

**Stat 451 Seasonal Project**

Air Passengers Analysis Report

Sijie Huang

Tianchen Wang

Shangkun Zuo

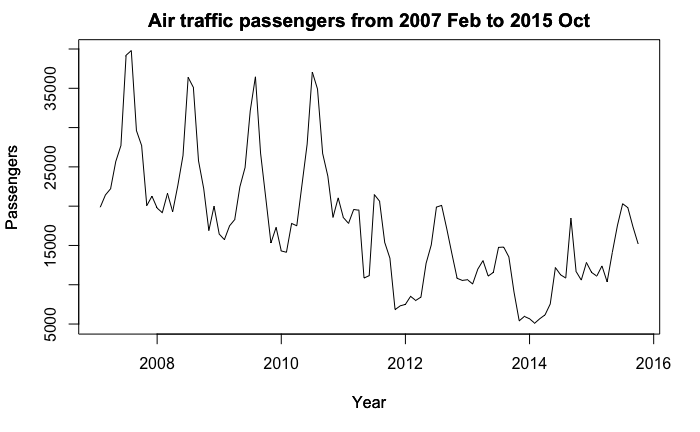
Executive Summary

The following is an analysis of Air Passengers. Since we want the data more professional, we use the home of the U.S Government’s open data to find data and the publisher is data.sfgov.org. Because the air traffic passengers are easily affected by seasons. Most time, we do know in the summer and winter vacation time, more people prefer to travel therefore there are likely more air passengers during this time. We only focus on air traffic passengers through time change, but the original dataset is very messy, we cut the dataset by picking one air company in a specific city in two specific terminals, thus we choose Air Canada in San Francisco in International Terminal A and Terminal E. There does not exist any missing observations in our dataset, each count of the air passengers is just for one terminal in one city in one month, so the amount it not big, and our total period is not that long, so after the filtering, we don’t need to rescale, and the thousands unit is not that much effect our output, therefore we assume the current length of the realization we use is great.

Our purpose of this time series model is to analyze the air traffic passengers amount change from 2007 February to 2015 October., and trying to predict the future trend. Since we want to choose a better fit model, we need to use d=1, D=1 for model SARIMA (12,1,1) (0,1,0)\_12 and gamma=0 (the log transformation).

**Table of Contents**

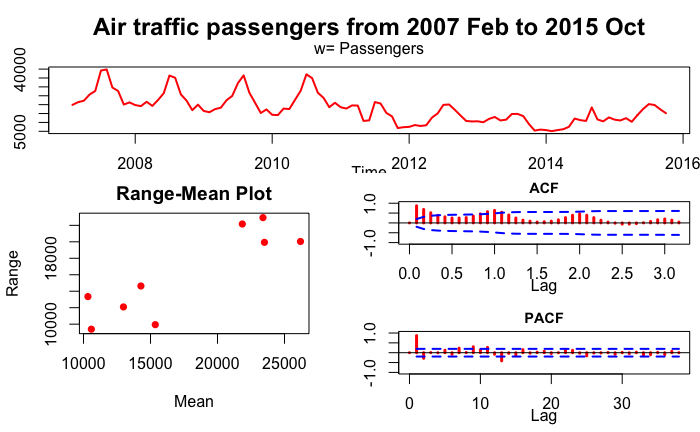
1. Data Description



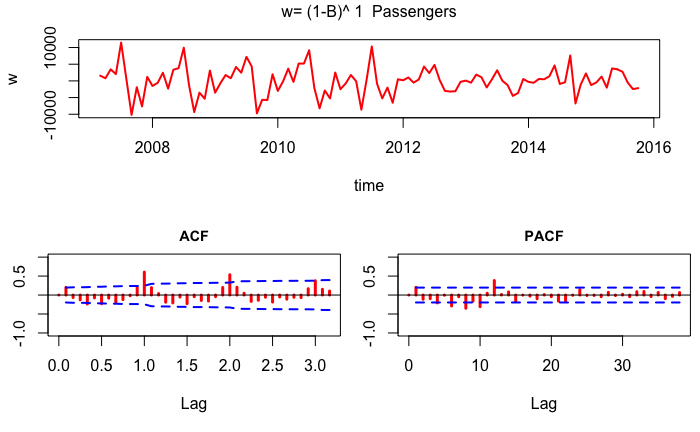
Overall, for our time series plot, there is a seasonal trend with about one peak per year, which is in summer time. Moreover, the distance between each peak is almost same. One problem we found in the plot is that the total passengers in 2012-2014 is less than other years, which could due to one crush occurred that year in San Francisco, also because these flights in SFO are usually delayed. Thus, the number of passengers went down in recent years may because the concern of safety and stand-by time.

