# Uber Demand in New York City: A Seasonal Time Series Data Analysis

Liang Hu

## 0. Summary

The project is titled *Uber Demand in New York City: A Seasonal Time Series Data Analysis*. The Uber demand data come from New York City Taxi & Limousine Commission. In R, we transform the dataset into a tsd object called uber.tsd. We identify the time series as a SARIMA(1, 1, 1)(0, 1, 1)$_7$ model, which performs well in diagnostic checking and predicting.

## 1. Introduction

Ride-hailing services such as Uber and Lyft are transforming the ways that people commute, especially in big cities. By simply clicking on the cellphones, users can request a cruising-around Uber car for a ride with much cheaper price than urban taxis. Just in the past a few years, Uber has become one of the important means of city transportation. We are curious about how the demand for Uber is changing over time.

The data about Uber requests in New York City were obtained from the New York City Taxi & Limousine Commission, who is responsible for managing this city's all types of taxis including Uber [NYC TLC, 2015]. This dataset recorded every Uber pickup that occurred in the city and collected the total number of Uber pickups per day.

Figure 1 shows the number of New York City Uber pickups (in thousand) by day from January 1, 2015 to June 30, 2015 (6 months, 181 observations). It is can be seen that there is seasonality with a period of about 7 days or 1 week. The demand for Uber during Friday, Saturday and Sunday is significantly higher than other weekdays. The reason for the phenomena might be that people have more recreational activities during weekends so that they do not want to or shall not drive their own cars (e.g., drinking alcohol).

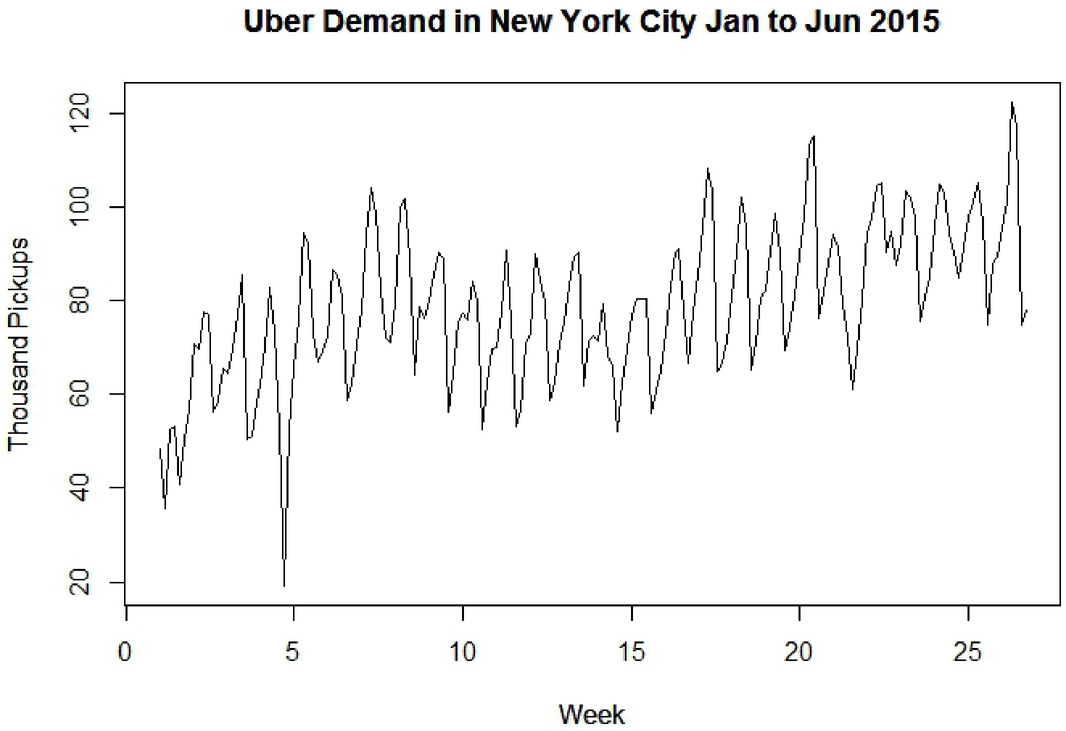


Figure 1. Uber demand in New York City between January and June 2015.

## 2. Tentative identification of the models

In this part, we use the *iden()* function to tentatively identify several possible models to fit the Uber demand time series. Figure 2 shows the results of identification without differencing and transformation. From the realization plot, it is can be seen that (1) the number of pickups in thousand has strong seasonality; (2) the mean changes over time but not significantly, while the variance seems to be constant; and (3) there is an outlier with a pretty low value during the 4th and the 5th week. The reason for the outlier might be bad weather such as heavy snow at the end of January, 2015. In the range-mean plot, the points cluster at the lower-right part, and there seems to be moderate correlation between the range and mean. Different transformation is then applied but we do not see significant change in the ACF and PACF of residuals. Therefore we decided not to use any transformation in the project. In the ACF plot, we can see significant spikes at lag 1, 7, 14, etc., and at their neighbors, indicating strong seasonality with a period of 7 days. In the PACF plot, there are some spikes standing out from lag 1 to lag 8.

