Zheqi Wu
PAIK SCHOENBERG, FREDERIC
STATS221: Time-Series Analysis
March 11, 2017

Time-Series Analysis of Traffic Collision in Los Angeles

1 Introduction

In 2010 the governments of the world declared 2011–2020 as the Decade of Action for Road Safety. *Global status report on road safety***[1]** shows that 1.24 million people were killed on the world's roads in 2010. This is unacceptably high. Road traffic injuries take an enormous toll on individuals and communities as well as on national economies. If we can find the pattern of the traffic collision and analyze what are potential reasons related to it, it will be instructive for the government to deal with the traffic problem. Moreover, the traffic collisions are closely related to our life that everyone should pay attention to it.

Intuitively, we can image the number of traffic collision might perform differently from weekday to weekend, as well as between holidays and normal days. Also, the weather might be a latent variable to have a huge influence on the series since it is prone to have traffic accidents in low visibility days such as rainy and snowing days. Consequently, the time-series data might have some weekly cycles or seasonal patterns in it.

The project is aiming to get a thorough understanding of the traffic collision time-series data, including the trend, seasonality, autocorrelations, periodic variations of the underlying phenomenon that produced the series and then forecast the number of traffic collision every day by fitting the series in ARIMA models with estimated parameters.

2 Data Description

The traffic collision information data set is download from DATA.GOV website which reflects every reported traffic collision incident in the City of Los Angeles dating back to 2010. (URL: https://catalog.data.gov/dataset/traffic-collision-data-from-2010-to-present)

The raw data includes several variables such as reported time, area and district location information, victim age, type of traffic collision and so on. Therefore, I clean the data removing incomplete or irrelevant parts, and collect the daily number of traffic collision information from 01/01/2013 to 12/31/2017 to construct my training set with 1827 observations. In addition, I hold the first month daily data in 2018 as the testing set. Finally, I aggregating the dataset at a daily level.

3 Regression and Data Exploration

In order to see time-series patterns, I decide to figure out the two basic classes of components: trend and seasonality.

Figure 1 shows the time plot of daily original data from 2013 to 2017. It seems like that there is an upward trend, so I fit the model via simple linear regression to estimate the trend. The p-value in Table 1 shows that the two parameters in the linear regression are both significant, so it is reasonable to believe the trend is fixed, indicating that the number of traffic collisions has a steady growth over the years.

Then I decompose the original series into trend, seasonal and random components. At the seasonal part, it is interesting to see that there is an obvious downward trend at the end of the year, due to thanksgiving festival and Christmas eve that people tend to stay at home and then the series rises to peaks in March and August. Also, it appears that the series remains steady, i.e., the amplitude of the seasonal is constant. Notice that the variance of the data is not increasing, it is not necessary to transform the data into log form to make it stationary.

To create a (possibly) stationary series, we can detrend the data or difference it. I plot the ACF and PACF of original series in Figure 3. It shows that the ACF decreases very slowly, so it appears that differencing may be a good choice. Indeed, the time plot of the first-differenced data in Figure 4 seems to be stationary with zero means. On the other hand, the time plot for the detrended data appears to be with high variability, which means that this might not be useful in the following analysis even it would have the better explanation than the differenced data. Therefore, I decide to fit the model estimating the first differences $y_t = x_t - x_{t-1}$.

4 Modeling Fitting and Diagnostics

After preprocessing nonstationary time series into a possibly stationary series in data exploration part, I plot the ACF and PACF of the first-differenced series in **Figure 5**. It

shows that the transformed data still has some weekly cycles while both two tail-off in the end, I assume that an ARIMA-model will serve my needs for next step.

I estimate a model-order to be 1 (i.e. D=1) because I use the first differenced series, but it is hard to identify preliminary values of the p, q in ARIMA model on account of the seasonal component effect on the plot. Considering that ACF and PACF might suggest not only one model but many, so I set them to be 1 and try different values later. Table 2 shows the candidate models and the corresponding AIC and BIC, and ARIMA(2,1,1) is chosen with the lowest AIC and BIC. The estimated model is $x_t = 0.0217 + 0.2728x_{t-1} -0.0315\, x_{t-2} - 0.9725\, w_{t-1} + w_t$. However, when I go deep into analysis if it is a preferred model for the data, the result is not good. **Figure 6** shows the standardized residual plot of the estimated model. Although we can see it from the Q-Q plot and time plot that the residual seems to be normally distributed with zero mean and steady variance, the ACF plot shows that there are still obvious weekly cycles in the residuals. In addition, the p-values in the Ljung-Box statistics plot are above the standard line after lag 1, which means it is significant and the residuals of the model are dependent.

At the same time, the forecasting part gives us the same result that the ARIMA(2,1,1) model can not have a good 30-step-ahead forecast for the traffic collision series. **Figure 7** shows this forecasting plot with red predictions and blue 95% confidence interval. The predictions are almost like a flat line, which fail to capture the patterns in the series. I further calculate the MSE of the 30 predictions with real data, the value is 185.98.

As I mentioned above, the traffic collision daily series is very likely to have a weekly cycle that the classic ARIMA model can not have a good fit on it. Looking at the sample ACF and PACF in **Figure 5**. From the seasonal component perspective, we can see it appears that the ACF is cutting off a lag 1s (s = 7), whereas the PACF is tailing off at lags 1s, 2s, … . So the preliminary value in the season is P = 0, Q = 1, s = 7. And from the non-Seasonal component perspective, it seems that ACF and PACF both are tailing off. Same as in the previous classic ARIMA analysis, we start with p = q = 1. So I try the seasonal ARIMA model around in order to find a better model with more accuracy. Table 3 shows the seasonal ARIMA model with different parameters and the corresponding AIC and BIC. Among all those models, in view of the seasonal ARIMA(2,0,1)(1,1,1)[7] has the lowest AIC 6.484 and BIC 5.497, I choose this model as our best seasonal ARIMA model and do the same diagnostics and forecasting.

Figure 8 shows the residuals diagnostics plot. Firstly, the residuals appear to be randomly scattered around zero with a roughly constant variance throughout most of the series. However, there are outliers which are distant from other residuals at every beginning of year. I guess it may be caused by yearly long cycles due to holidays that the seasonal arima model cannot eliminate this influence. Secondly, the Q-Q plot suggests that sets of residuals very likely come from Normal distributions. Thirdly, most of the ACF values of the residuals are not significant and the weekly cycle which is

obvious in the classic ARIMA model is hard to recognize. Finally, Ljung-Box statistics plot shows that the residuals are independent of each other since all the P-values are not significant. Therefore, I can conclude that the seasonal ARIMA(2,0,1)(1,1,1)[7] might be better than ARIMA model above.