1. Choosing a Model

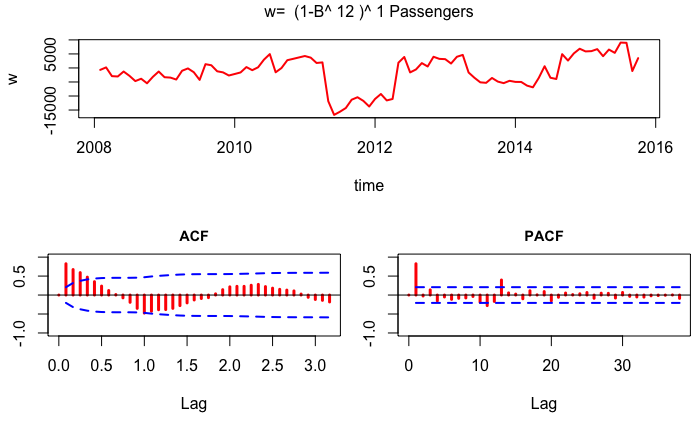
Original iden function



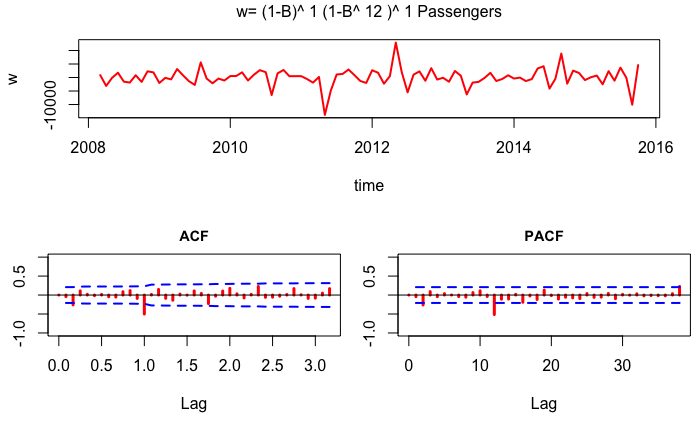
iden d=1



iden D=1

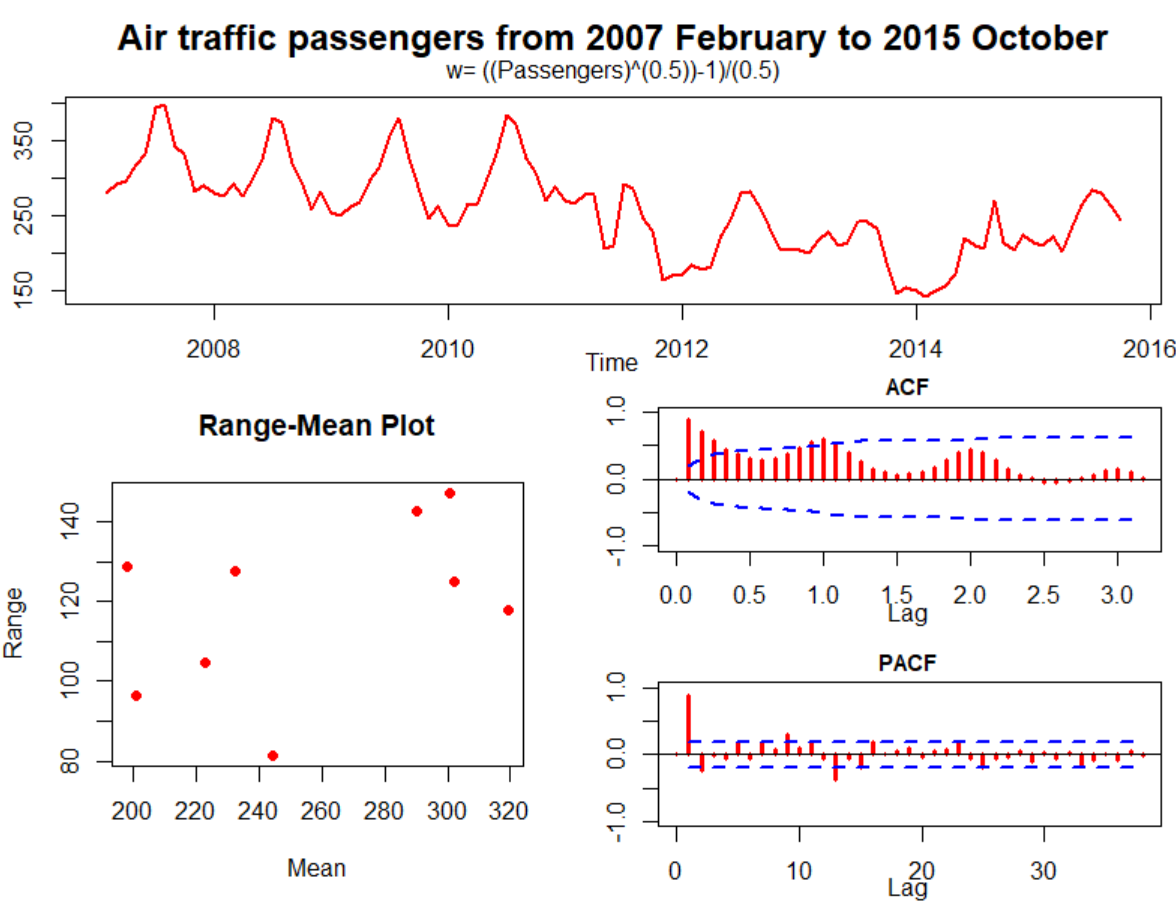


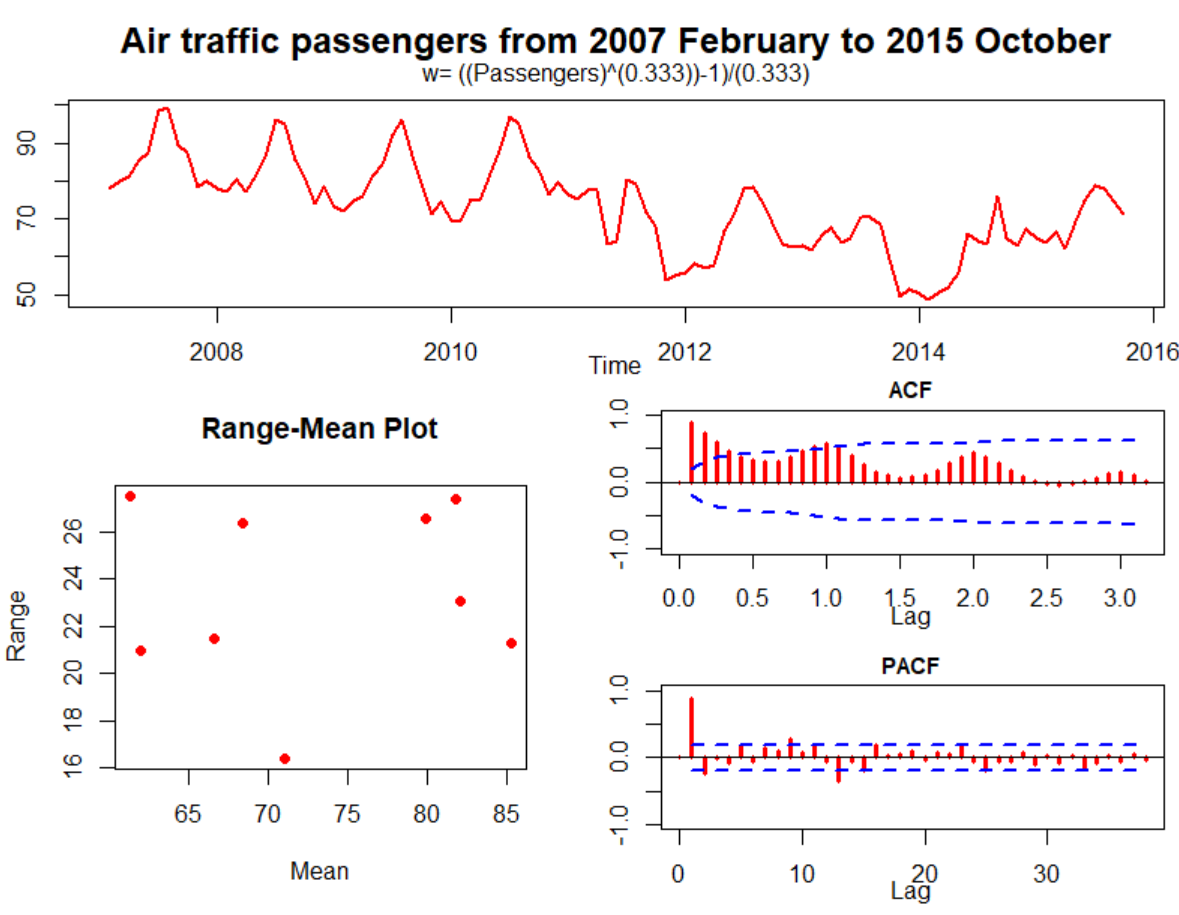