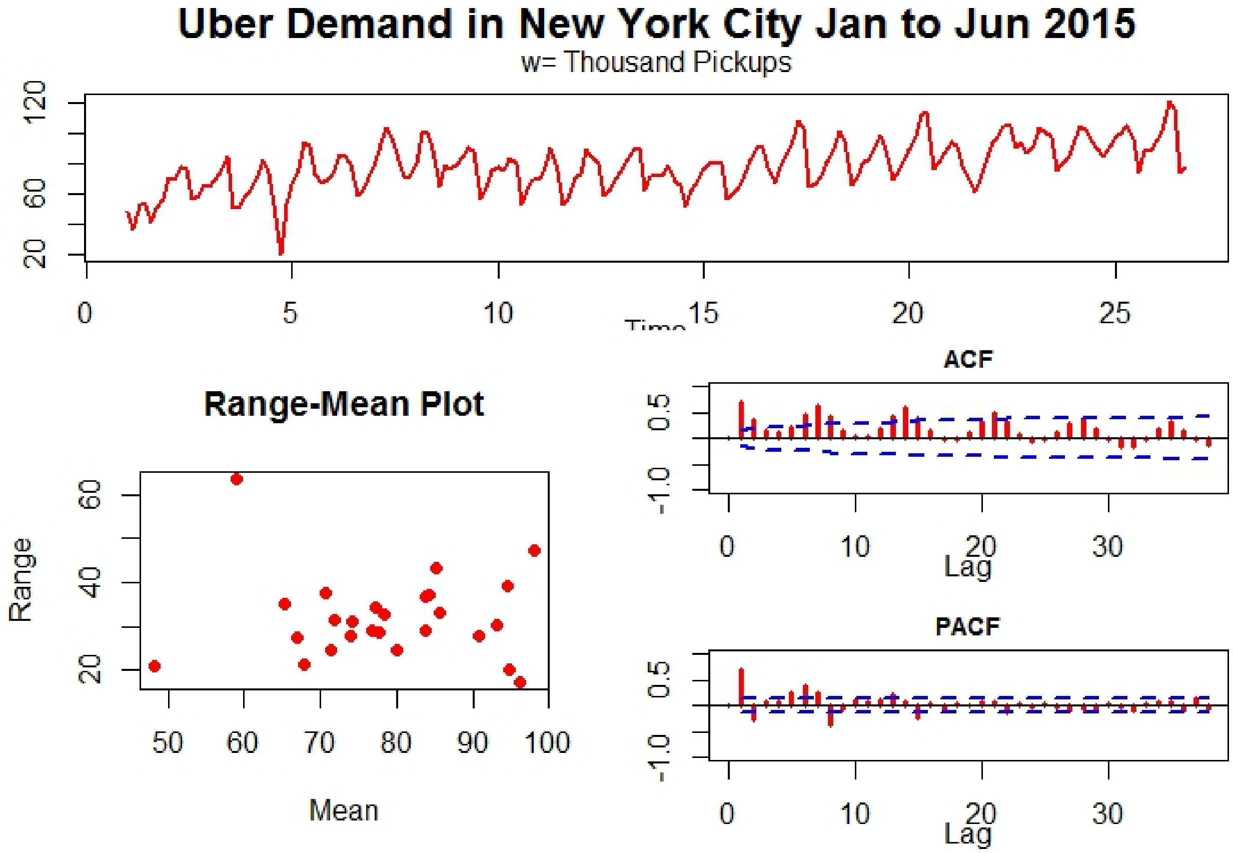


Figure 2. Identification of Uber demand time series without difference.

The second step is to take 1 regular difference (d=1). The results are displayed in Figure 3. The data look like more stable. In the ACF plot, the statistically significant spikes show at lag 7, 14, 21, etc., which indicates 1 seasonal difference might be needed. In each season, ACF at regular lags dies down. PACF at seasonal lags as well as at regular lags die down. Additionally, we can also see many significant spikes show up in the first season.