Figure 9 shows the forecasting plot. Rather than a flat line, the seasonal ARIMA model can actually capture some of the variations of the time series. And the MSE 146.72 is smaller than classic ARIMA model. However, there is a sudden drop at the beginning of 2018 which surprised me that even the seasonal ARIMA model cannot capture its variation, so the MSE is not as low as I except. Looking at the original time plot of the data, we can see clearly that there seems to be a yearly pattern, i.e., there is a sudden drop at the beginning of each year. Unfortunately, the seasonal ARIMA model cannot capture it.

5 Spectral Analysis

This part focuses on the frequency domain approach to analysis our traffic collision series to investigate the periodic components. In order to eliminate the trend influence on the following analysis, I use the detrended data via simple linear regression.

At first, I plot the raw periodogram displayed in **Figure 10**, where the frequency axis is labeled in multiples of 1/365. It is quite apparent we can see the three predominant frequencies. But in case those peaks are caused by the noise, then we use the modified Daniell kernel to smooth the periodogram displayed in **Figure 11**. Indeed, we have same result that three main **w** at 52/365, 104/365, 2/365, which indicates that the traffic collision series has three main periods 7 days, 3.5 days and 182.5 days. In this way, the plot suggests that the weekly cycle of traffic collision are predominant in the series while it has the minor half-yearly cycle. Therefore, the spectral analysis results are in accord with the analysis in ACF and PACF in the ARIMA model and our intuition understanding.

The second approach is to estimate the spectral density via AR model with lowest AIC. **Figure 12** shows the plot of AIC and BIC for various AR models, and AIC reach the minimum at lag 41, so I choose to use the spectral density of AR(41) as the approximation displayed in **Figure 13** with familiar three main peaks. It is interesting to see that even if we use the different method to do the spectral analysis, we still have the same result.

6 Conclusions and Next steps

In this project, we explore the traffic collision time-series data with trend and seasonality giving a real-world interpretation. Then we choose the seasonal ARIMA(2,0,1)(1,1,1)[7] model to do forecasting, which has a better fit of data and smaller MSE compare to ARIMA(2,1,1). In the spectral analysis, we find out that the weekly cycle of traffic collision are predominant in the series while it has the minor half-yearly cycle.

One main shortage of the seasonal ARIMA model is that it cannot fit with long seasonality. That is to say, the multiplicative seasonal ARIMA model assumption seems to work well with many seasonal time series data sets while reducing the number of parameters that must be estimated. The problem is that there are m - 1 parameters to be estimated for the initial seasonal states where m is the seasonal period. So for large m, the estimation becomes almost impossible. Maybe we can use a Fourier series approach where the seasonal pattern is modeled using Fourier terms with short-term time series dynamics allowed in the error.**[2]** In addition, we can introduce other variables such as weather and location of traffic collision to cooperate with the original dataset and get better predictions.

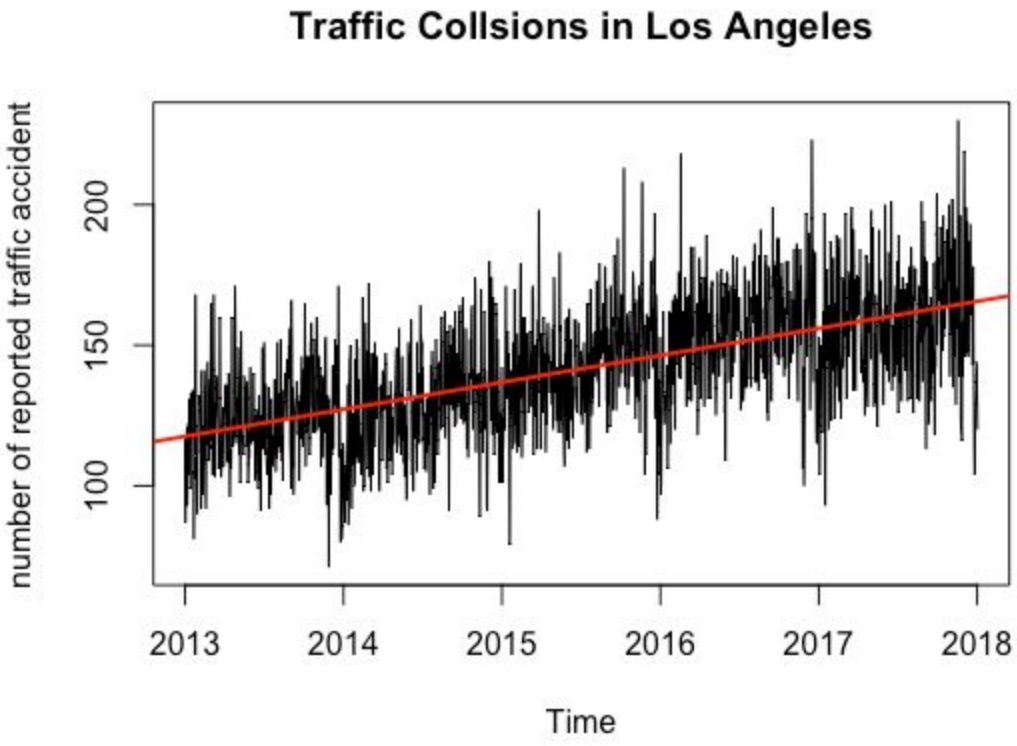

Figure 1

Coefficients	Estimate	Std.Error	t.value	pr(>|t|)
(Intercept)	-1.923e+04	5.940e+02	-32.37	<2e-16 ***
time(dat)	9.611e+00	2.947e-01	32.61	<2e-16 ***

