SARIMA}(2,1,0) \times (0,1,1)_{12}$	964.1645
$\text{SARIMA}(0,1,2) \times (0,1,1)_{12}$	946.6549
$\text{SARIMA}(2,1,2) \times (0,1,1)_{12}$	946.17

2.2.1 Diagnostics and Q-Q plot of SARIMA(2,1,2)(0,1,1)[12]

According to `Figure` 7 and `Figure` 8 , the residual distribution seems like randomly distribute, and most of lag in ACF plot are not significant. Meanwhile, the Q-Q plot shows that the residuals does not include much useful information. Thus,this SARIMA(2,1,2)(0,1,1)[12] model is adequate enough for fitting.

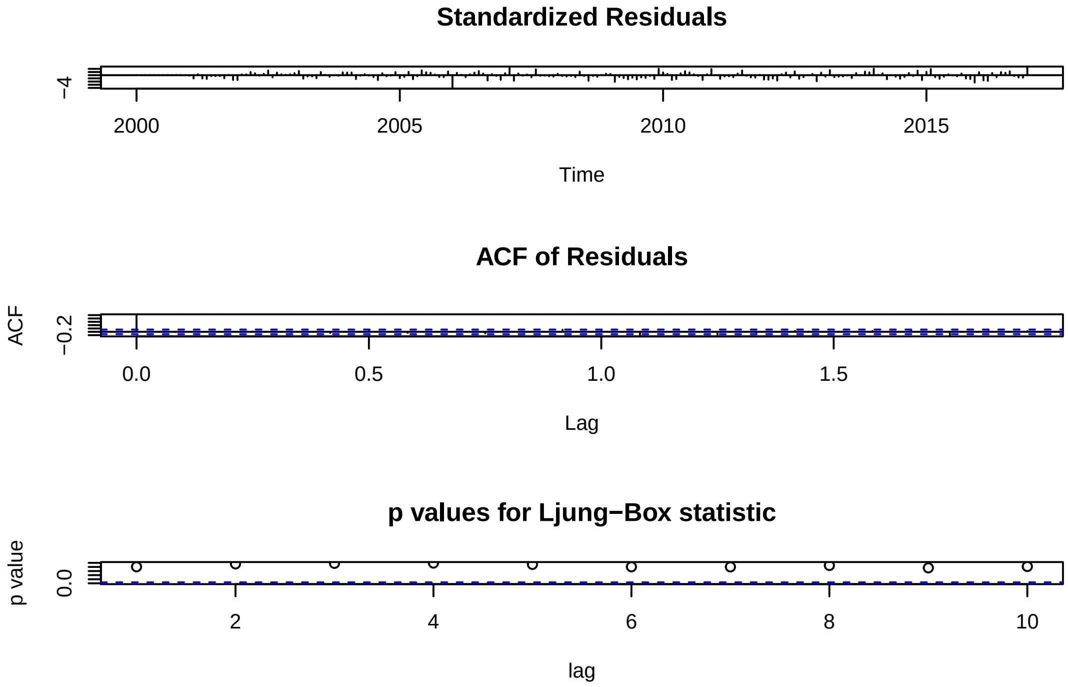

Figure 7: Diagnostics of SARIMA(2,1,2)(0,1,1)[12]

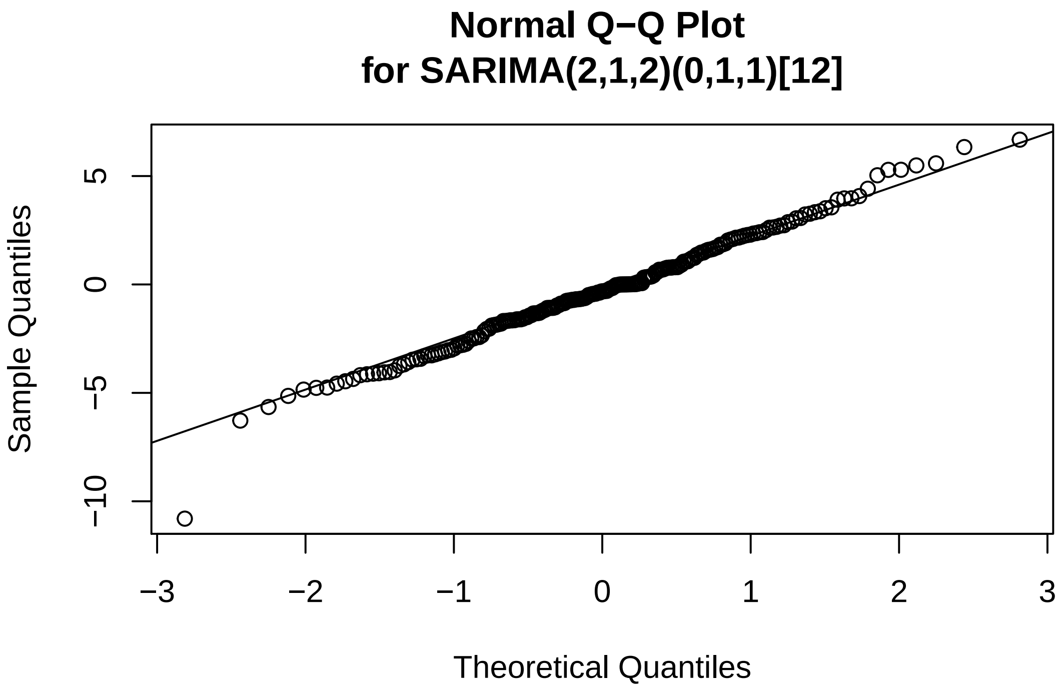

Figure 8: QQ plot of SARIMA(2,1,2)(0,1,1)[12]