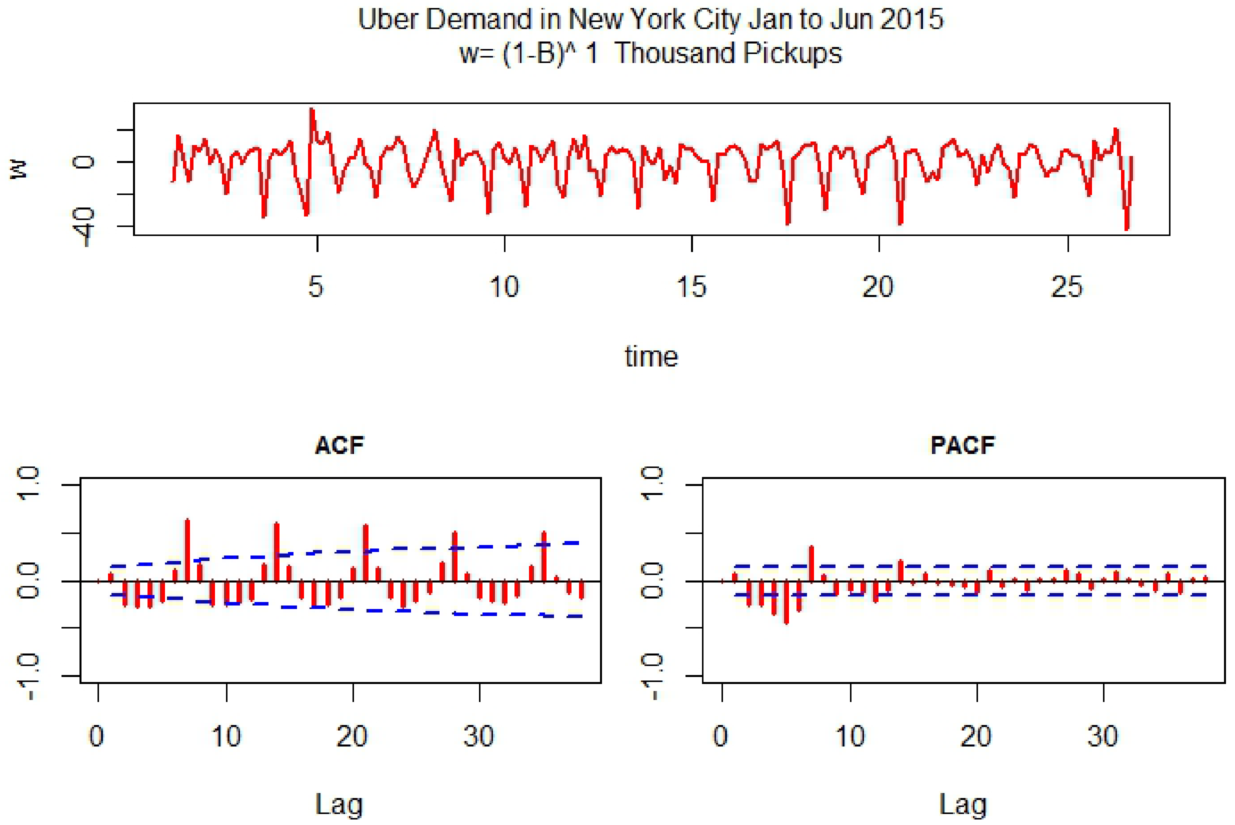


Figure 3. Identification of Uber demand time series with 1 regular difference.

The third step is to take 1 seasonal difference (D=1). From the results in Figure 4, we can see the time series is not stationary, because the variance tends to be smaller in the middle. In the ACF plot, the ACF at both regular lags and seasonal lags die down. In the PACF plot, PACF at regular lags cuts off after 1 lag, while PACF at seasonal lags dies down. Therefore we decide to try a **SARIMA(1, 0, 0)(1, 1, 1)$_7$** model to fit the time series.

The fourth step of identification is to take both 1 regular difference and 1 seasonal difference (d=1, D=1). Figure 5 shows that the time series is generally stationary, though the variance near the outlier is a little larger. ACF and PACF have significant improvement. In the ACF plot, the ACF at seasonal lags cuts off after 1 seasonal lag, but it is not easy to tell whether the ACF at regular lags is cutting off or dying down. In the PACF plot, the PACF at both seasonal lags and regular lags dies down. Therefore, a **SARIMA(1, 1, 1)(0, 1, 1)$_7$** model and a **SARIMA(0, 1, 1)(0, 1, 1)$_7$** model will be tried and compared.

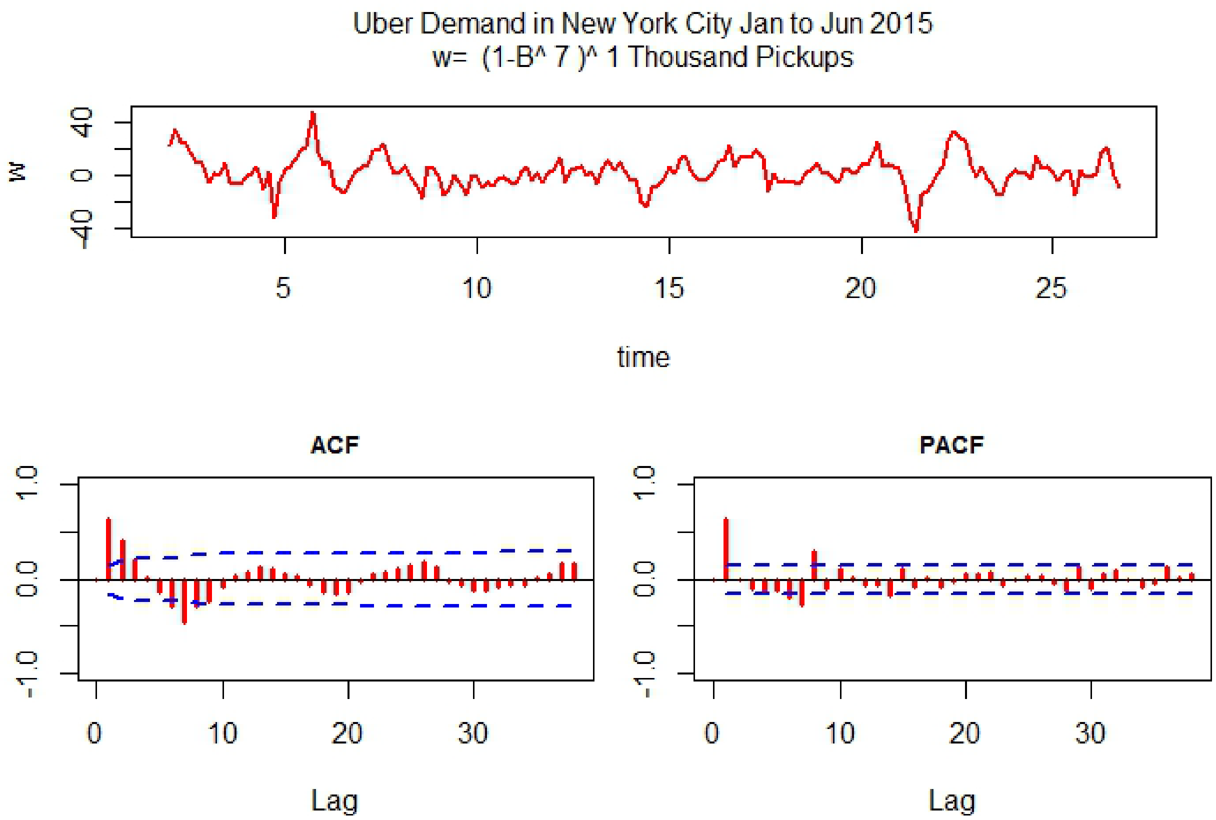


Figure 4. Identification of Uber demand time series with 1 seasonal difference.