Table 1

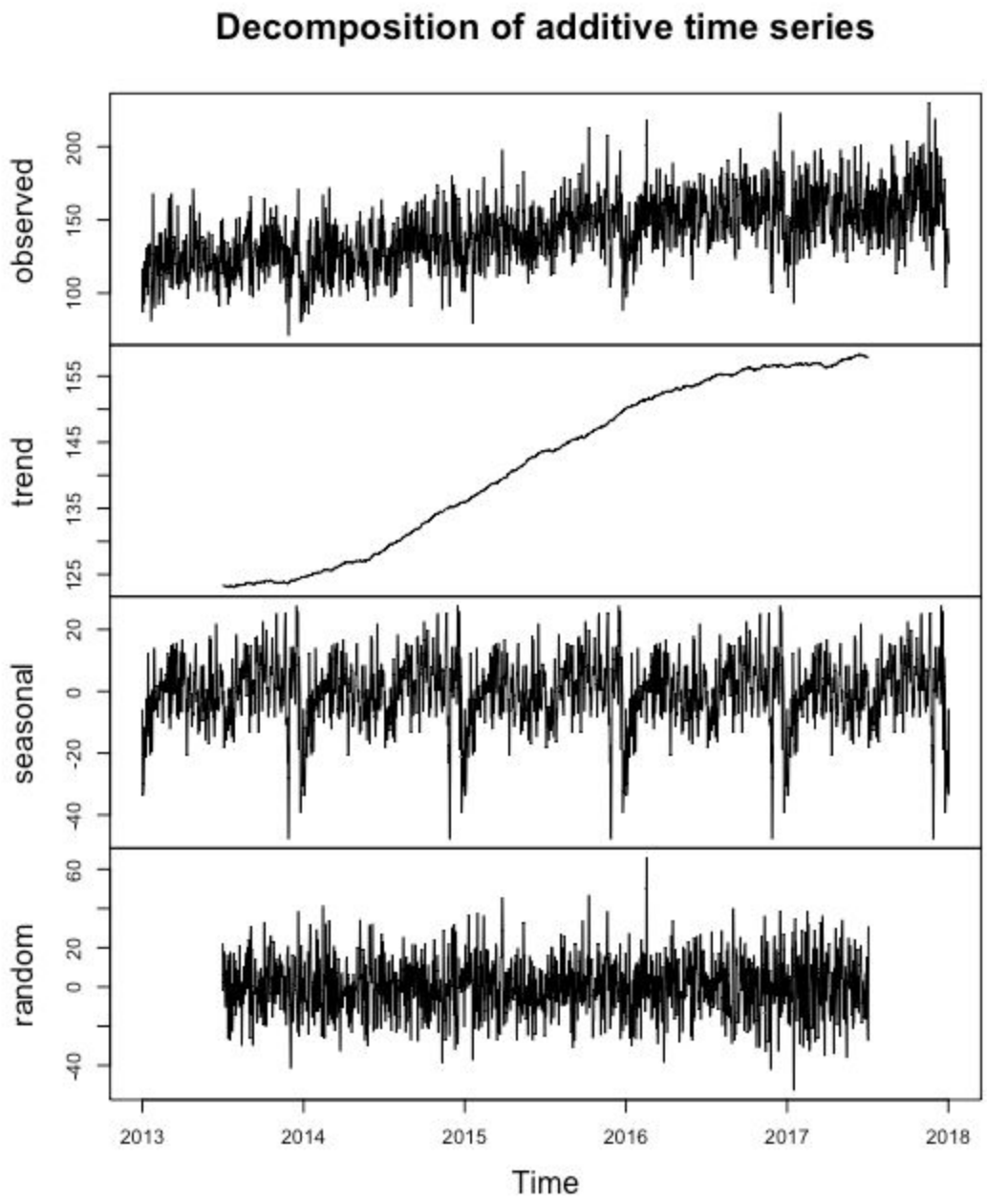
Decomposition of additive time series
observed
100
150
200
trend
125
135
145
155
seasonal
-40
-20
0
20
random
-40
0
20
60
2013
2014
2015
2016
2017
2018
Time

