CSC 425

Virginia Jiang

Guy Dor

**Speeding Violations in Chicago**

**Non-Technical Summary**

The dataset we chose to analyze was the daily volume of scanner-recorded speeding violations in Chicago from Kaggle (<https://www.kaggle.com/chicagopolice/speed-violations>). Even though the website states the data is from July 1, 2014 through December 31, 2016, it now is actually through August 31, 2017. The data reflects recordings of all violations, including those where a citation was not issued. So, we can assume that human intervention has not skewed the results.

Through our analysis, we were able to find multiple statistical models that could be used to forecast speeding violations at the daily, weekly, and monthly level. While each of these models were tested to have varying levels of accuracy, we do not rule one model out over another because the results tell a different story for each level of periodicity.

At the daily level, we found it difficult to create a strong model but were able to find intricacies that allow us to forecast out in 7-day periods.

At the weekly level, we were able to find a strong model, but were reluctant to compare it to the daily model as it is virtually a model on a different dataset and forecasts a different level of detail.

At the monthly level, in trend to what happened at the weekly level, we found a model with an even stronger accuracy. Once again, this was based on smoother data, so comparing the model to the weekly or daily models is not advised.

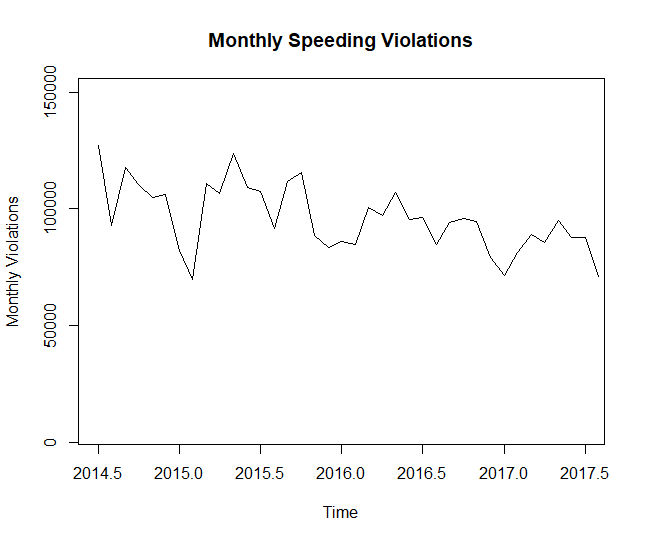
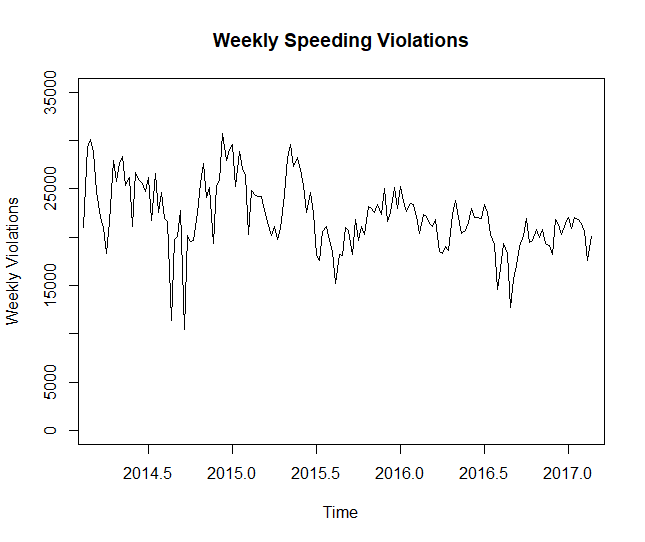
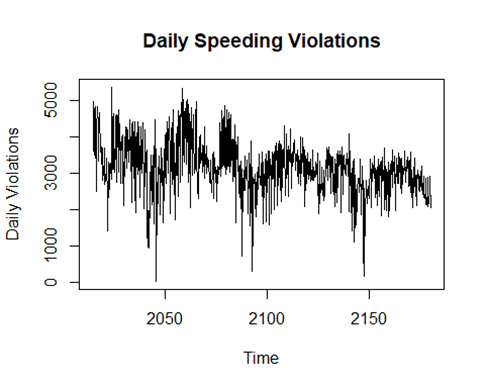
Overall, we discovered that speeding can be considered a highly seasonal event. People tend to speed less on the weekends and during the months of August and January, which are months of the highest and lowest temperature, respectively. In contrast, People tend to speed more on weekdays and during spring and early summer months (i.e. April through June).