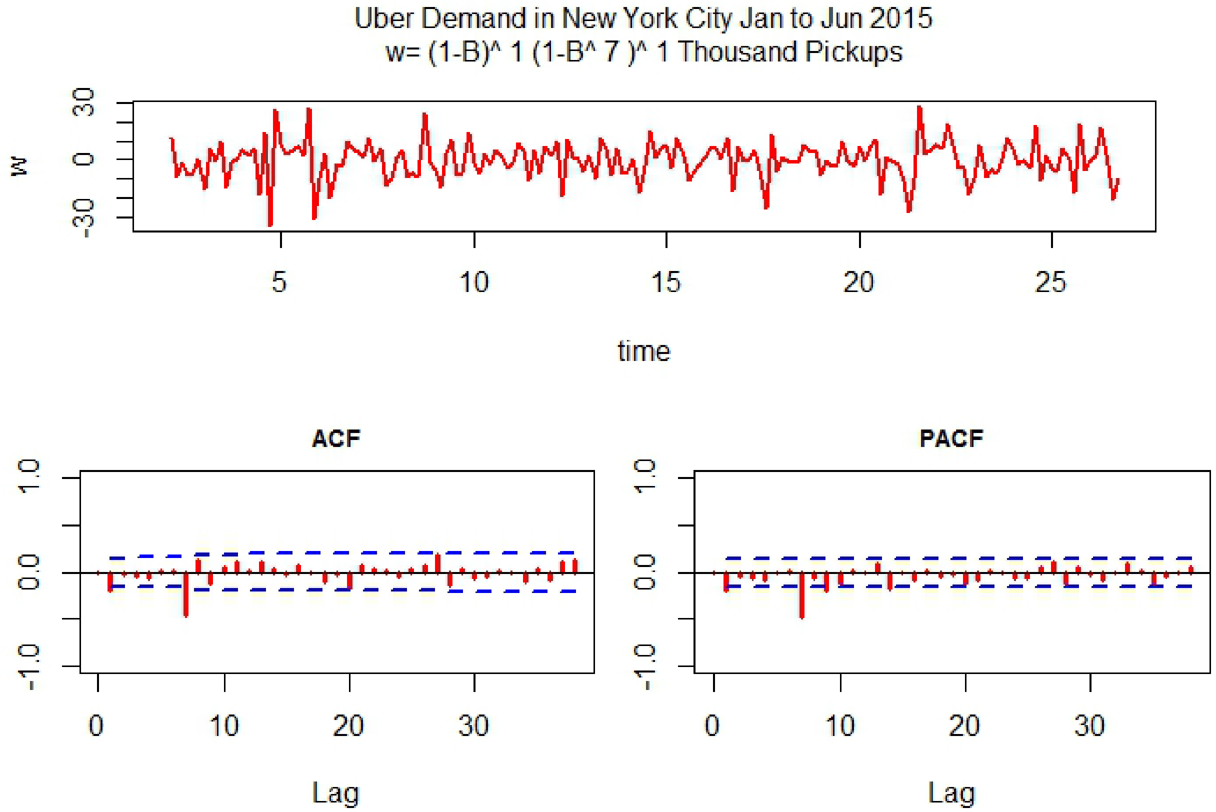


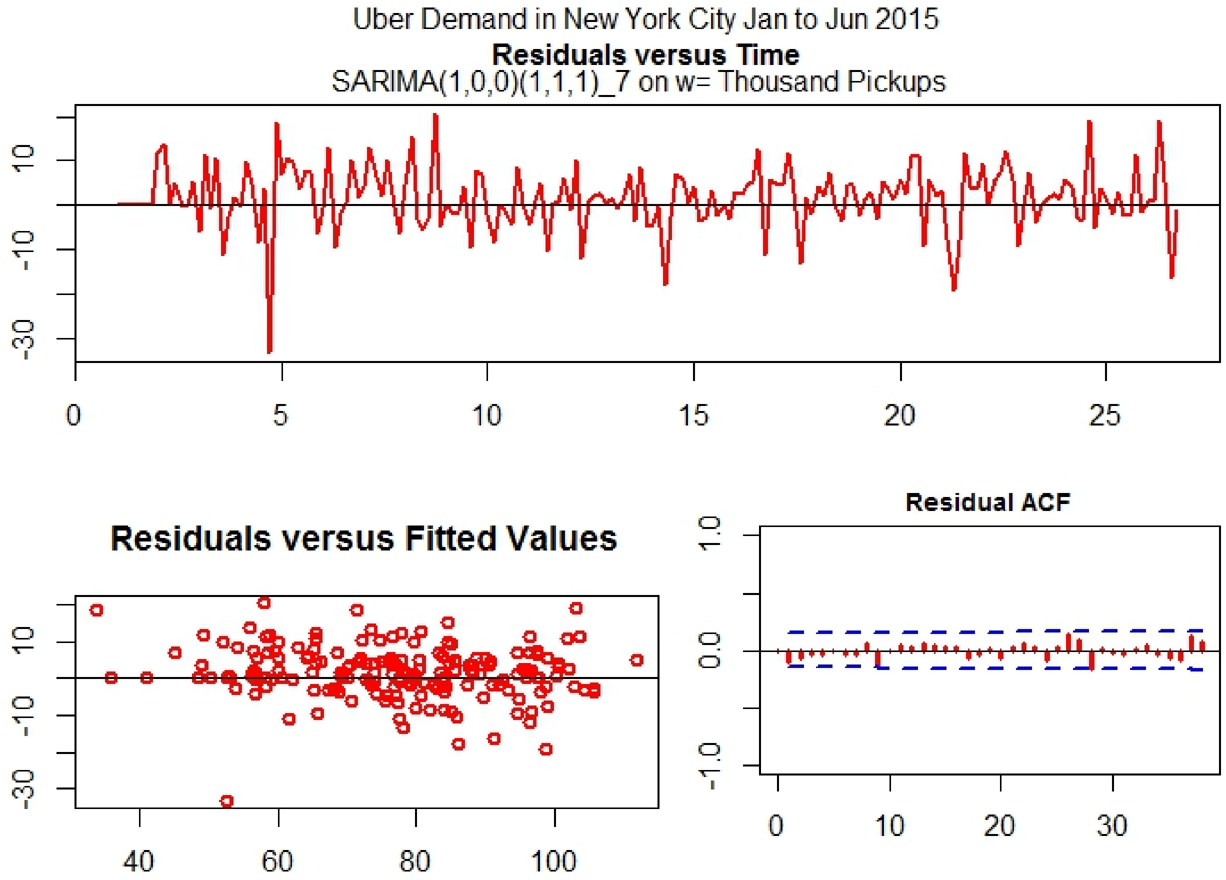
Figure 5. Identification of Uber demand time series with 1 regular difference and 1 seasonal difference.