Figure 2

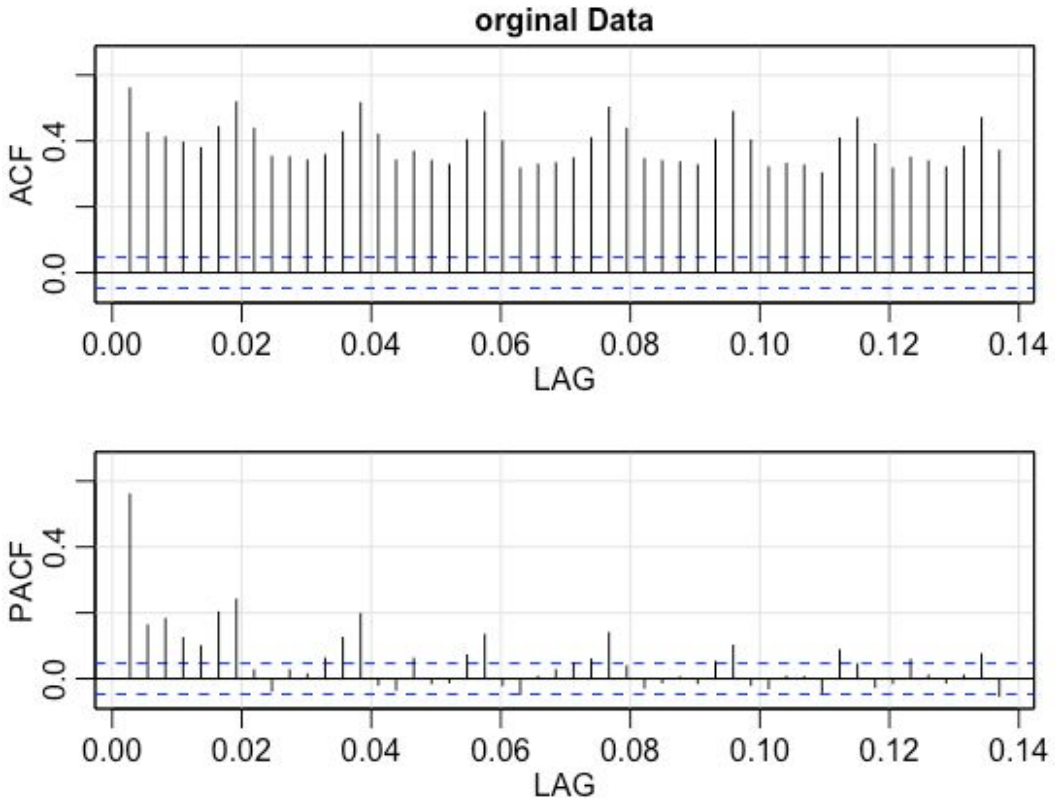

Figure 3

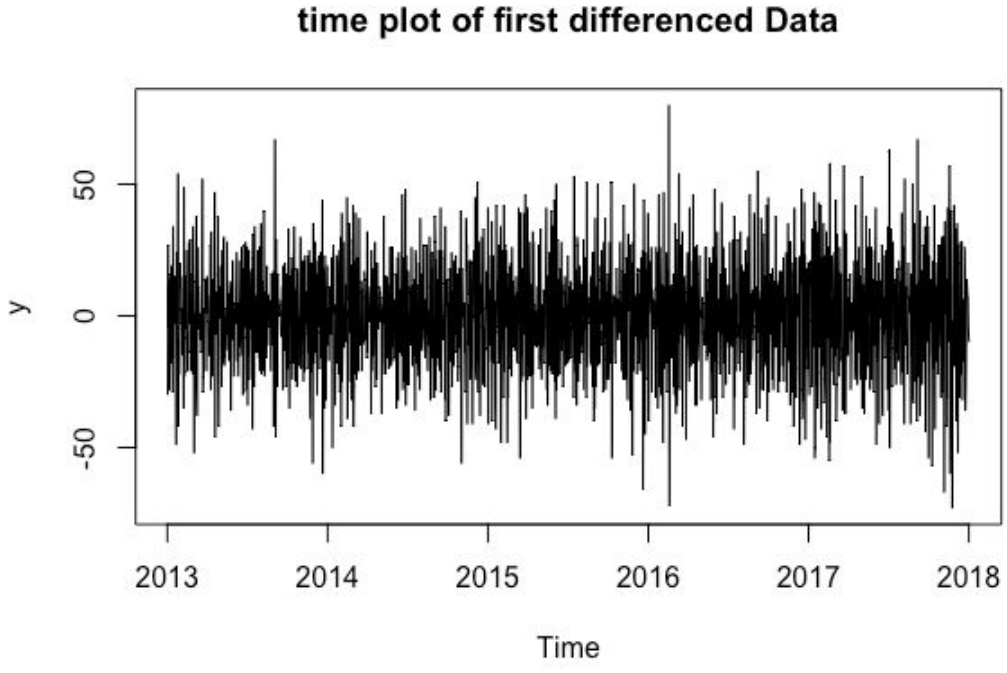

Figure 4

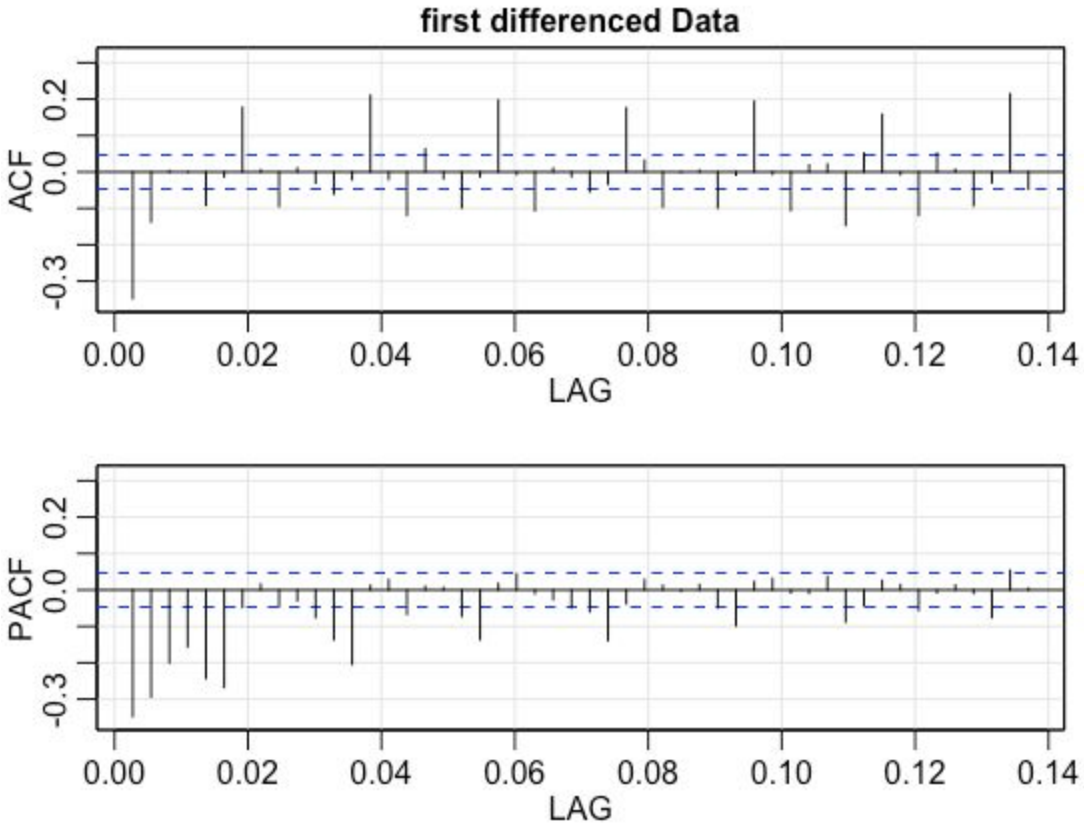

Figure 5

ARIMA model´	AIC	BIC
(2,1,1)	**6.706121**	**5.718186**
(2,1,2)	6.706919	5.721999
(2,0,1)	6.707929	5.719994
(2,0,2)	6.707674	5.722755
(1,0,2)	6.763829	5.772877

Table 2

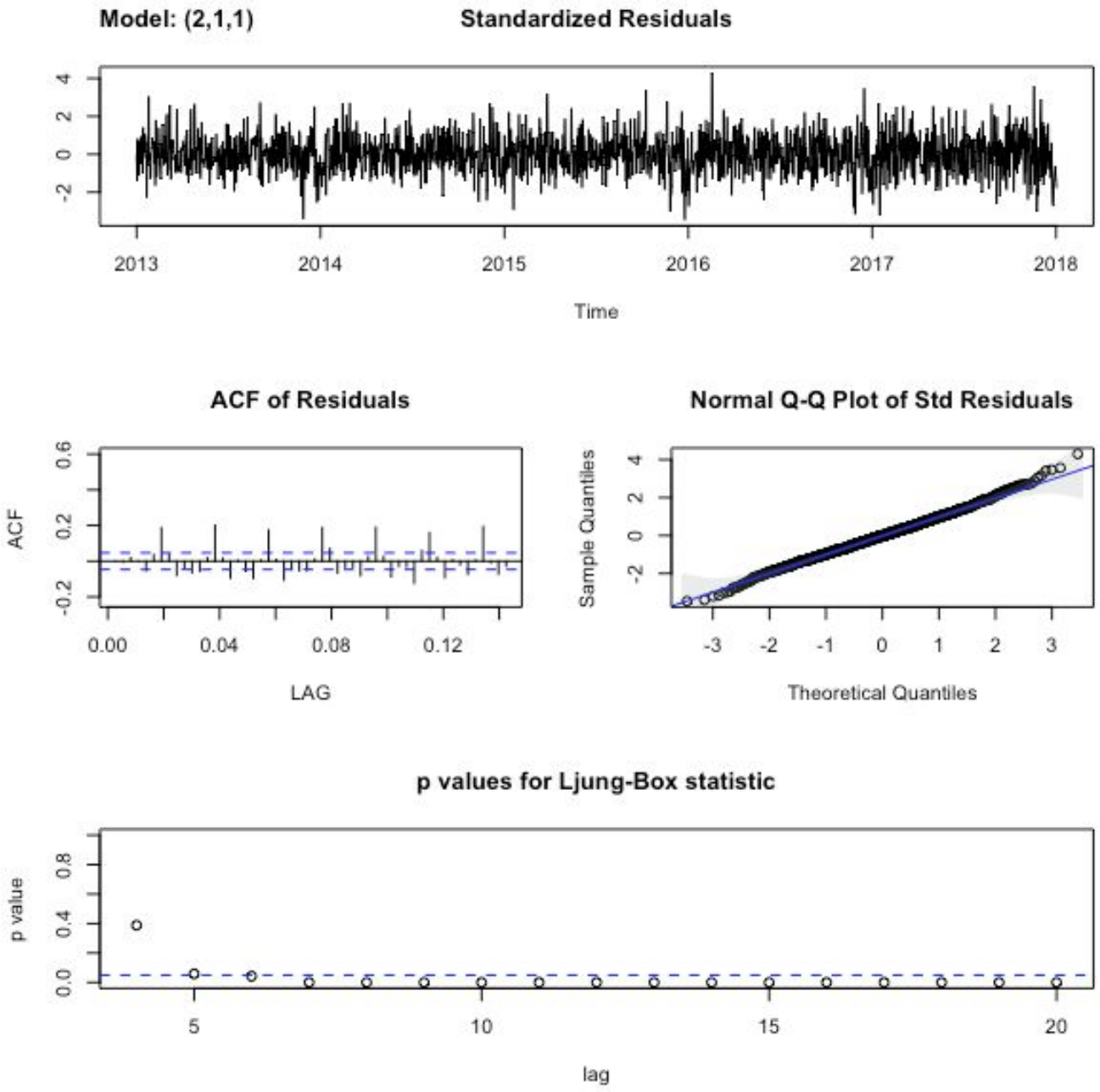
Model: (2,1,1)
Standardized Residuals
-2
0
2
4
2013
2014
2015
2016
2017
2018
Time
ACF of Residuals
ACF
-0.2
0.2
0.6
0.00
0.04
0.08
0.12
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
-2
0
2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
0.4
0.8
5
10
15
20
lag