iden d=1, D=1

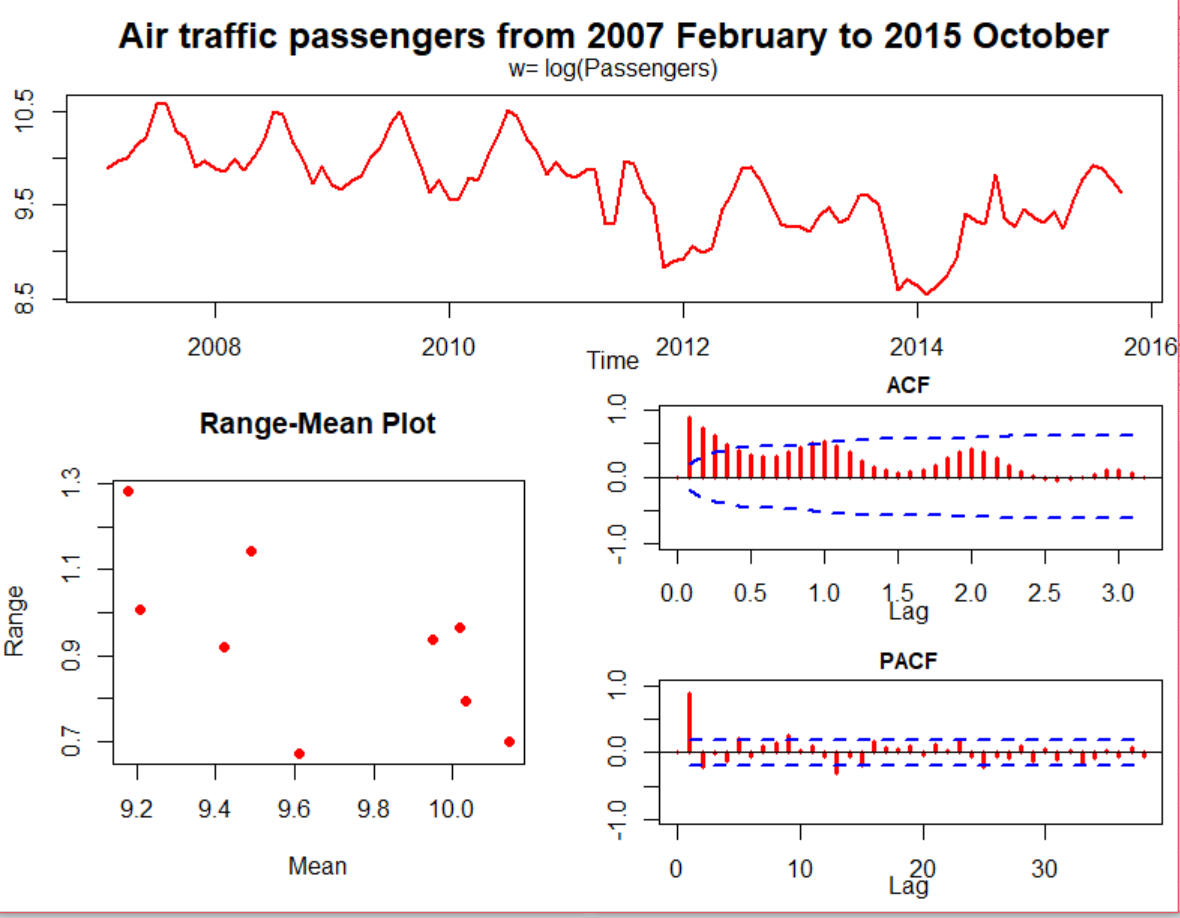


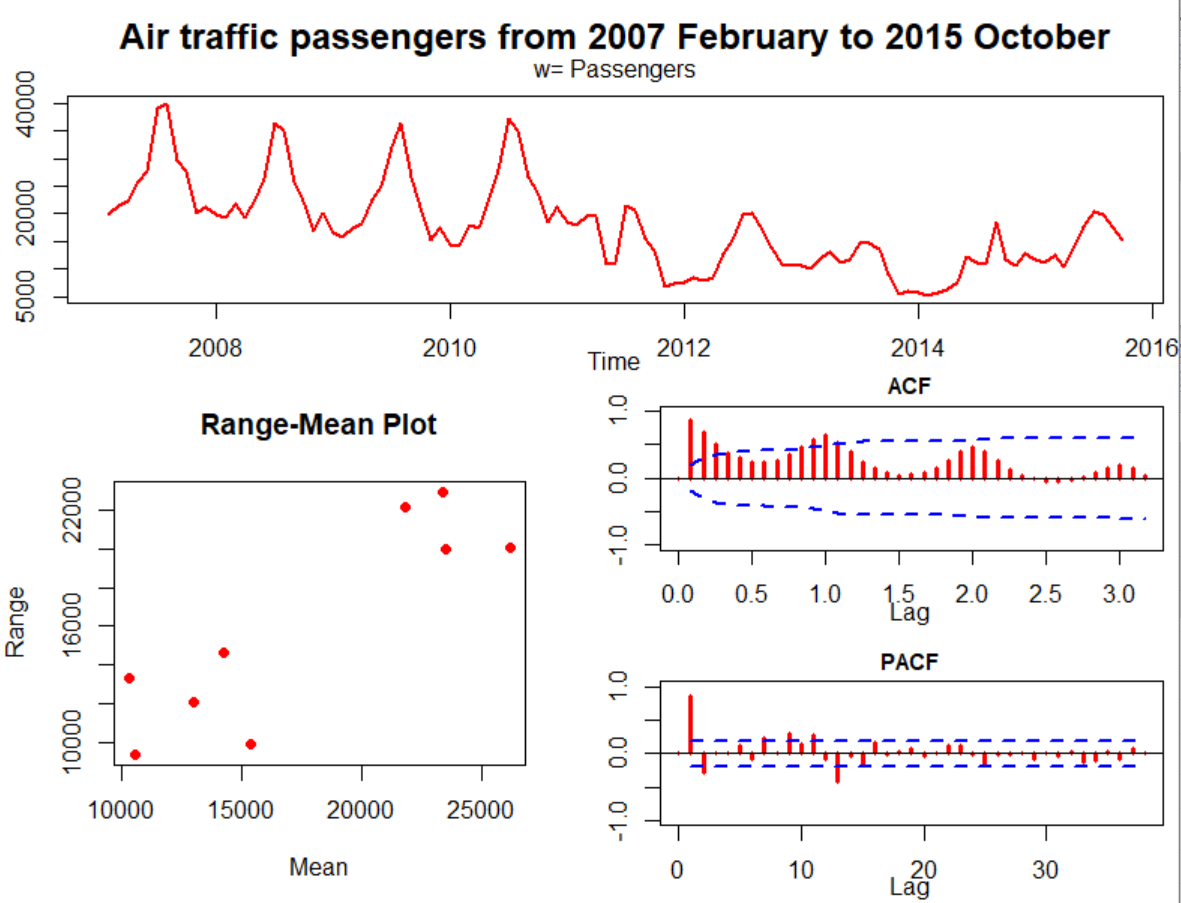
From the three iden functions, it is clear that we need to take both seasonal and non-seasonal difference. There are many spikes appeared