## 3. Estimation of the models

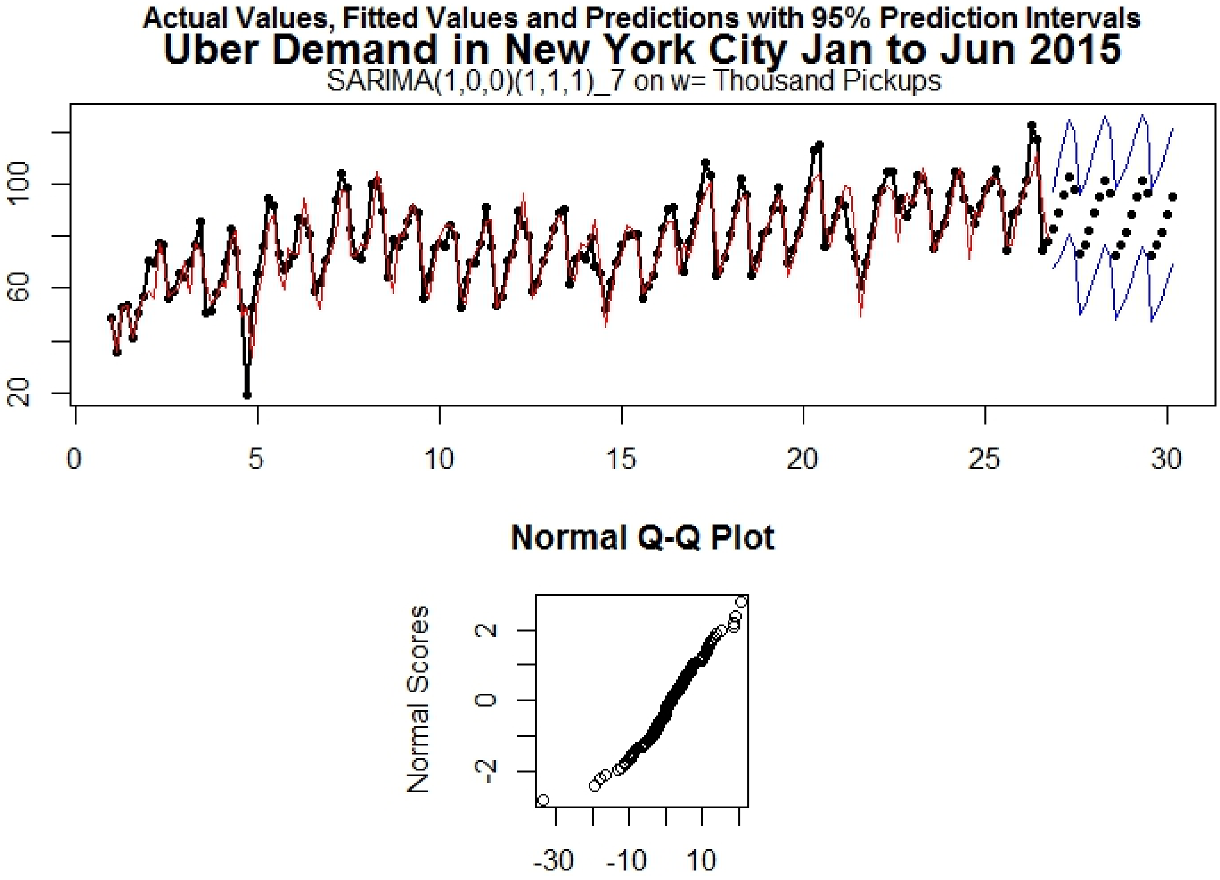
This section compares the estimation results of model 1 **SARIMA(1, 0, 0)(1, 1, 1)**$_7$, model 2 **SARIMA(1, 1, 1)(0, 1, 1)**$_7$, and model 3 **SARIMA(0, 1, 1)(0, 1, 1)**$_7$. The function esti() is used.

### 3.1. Model 1 SARIMA(1, 0, 0)(1, 1, 1)$_7$

In model 1 we only take 1 seasonal difference. The estimation results are given in Figure 6. From the plot of residuals versus fitted values, it can be seen that generally the residuals distributed randomly along the cross line, but on the left the residuals tend to locate above zero. The residual ACF looks not bad, but there are a few spikes at lag 9, 26 and 28 approaching the warning limit. From figure 6(b), the model generally fits the time series well, but some extreme values such as the very low outlier and a few larger observations are not well fitted. The given predictions tend to be along the average. Q-Q plot indicates the residuals do not follow the normal distributions very well.