As will be demonstrated in the technical summary, the models that were chosen at each level were discovered independently of each other and were not used to reinforce data at different levels. This was done to remove bias of performance from one level of data over another, as well as to maintain that the different levels of data should be treated as different datasets.

**Technical Summary**

**Exploratory analysis of the Data**

We decided to explore the data at three different levels of periodicity: daily, weekly, and monthly. We aggregated the data at daily, weekly, and monthly levels. We have 1,156 observations at daily level, 165 observations at weekly level, and 38 observations at monthly level. The time plots tended to smooth as the levels increased from daily to monthly, and at a first glance series appeared to be non-stationary.



As mentioned above the time series appear to be non-stationary. We then checked the ACFs and PACFs, and they all showed some level of decay, slower at the daily level and faster at the weekly and monthly levels. Dickey-Fuller tests also confirmed the three types of time series were non-stationary.

We used differencing to stationarize the three series. For our daily and monthly data we used first differencing, and our for weekly we used seasonal difference with lag of 52.

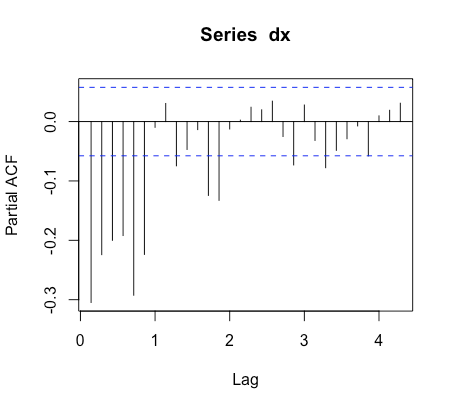
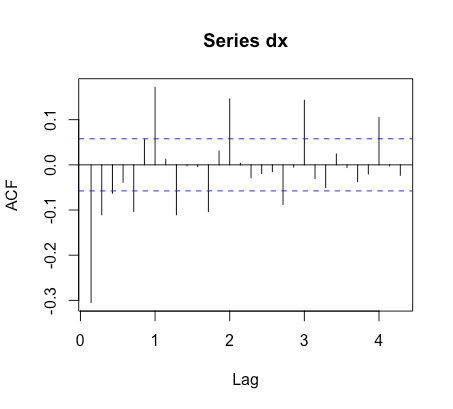
Then we performed the Dickey Fuller Test on the differenced data and we reject the null hypothesis and confirm that the series were now stationary and we can use the differenced data to find a time series model.

We used the Ljung Box Tests to check for autocorrelation, and we found autocorrelation at each of the three levels.

We then decided to approach modeling for each level independently.

**Daily Level:**

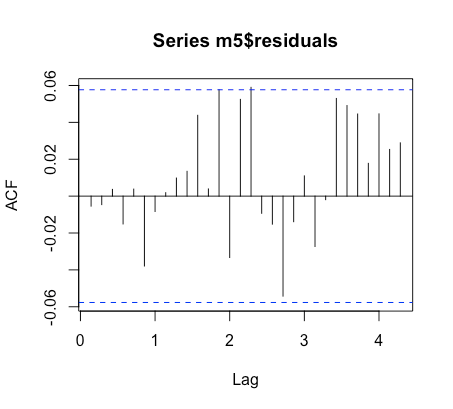
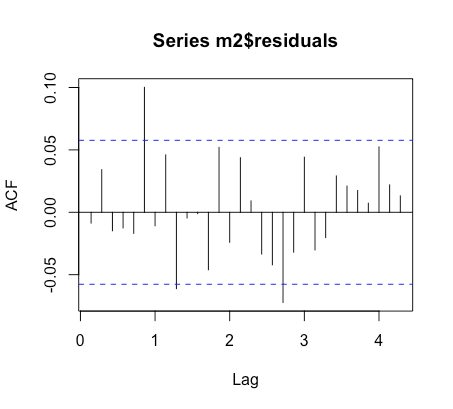
**Model fitting**After differencing the data, we see there is a moving average component from the ACF plot and an autoregressive component from the PACF plot. We used auto.arima with BIC criteria to find the best model and we get the model ARIMA(1,0,1)(2,0,0)[7] with zero mean. However, from the ACF plot we see that there may be a MA(7) component. We built another model ARIMA(1,0,7)(2,0,0)[7] with non-zero mean.