when we take non seasonal difference=1 on both acf and pacf plots. Moreover, a strong autocorrelation showed in acf plot when we only take the seasonal difference. If we take both d=1, D=1, it appears that there is only one spike on acf and pacf plot, the rest of lags are pretty clear. Therefore, d=1 and D=1 would be a better choice compared to the other two.

Check for Transformation

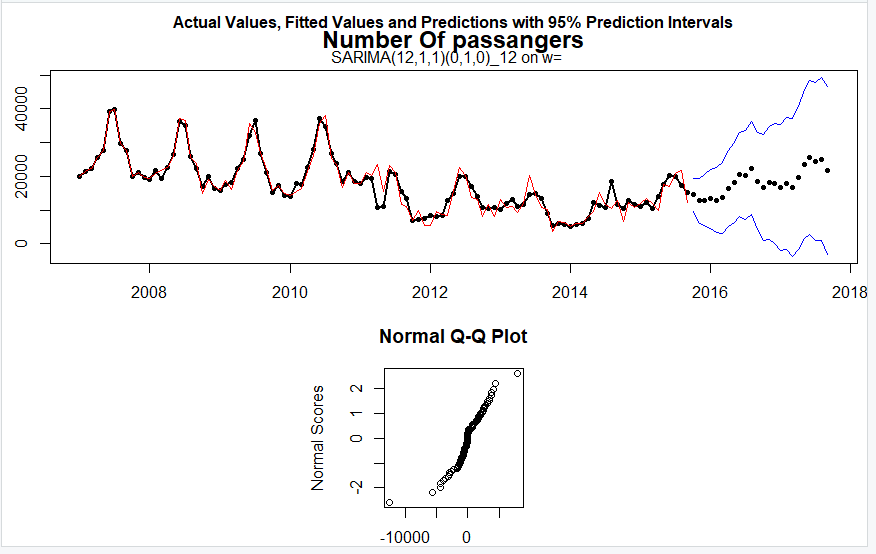
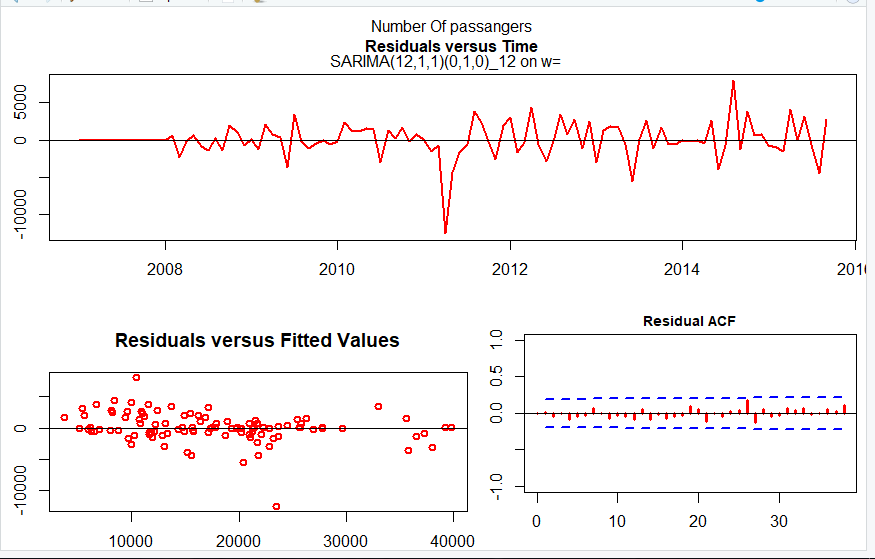