(a)

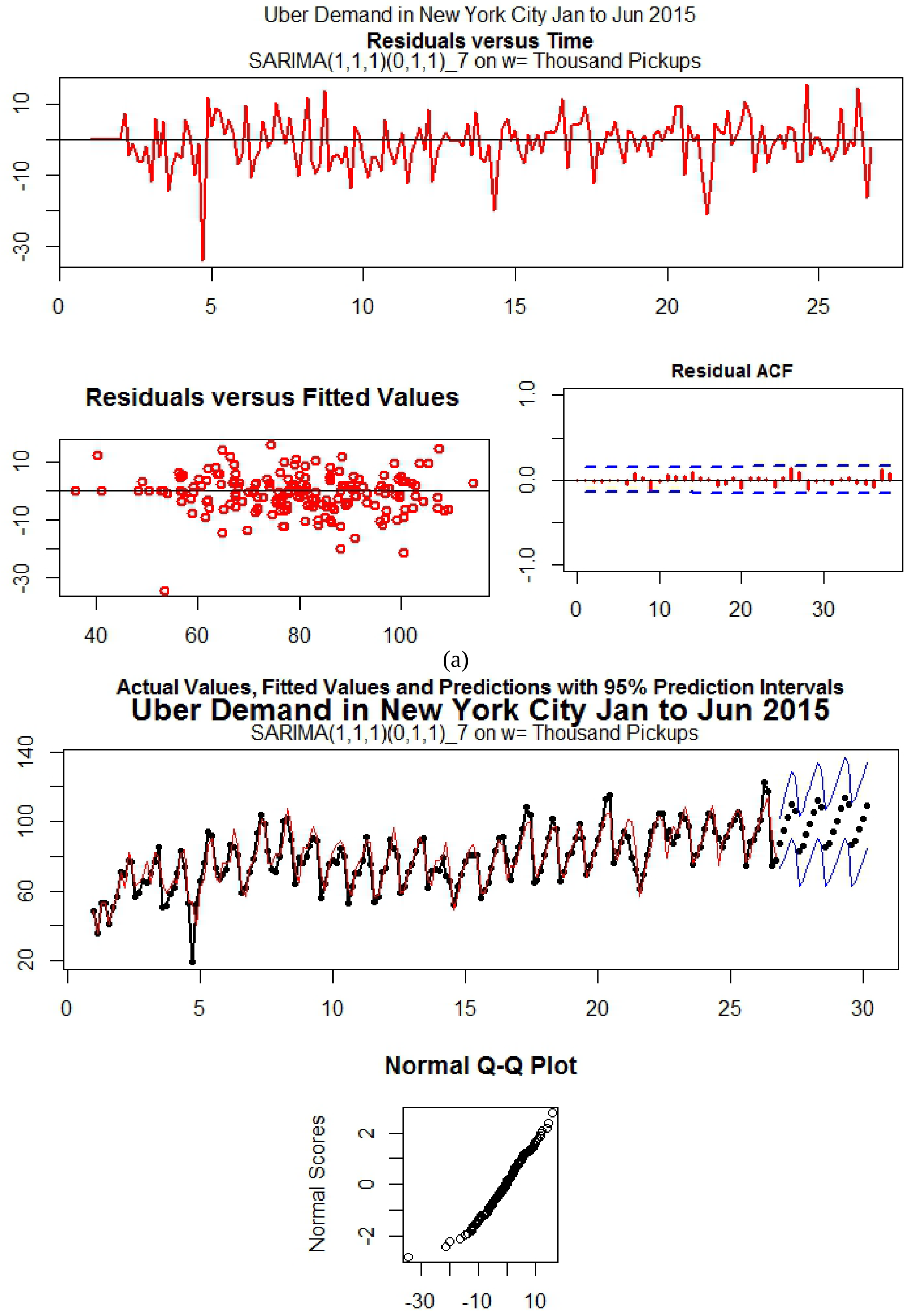


(b)

Figure 6. Estimation results of model 1.

### 3.2. Model 2 SARIMA(1, 1, 1)(0, 1, 1)$_7$

The estimation results of model 2 SARIMA(1, 1, 1)(0, 1, 1)$_7$ is showed in Figure 7. From the plot of residuals versus fitted values, we can see the residuals are randomly located along the line, though there is one residual with a very low value. This might be due to the outlier. The residual ACF looks clean. Similar to model 1, the spike at lag 9 is very close to the warning limit. By looking at Figure 7(b), the model fits the time series pretty well. There is an upward trend in the predictions, which is more realistic than the predictions of model 1, because more and more people are accepting shared mobility and choosing Uber. The residuals are normally distributed, indicated by the Q-Q plot.