****

**Residual analysis and model diagnostics**

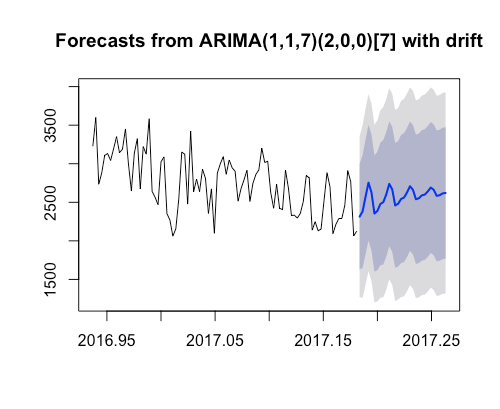
The model ARIMA(1,0,1)(2,0,0)[7] has AIC of 17823.04 and BIC of 17853.36. All four coefficients for this model are significant. However, the ACF plot of the residual showed there is autocorrelation; the Ljung Box Test for the residuals also confirmed there is autocorrelation in the residuals.

The model ARIMA(1,0,7)(2,0,0)[7] with non-zero mean has AIC of 17795.95 and BIC of 17856.57. Two of the coefficients, AR1 and MA5, are not significant for this model. The residual ACF plot showed the autocorrelation is not significant and it passed the Ljung Box Test at lag 22.



**Forecast analysis and Backtesting**

The model ARIMA(1,0,1)(2,0,0)[7] has MAPE of 16.88694%. The model ARIMA(1,0,7)(2,0,0)[7] with non-zero mean has a MAPE of 16.1271%. ARIMA(1,0,7)(2,0,0)[7] with non-zero mean is a more complex model, but it has a lower AIC and lower MAPE, and more importantly we did not find strong autocorrelation in the residuals. We decided to use ARIMA(1,0,7)(2,0,0)[7] with non-zero mean to forecast. The forecast plot showed the past 90 days and forecast of 30 days ahead. It showed a similar pattern from our past data and it slowly converges to the mean.



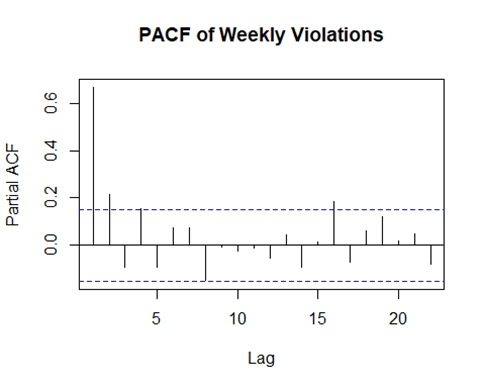
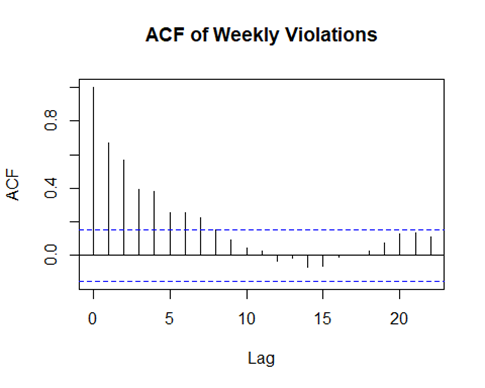
**Analysis of the results and discussion**

At the daily level we can see details of how the violations vary in each day of the week. When we zoom in on our forecast plot we can see that the weekends have fewer violations than on weekdays. While this is a good model that is able to forecast daily trends, the drawback is that we cannot forecast out too far ahead, as they will eventually converge to the mean. To understand better the long-term trends, we can look at the data at the weekly and monthly level.