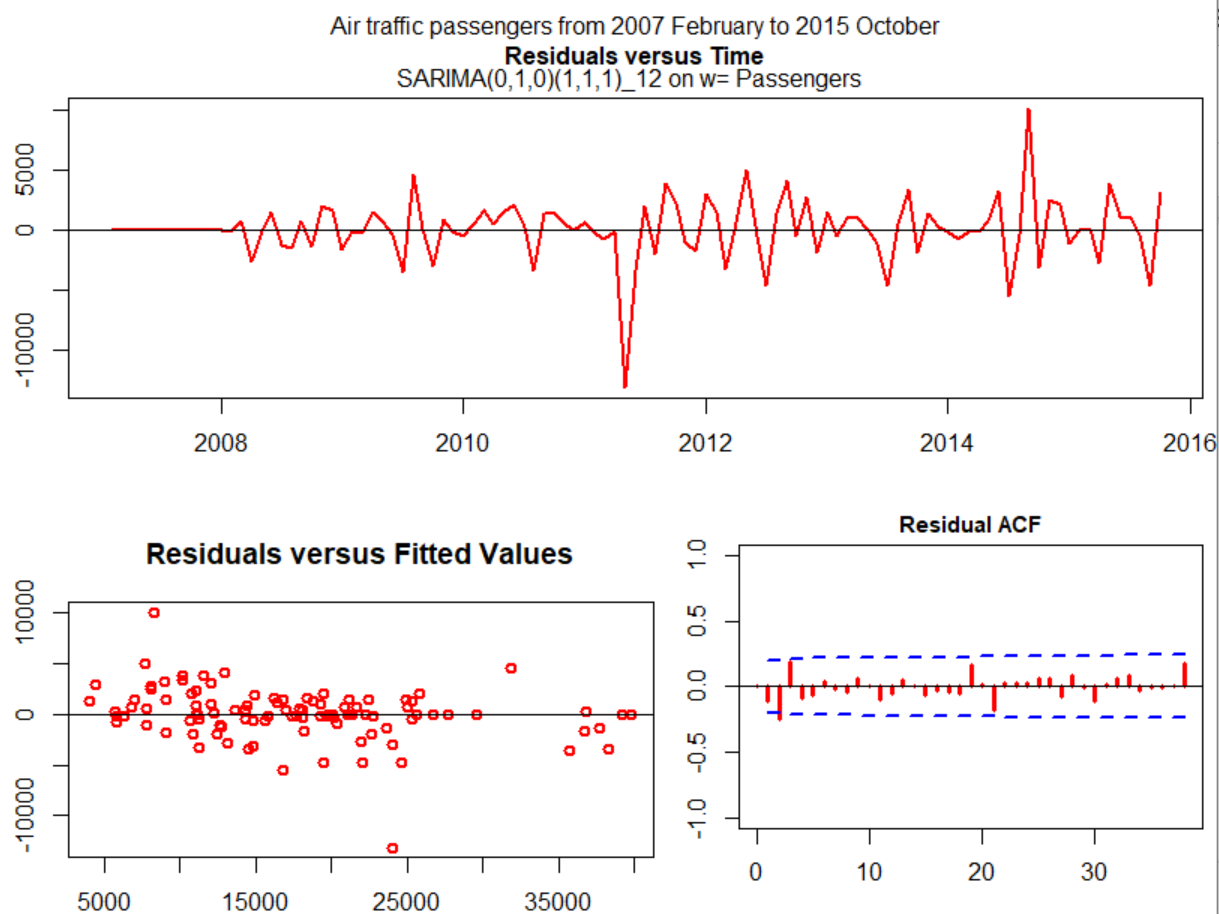


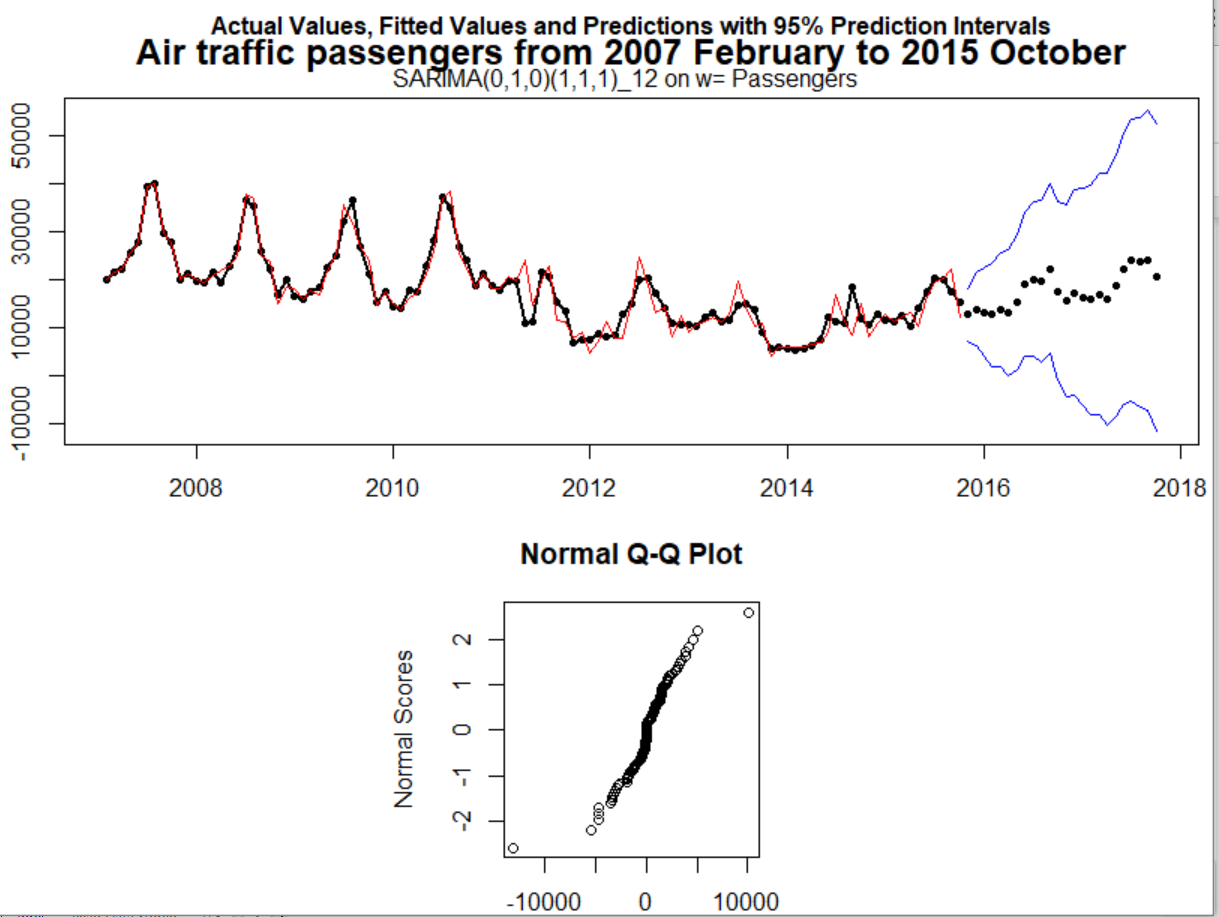




We use iden() to make the range plot to determine if we need to use transformation. Firstly, it’s plot without any transformation. In the range-man plot, the range and mean are pretty big and the plot are gathering in the lower y and higher y-axis. With the log transformation, the range and the mean become much more small, and the pattern are randomly distributed. With the square root transformation, there’s a similar pattern as reciprocal cube-root, which is random distributed and without exact pattern. But the range of reciprocal cube is smaller than the square root transformation. Compare to no transformation, log transformation and the reciprocal cube transformation, using log transformation(gamma=0) is better.