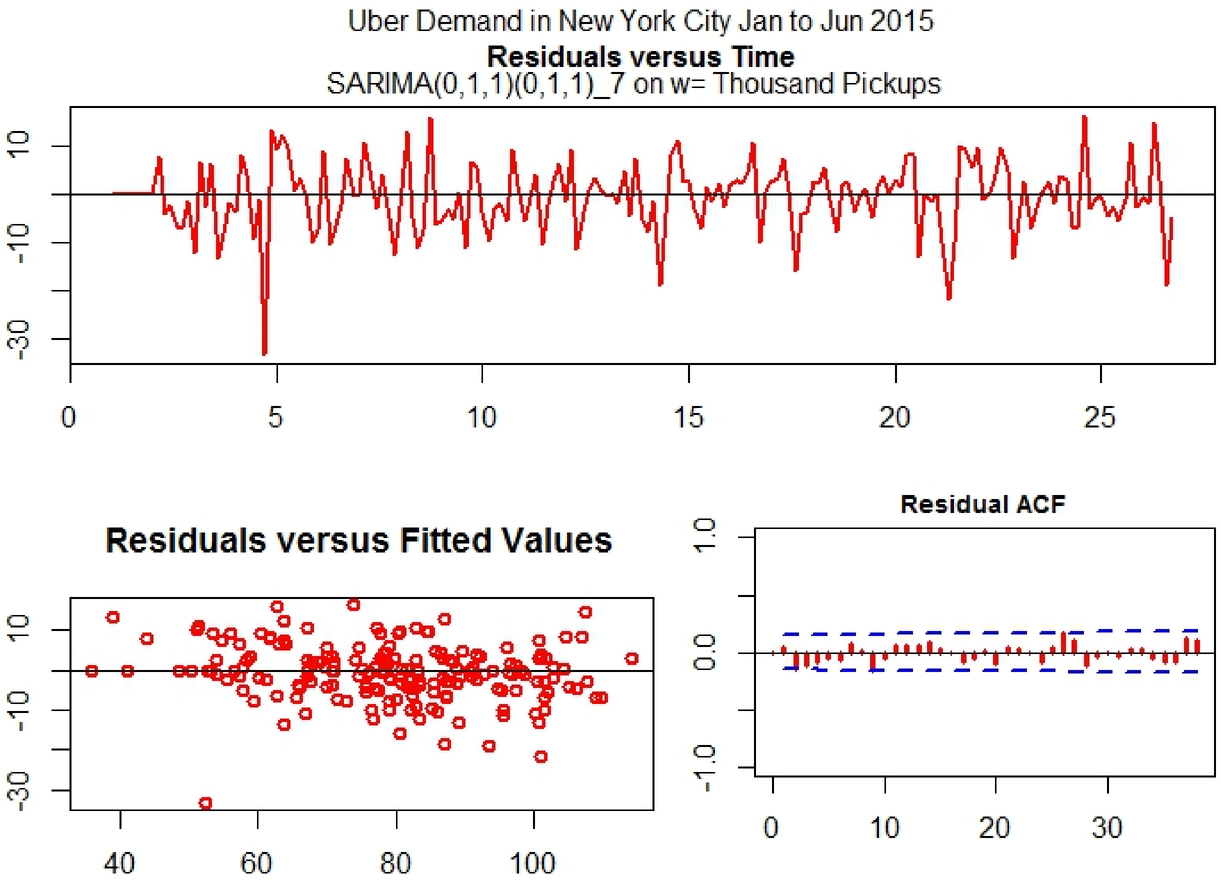
(a)

(b)

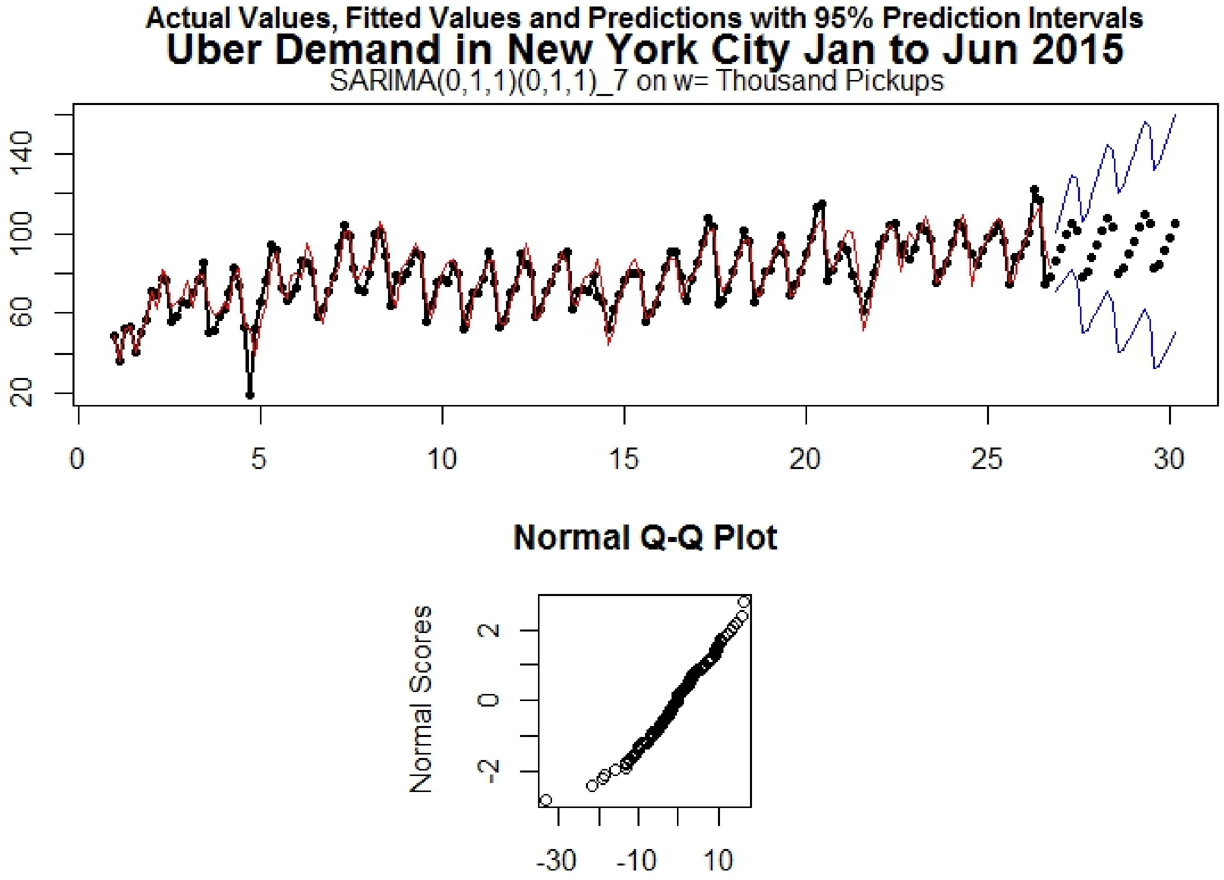
Figure 7. Estimation results of model 2.