**Weekly Level:**

**Model fitting**

Looking at the ACF and PACF plots of the weekly data suggest that there is possibly both autoregressive and moving average components to model, as well as possible indications of a seasonal component.

****

We evaluated multiple models based on the output of differencing at multiple levels using auto.arima for the original weekly data and the 52 week seasonal differenced data.

Original Weekly Data: We got two auto.arima models with and without using parameters stepwise=FALSE and approx=FALSE. stepwise=FALSE allows the process to evaluate more potential models, while approx=FALSE forces it to evaluate the likelihood more accurately for each model.

M0=ARIMA(3,1,2)(1,0,0)[52]

M1=ARIMA(2,1,2)(1,0,0)[52]

M1 had lower AIC and BIC

Seasonal differenced data (52 periods): We got two auto.arima models with and without using parameters stepwise=FALSE and approx=FALSE.

M00=ARIMA(2,0,0) with non-zero mean

M11=ARIMA(1,0,2) with non-zero mean

M11 had lower AIC and BIC and could be translated back to the original data as M12 below:

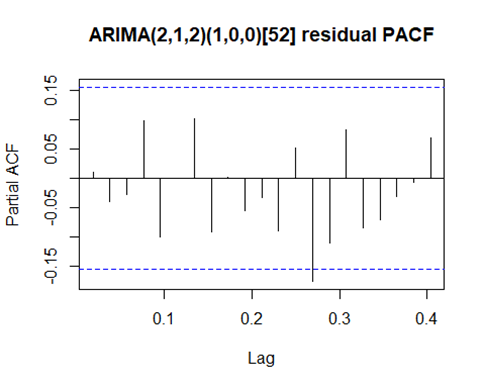
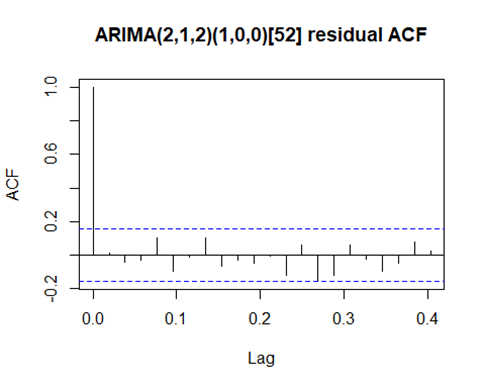
M12=ARIMA(1,0,2)(0,1,0)[52] with drift

Since M1 and M12 had the lowest AIC and BIC out of the models evaluated, we decided to move on to Residual analysis

**Residual analysis and model diagnostics**

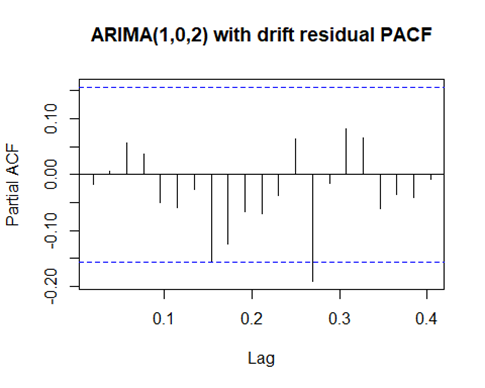
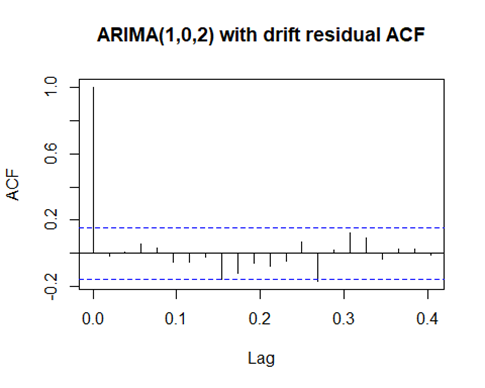
From studying the residuals of M1, we see possible autocorrelation at lag 14, but the Ljung Box tests quickly indicate that this is not the case. The same can be said for M12.

M1 residual ACF and PACF



Ljung-Box tests failed to reject null hypothesis at 26 and 52 lags.