esti() Graphical Output and Forecast for Models 





We use esti() function fit 5 models, which are SARIMA (0,1,12) (0,1,0), SARIMA(0,1,0) (1,1,1), SARIMA (12,1,1,) (0,1,0), SARIMA (2,1,2) (0,1,0), and SARIMA (1,1,1) (0,1,0). By simply looking at the plots we have, SARIMA (0,1,0) (1,1,1) model is likely the best model we found so far. There is a significant spike in SARIMA (2,1,2) (0,1,0). In SARIMA (12,1,1) (0,1,0), SARIMA (0,1,12) (0,1,0), the fitted value versus residual value plot do not have a normal distributed pattern, which indicates a violation of constant variance. In our chosen model, SARIMA (0,1,0) (1,1,1), there is no spike in residual acf plot, and the residual versus fitted values also tend to be normal distributed, which indicates the constant variance is met. We can see a slight curve in QQ plot even it’s better than the SARIMA (12,1,1) (0,1,0), however, we can ignore it since the sample size is fairly large. Also, the forecasting is almost fit the trend.

Table to Compare Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SARIMA(2,1,2)(0,1,0) | SARIMA(12,1,1) (0,1,0) | SARIMA(1,1,1)(0,1,0) | SARIMA(0,1,0)(1,1,1) | SARIMA(0,1,12)(0,1,0) |
| box-cox(gamma) | 1 | 1 | 1 | 1 | 1 |
| Regular differences | d=1 | 0 | 0 | 1 | 1 |
| Seasonal differences | 1 | 1 | 1 | 1 | 1 |
| Ф1 | -0.430 | 0.09 | 0.843 |  |  |
| θ1 | -0.426 | 0.22 | 0.999 |  | 0.08 |
| θ1(seasonal) | 0 | 0 | 0 | -0.425 | 0 |
| Sig. pk(a) | 1 | 0 | 1 | 0 | 1 |
| S | 3122.221 | 2524.776 | 3144.307 | 2723.147 | 2634.613 |
| AIC | 1751.847 | 1735.386 | 1751.081 | 1726.698 | 1742.41 |
| -2log(Likelihood) | 1741.847 | 1707.386 | 1745.081 | 1720.698 | 1716.41 |
| Ljung-Box x26 | 2.192 | 6.783 | 7.018 | 13.86 | 12.89 |

By looking at the table, there are five reasonable models we have so far. By picking the best model overall, we first eliminate the model SARIMA (2,1,2) (0,1,0), SARIMA (1,1,1) (0,1,0), and SARIMA (0,1,12) (0,1,0) since they all have at least one significant spike, and the model SARIMA (2,1,2) (0,1,0) does not have a significant Ljung-box statistic (2.192). Moreover, in SARIMA (0,1,12) (0,1,0) model, the value of S does not improve much, which indicates the overfit of the model. By comparing the model SARIMA (12,1,0) (0,1,0) and SARIMA (0,1,0) (1,1,1), it is hard to tell from the output since their values on the table are similar; they both do not have a spike in their residual acf; their Ljung-Box statistics are significant. Moreover, their values of S, AIC, likelihood are reasonable small, compared to the other three models. Therefore, these two models are the best two we end up with. However, if we count the plots we have previously, model SARIMA (0,1,0) (1,1,1) is better since its residuals versus fitted values look more reasonable than SARIMA (12,1,0) (0,1,0).

1. Conclusion

After the analysis of Air Passengers above, we do believe the SARIMA (0,1,0) (1,1,1) is the best model. Compare to all iden() we made, we think that the log transformation(gamma=0) is the best, and both d=1 and D=1 should be used. By Comparing the table of our four models, considering the significant spike, our best two models are SARIMA (12,1,0) (0,1,0) and SARIMA (0,1,0) (1,1,1), but since model SARIMA (0,1,0) (1,1,1) has more reasonable residuals versus fitted values which is normal distributed, we pick SARIMA (0,1,0) (1,1,1). We pick it also because there is no spike in residual acf plot and it has good fit forecasting for the trend. Consequently, SARIMA (0,1,0) (1,1,1) is our best model to do the time series analysis and forecasting.

1. Works Cited

<https://catalog.data.gov/dataset/air-traffic-passenger-statistics>

<https://www.usatoday.com/story/travel/news/2013/07/06/airline-crash-san-francisco/2495099/>

<https://www.nbcbayarea.com/news/local/An-Overview-of-NBC-Bay-Areas-Investigations-into-Airport-Incidents-468891033.html>

<https://www.nerdwallet.com/blog/studies/sfo-airport-delay/>