### 3.3. Model 3 SARIMA(0, 1, 1)(0, 1, 1)$_7$

Model 3 has only moving average terms. The estimation results can be found in Figure 8. There seems to be a slightly downward trend in the plot of residuals versus fitted values, that is, more positive residuals on the left and more negative residuals on the right. The residual ACF looks not bad, but not as clean as model 2 residual ACF. Statistically residuals show up at lag 1, 9 and 26. From Figure 7(b), it is seen that model 3 does not fit the time series very well, especially on the left. The predictions are generally along the average, and the 95% prediction intervals become wider and wider. Q-Q plot shows the residuals follow normal distributions.



(a)



(b)

Figure 3. Estimation results of model 3.

3.4. Parameter estimations

Table 1 lists the estimated model parameters of the 3 models. No transformation is used. Model 3 is the most parsimonious model with 2 parameters. Model 2 has the smallest AIC, -2(Log Likelihood), S and Ljung-Box statistics. The 3 parameters of model 2 have large t-ratios.

Table 1. Estimations of model parameters.

| Parameters | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Transformation | None | None | None |
| Regular differences d | 0 | 1 | 1 |
| Seasonal differences D | 1 | 1 | 1 |
| $\phi_1$ | 0.78<br>(12.9) | 0.61<br>(7.0) | — |
| $\theta_1$ | — | 0.92<br>(18.0) | 0.29<br>(3.0) |
| $\Phi_1$ (Seasonal) | 0.07<br>(0.7) | — | — |
| $\Theta_1$ (Seasonal) | 0.86 | 1.00 | 1.00 |

| | | | |
|---|---|---|---|
| | (11.8) | (9.9) | (11.7) |
| AIC | 1211.228 | 1197.564 | 1209.603 |
| -2(Log Likelihood) | 1203.228 | 1189.564 | 1203.603 |
| S | 7.48 | 7.01 | 7.34 |
| # parameters | 3 | 3 | 2 |
| Ljung-Box $\chi^2$ (20dof) | 13.23 | 12.33 | 23.20 |
| p-value for L-B | 0.87 | 0.90 | 0.28 |

### 3.5. Determination of models

Compared with model 1 and model 3, model 2 SARIMA(1, 1, 1)(0, 1, 1)$_7$ has better residuals versus fitted values, cleaner residual ACF, better fitting results. In addition, the predictions have an upward trend which is more realistic. Model 2 also performs better in diagnostic checking and has statistically significant parameters. Therefore, we decide to use model 2 to fit the Uber demand time series.

## 4. Forecasting

This section will forecast and compare the number of pick-ups one step ahead. Model 2 best fits the time series, and model 1 comes the next.

The unscrambled format of model 2 SARIMA(1, 1, 1)(0, 1, 1)$_7$ is

$$Z_t = 1.61Z_{t-1} - 0.61Z_{t-2} + Z_{t-7} - 1.61Z_{t-8} + 0.61Z_{t-9} - 0.92a_{t-1} - a_{t-7} + 0.92a_{t-8} + a_t \quad (1)$$

The time series has totally 181 observations. We want to forecast one step ahead, which is $Z_{182}$, from the forecast origin 181. Therefore,

$$\hat{Z}_{181}(1) = 1.61Z_{181} - 0.61Z_{180} + Z_{175} - 1.61Z_{174} + 0.61Z_{173} - 0.92[a_{181}] - [a_{175}] + 0.92[a_{174}] + [a_{182}]$$

= 87.396 thousand pick-ups (2)

Also, $Variance = \sigma_a^2 = 46.42$ (3)

Therefore the 95% confidence interval for $\hat{Z}_{181}(1)$ is $87.396 \pm 1.96 \times \sqrt{46.42} = [74.042,100.75]$.

The unscrambled format of model 1 SARIMA(1, 0, 0)(1, 1, 1)$_7$ is

$$Z_t = 0.78Z_{t-1} + 1.07Z_{t-7} - 0.835Z_{t-8} - 0.07Z_{t-14} + 0.0546Z_{t-15} - 0.86a_{t-7} + a_t \quad (4)$$

The forecast of one step ahead $Z_{182}$ is

$$\hat{Z}_{181}(1) = 0.78Z_{181} + 1.07Z_{175} - 0.835Z_{174} - 0.07Z_{168} + 0.0546Z_{167} - 0.86[a_{175}] + [a_{182}]$$

=82.522 thousand pick-ups (5)

Also, $Variance = \sigma_a^2 = 51.40$ (6)

Therefore the 95% confidence interval for $\hat{Z}_{181}(1)$ is $82.522 \pm 1.96 \times \sqrt{51.40} = [68.470, 96.574]$.

In summary, at one step ahead, model 2 has larger prediction and slightly narrower confidence interval than model 1.

## 5. Conclusions

This seasonal project analyzes the time series of Uber demand in thousand pick-ups that occurred in New York City from January to June 2015. It is found the time series has seasonality of 7 days, and the demands on Fridays and weekends are higher than the other days. Three seasonal models are estimated and compared, and the results show that the SARIMA(1, 1, 1)(0, 1, 1)$_7$ model taking no transformation, 1 regular and 1 seasonal difference best fits the time series. At one step ahead prediction, this model has larger prediction and slightly narrower confidence interval than the SARIMA(1, 0, 0)(1, 1, 1)$_7$ model.

Data source:

NYC Taxi & Limousine Commission. https://github.com/fivethirtyeight/uber-tlc-foil-response.