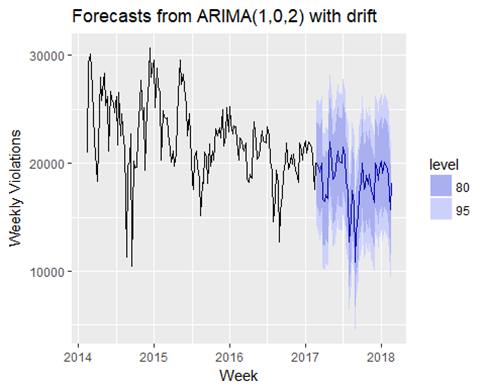
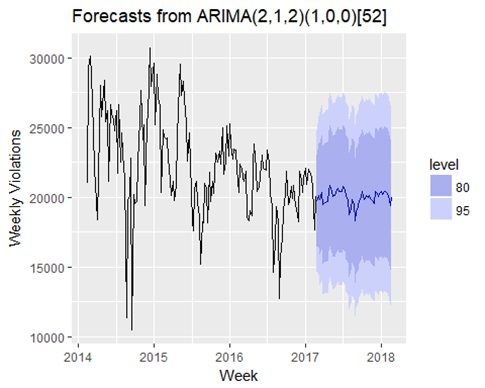
M12 residual ACF and PACF



Ljung-Box tests failed to reject null hypothesis at 26 and 52 lags.

**Forecast analysis and Backtesting**

While M12 visually looks like a better model to forecast in regard to the ability to account for the intricacies of the data, we decided to choose M1 after backtesting due to lower error measurements.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | RMSE | MAE | MAPE | SMAPE |
| M1 | 2064.699 | 1611.693 | 9.05% | 8.85% |
| M12 | 2484.861 | 2007.186 | 10.60% | 10.63% |

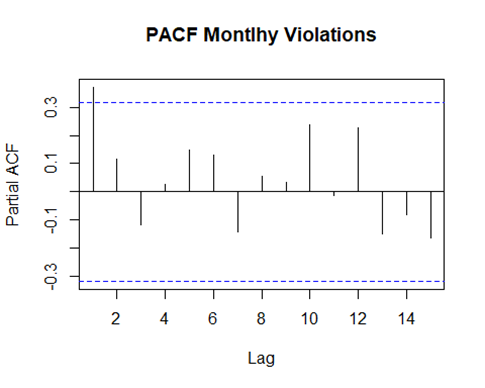
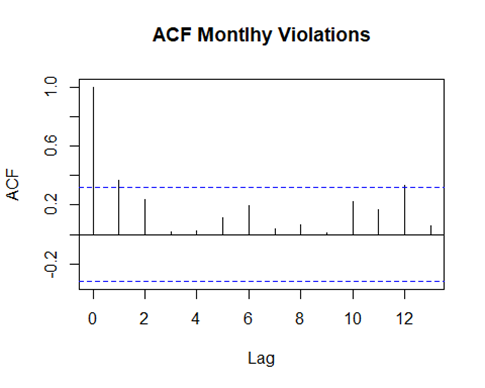
**Analysis of the results and discussion**

We can see that the M1 model is able to functionally forecast and represent the data in a way that we could possibly make predictions at the weekly level. We also see that the AIC/BIC criteria and the accuracy metrics are better than at the daily level. This is to be expected since the data was smoothed by lowering the amount of observations, which tells us that we should not compare this model to the model chosen at the daily level.

**Monthly Level**

**Model fitting**

Looking at the ACF and PACF plots of the monthly data allows us to determine a possible MA(1) component.



We evaluated two models using auto.arima, one with evaluation of the data itself and another with a coerced seasonal differencing component.