3 Compare ARIMA and Non-ARIMA Models accuracy

Criteria	**ARIMA(2,1,2)**	**SARIMA(2,1,2)(0,1,1)[12]**	**Holt-Winters**
MAE	5.5304627	3.1400633	4.4896086
RMSE	6.9840571	4.1691231	5.5226673
MAPE	5.2210931	2.985482	4.390496

4 Forecasting

Since SARIMA(2,1,2)(0,1,1)[12] has the smallest MAE,RMSE and MAPE, therefore, SARIMA model is the most compatible for our project.

From `Figure` 9, it shows the prediction pattern of production of electric and gas. The overall pattern is same as before, it has both seasonal and trend fetures.

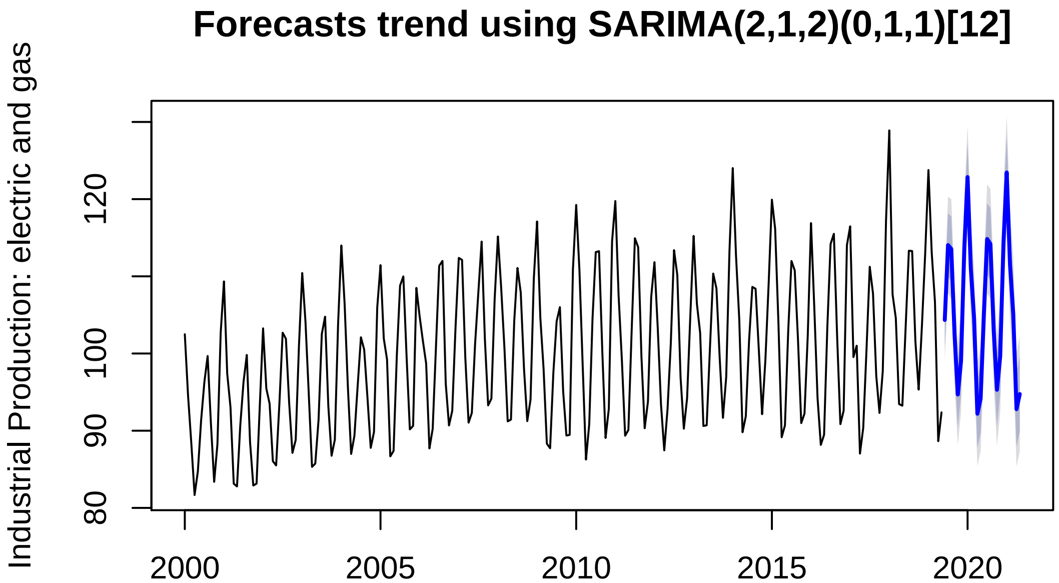

Figure 9: Forecast future trend of production

According to `Figure` 10, by comparing the actual data to the prediction data, it is clear to see SARIMA(2,1,2)(0,1,1)[12] works well, and the predicted values are in the 80% confidence intervals.

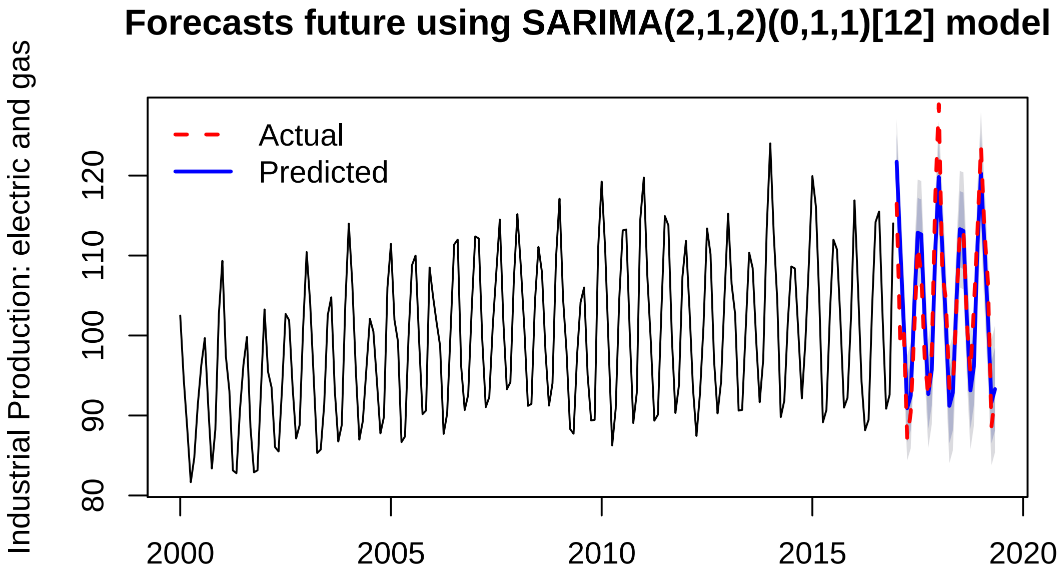

Figure 10: Compare actual and prediction

5 Reference

https://www.kaggle.com/sadeght/industrial-production-electric-and-gas-utilities