Figure 6

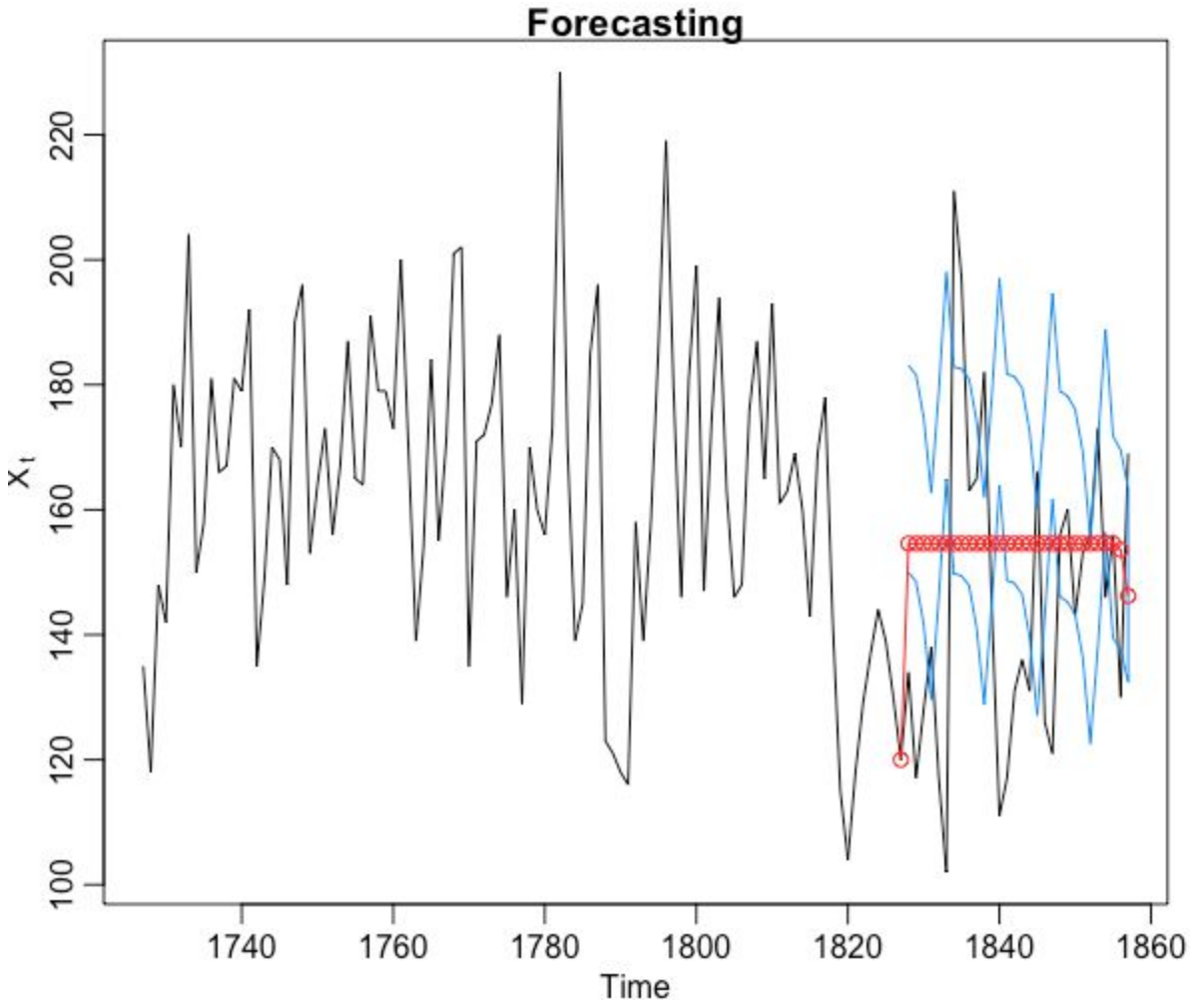

Figure 7

ARIMA model	AIC	BIC
(2,0,1)(1,1,1)[7]	6.484181	5.49734
(2,0,1)(0,1,1)[7]	6.484315	5.499396
(1,0,1)(0,1,1)[7]	6.485275	5.502278
(1,0,1)(1,1,1)[7]	6.486009	5.50109
(2,1,1)(0,1,1)[7]	6.492648	5.504713
(2,1,1)(1,1,1)[7]	6.493552	5.508633