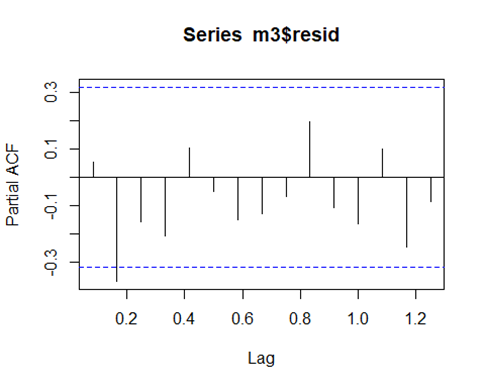
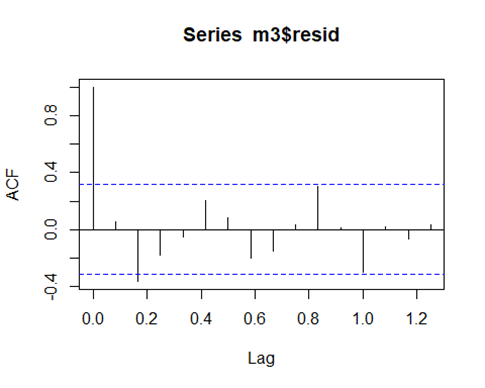
M3=ARIMA(0,1,1)(1,0,0)[12]

M4=ARIMA(0,0,0)(0,1,0)[12] with drift

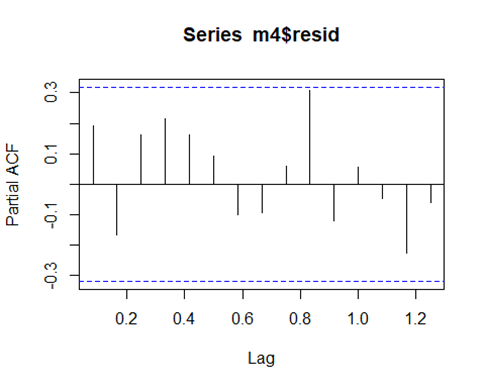
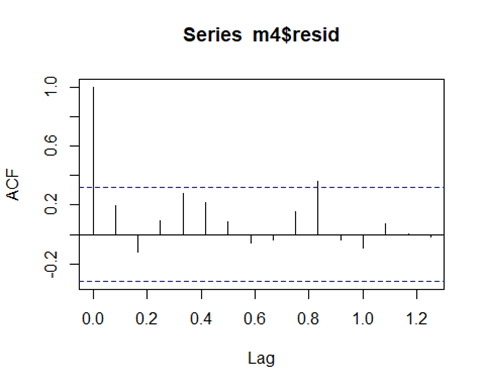
Even though the model evaluated for M4 is basically white noise, we decided to compare the results to M3 since it had a lower AIC and BIC than M3.

**Residual analysis and model diagnostics**

From studying the residuals of M3, we see possible autocorrelation at lag 2, but the Ljung Box tests quickly indicate that this is not the case. The same can be said for M4 at lag 10.



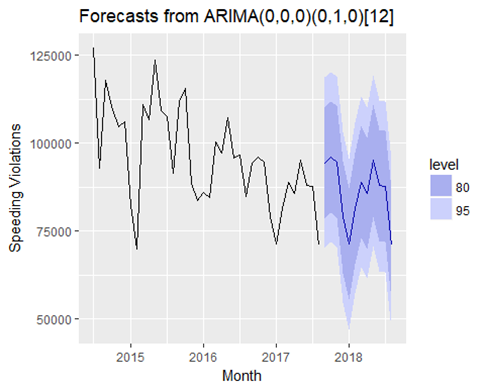
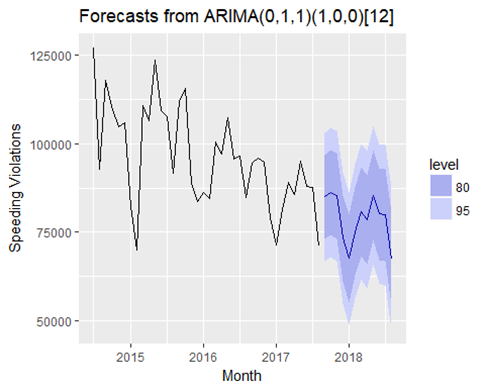
Ljung-Box tests failed to reject null hypothesis at 6 and 12 lags.



Ljung-Box tests failed to reject null hypothesis at 6 and 12 lags.

**Forecast analysis and Backtesting**

What is interesting to observe is that the forecasts between the two models behave almost identically.



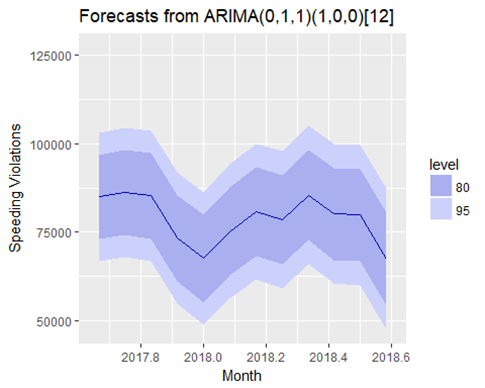
However, when evaluating the backtesting, we see that M3 is still a superior model to the white noise M4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | RMSE | MAE | MAPE | SMAPE |
| M3 | 6076.338 | 3900.595 | 4.98% | 4.74% |
| M4 | 10531.4 | 9940.429 | 11.79% | 11.04% |

**Analysis of the results and discussion**

The monthly model M3 had the best level of accuracy when looking at MAPE (comparing based on MSE or MAE is not possible when looking at different periodicities).

After selecting M3 as our winner, we decided to look at the forecast a little more thoroughly.



Not only do we get a simpler forecast from smoother data, but we can also observe the seasonal patterns that follow the data. This allows us to make some inferences about speeding behavior during a given season. For example, people either drive less or are more careful during extreme weather (i.e. when it is too hot or when there is ice during August and January, respectively). Essentially, people are more likely to speed when given the right conditions to do so. While that statement may seem somewhat redundant, the model simply reinforces the point.

**Conclusion**

After comparing the three models, all three have their advantages and disadvantages. As the level increased from daily to monthly, the models became less complex and had better accuracy. At the same time, the number of observations decreases as the level increases, which means that while the higher-level models may have better accuracy, they may also be more susceptible to “black swan” events. We would need to have as many observations on the monthly and weekly levels as on the daily levels to be as confident in the higher-level models as we are in in the lower level models. That being said, we are still confident in the three models with the information available to us.

Project Appendix

# Loading Libraries and setting local working Directory

library(tseries)

library(fBasics)

library(forecast)

library(lmtest)

library(ggplot2)

library(astsa)

library(fUnitRoots)

library(lubridate)

library(scales)

library(gridExtra)  
setwd("C:\\Documents\\CSC 425\\Project")

# DAILY DATA

# read data

myd=read.table("Speed\_Camera\_Violations.csv",header=T, sep=',')

# aggregrate data by day

smyd <- aggregate(x = myd[c("VIOLATIONS")],FUN = sum,  
 by = list(VIOLATION.DATE = myd$VIOLATION.DATE))

# sort data by day

dsmyd = smyd[order(as.Date(smyd$VIOLATION.DATE, format="%m/%d/%Y")),]

head(dsmyd)

## VIOLATION.DATE VIOLATIONS  
## 545 07/01/2014 4956  
## 549 07/02/2014 3977  
## 553 07/03/2014 4852  
## 557 07/04/2014 3695  
## 561 07/05/2014 3489  
## 565 07/06/2014 3599

basicStats(dsmyd$VIOLATIONS)

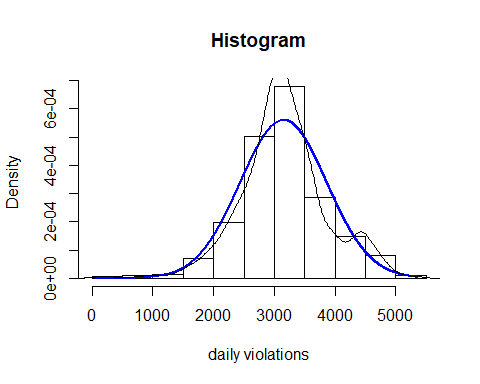
## X..dsmyd.VIOLATIONS  
## nobs 1.156000e+03  
## NAs 0.000000e+00  
## Minimum 3.200000e+01  
## Maximum 5.358000e+03  
## 1. Quartile 2.759500e+03  
## 3. Quartile 3.522750e+03  
## Mean 3.154998e+03  
## Median 3.123500e+03  
## Sum 3.647178e+06  
## SE Mean 2.093201e+01  
## LCL Mean 3.113929e+03  
## UCL Mean 3.196067e+03  
## Variance 5.065006e+05  
## Stdev 7.116885e+02  
## Skewness -4.546200e-02  
## Kurtosis 1.131251e+00

vts= ts(dsmyd[,2], start=c(2014,7,1), freq=7)