Table 3

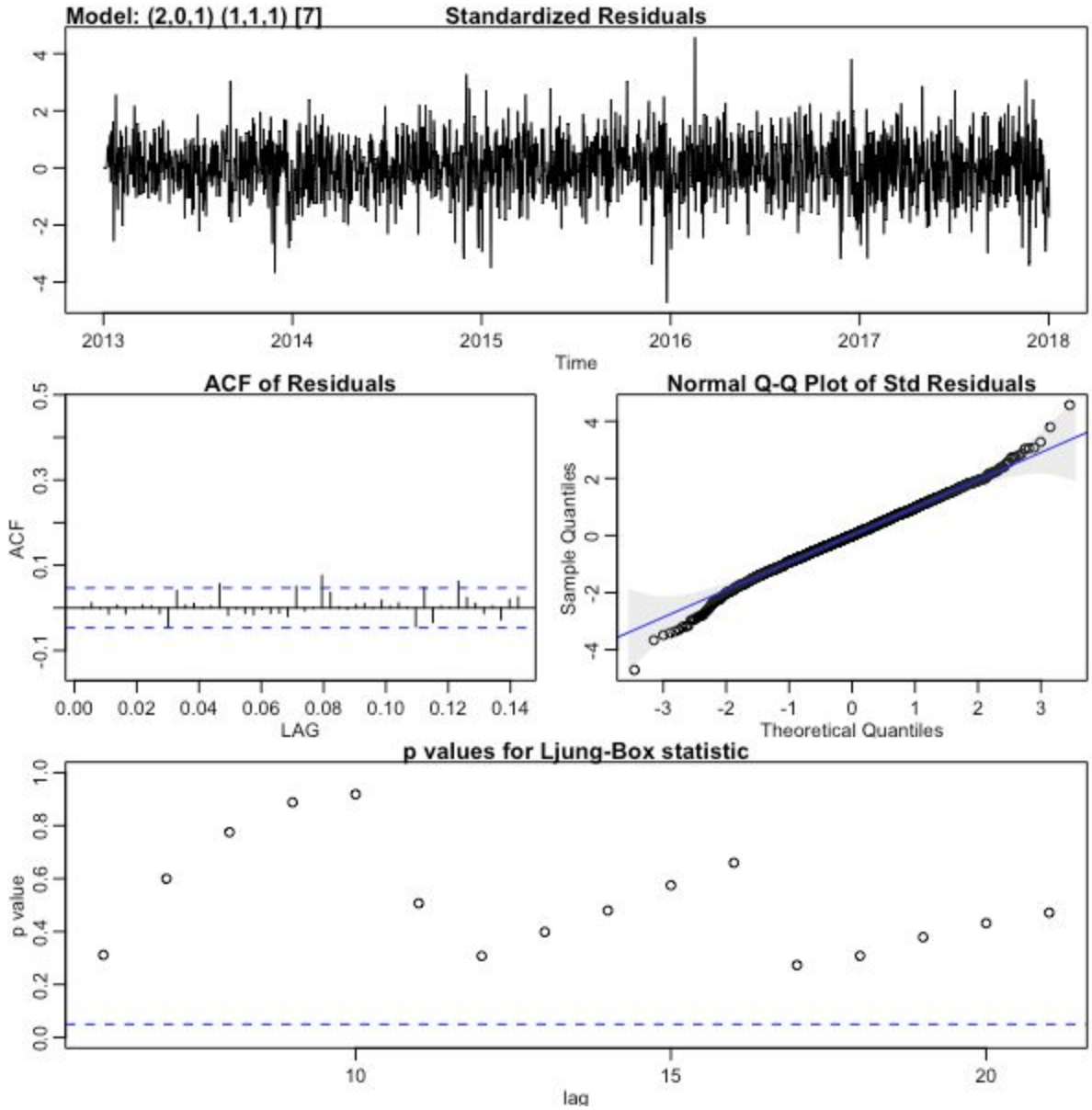
Model: (2,0,1) (1,1,1) [7]
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
lag

Figure 8

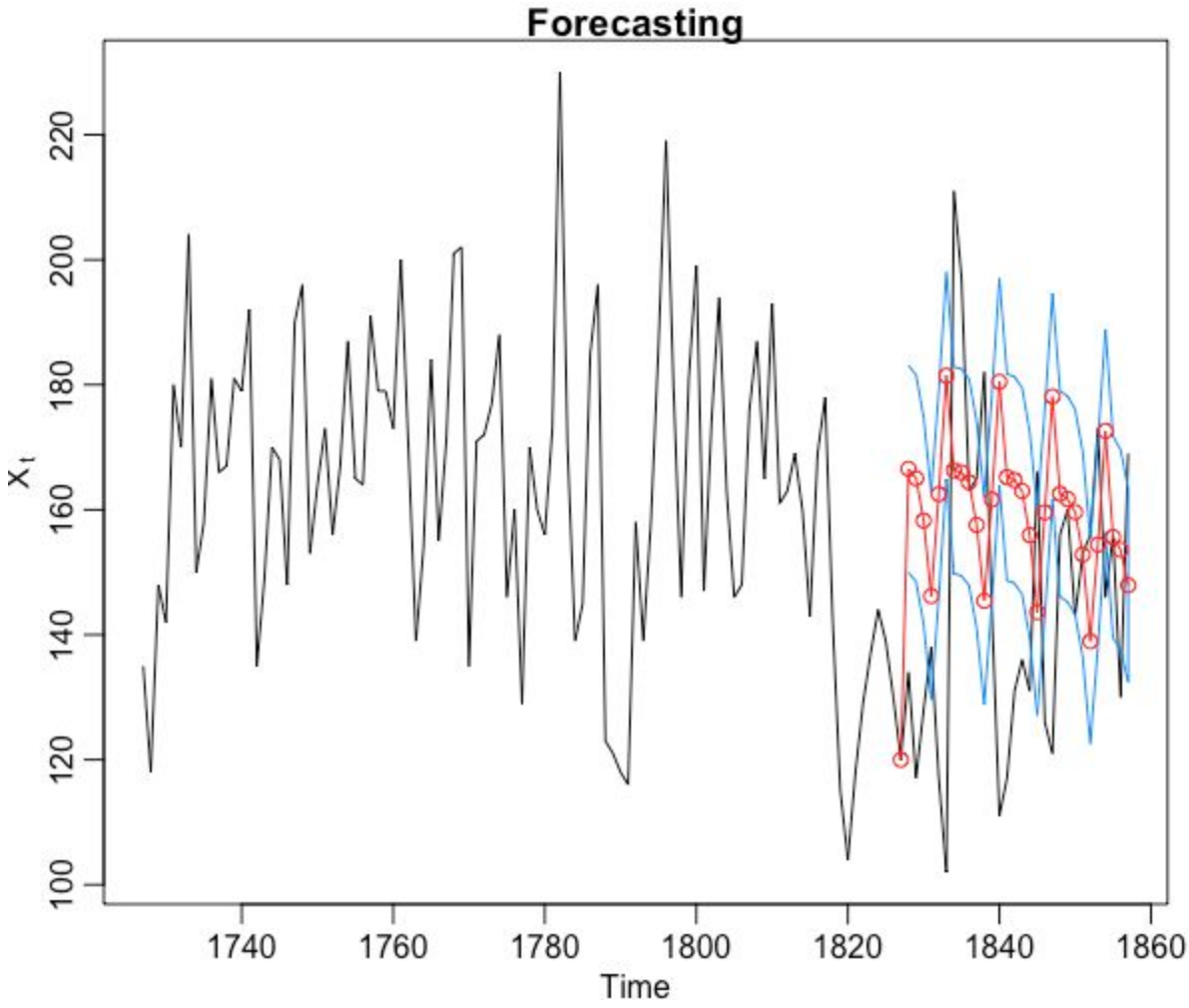
Forecasting
X_t
100
120
140
160
180
200
220
1740
1760
1780
1800
1820
1840
1860
Time

Figure 9

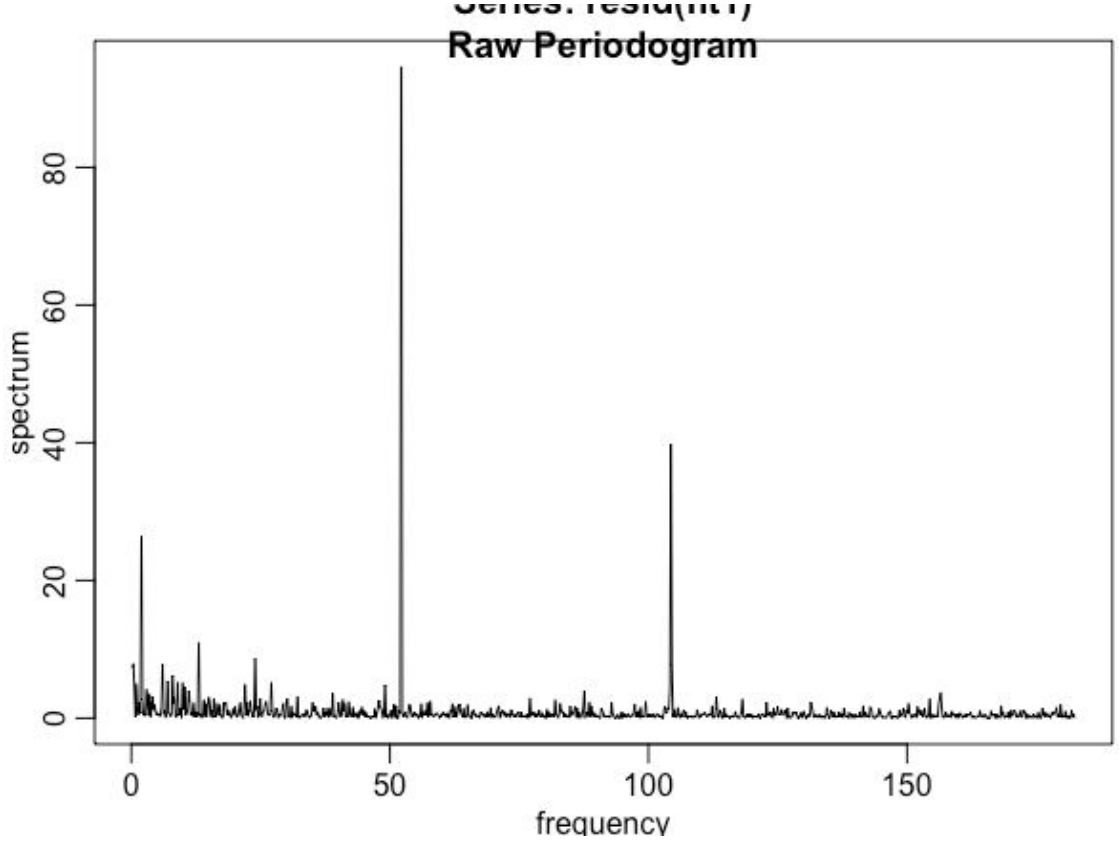

Figure 10

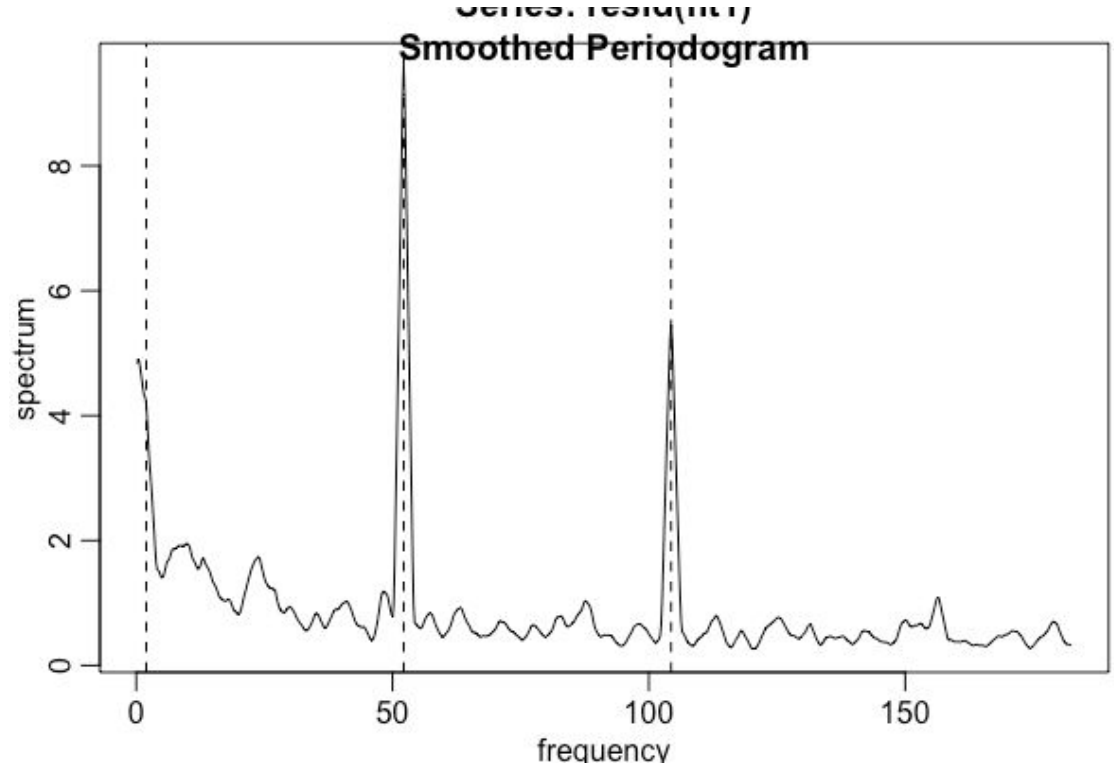

Figure 11

time series	w	period
Traffic collision	52/365	7 days
	104/354	3.5 days
	2/365	182.5 days

Table 4

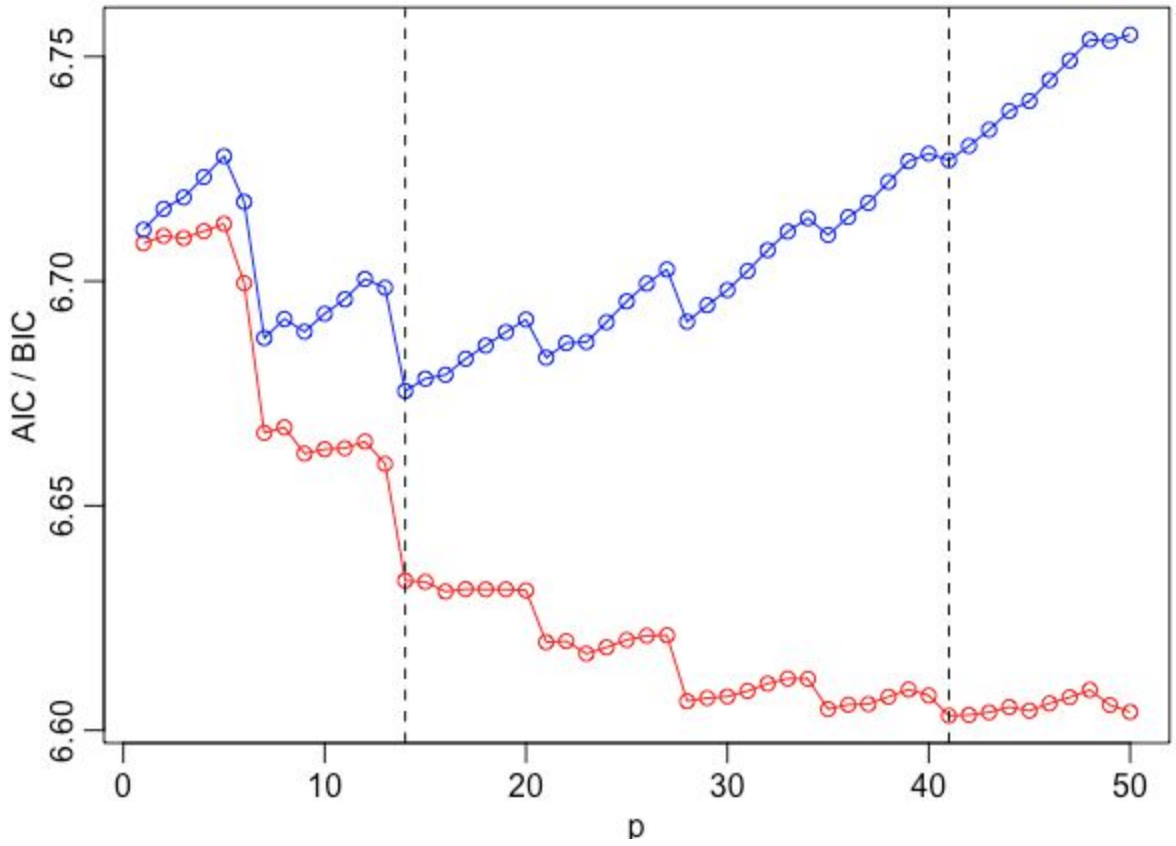

Figure 12

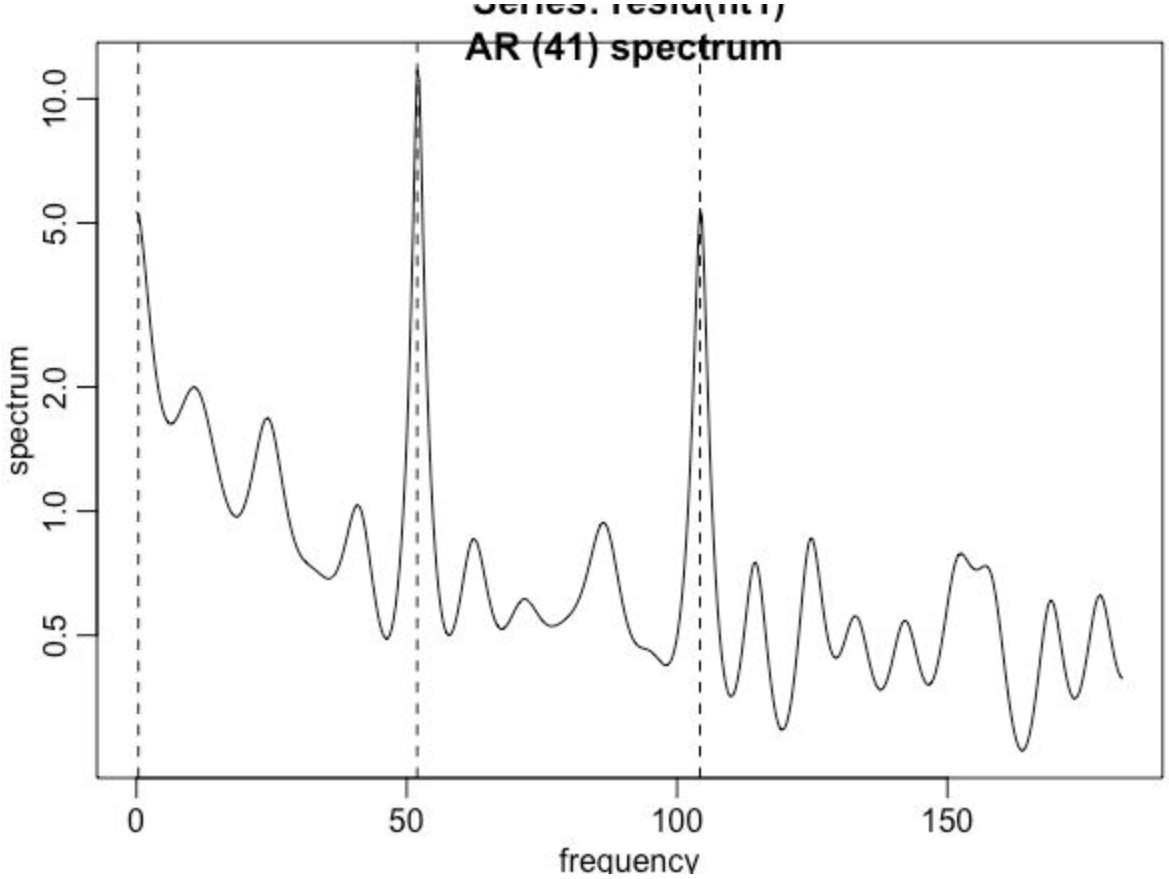

Figure 13

References

[1] World Health Organization, Global status report on road safety 2013
[2] Hyndsight, Forecasting with long seasonal periods 2010, https://robjhyndman.com/hyndsight/longseasonality/