# histogram

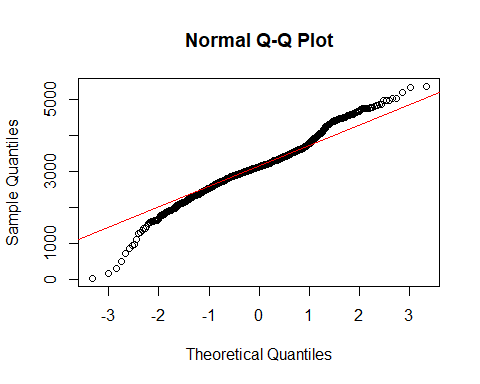
hist(dsmyd$VIOLATIONS, xlab="daily violations", prob=TRUE, main="Histogram")  
xfit<-seq(min(dsmyd$VIOLATIONS),max(dsmyd$VIOLATIONS),length=5500)  
yfit<-dnorm(xfit,mean=mean(dsmyd$VIOLATIONS),sd=sd(dsmyd$VIOLATIONS))  
lines(xfit, yfit, col="blue", lwd=2)  
lines(density(dsmyd$VIOLATIONS, col= 'blue', lwd =2))

## Warning: In density.default(dsmyd$VIOLATIONS, col = "blue", lwd = 2) :  
## extra arguments 'col', 'lwd' will be disregarded



# qq plot

qqnorm(dsmyd$VIOLATIONS)  
qqline(dsmyd$VIOLATIONS, col = 2)

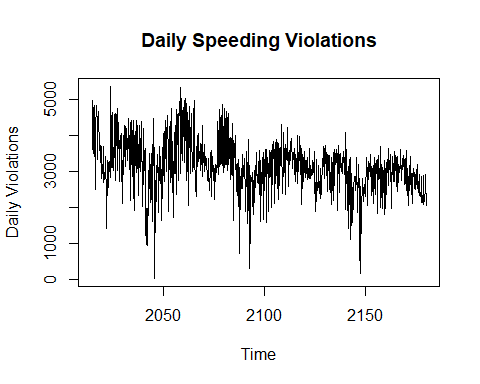


# normality test

normalTest(dsmyd$VIOLATIONS,method=c("jb"))

##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 62.8218  
## P VALUE:  
## Asymptotic p Value: 2.287e-14   
##   
## Description:  
## Sat Mar 10 13:40:02 2018 by user: guy.dor

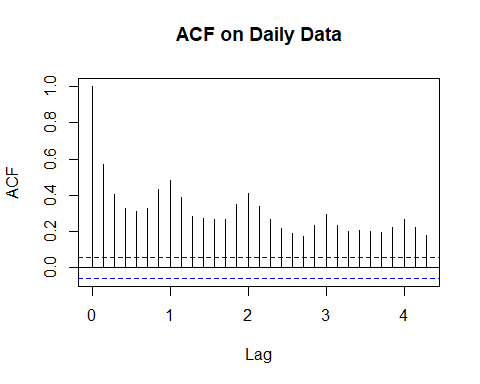
plot(vts, ylab='Daily Violations', main = 'Daily Speeding Violations')



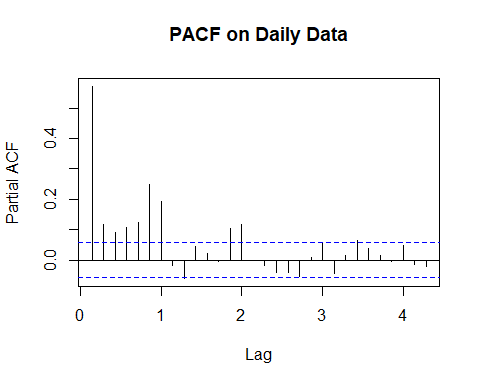
acf(vts, plot=F)

##   
## Autocorrelations of series 'vts', by lag  
##   
## 0.000 0.143 0.286 0.429 0.571 0.714 0.857 1.000 1.143 1.286 1.429 1.571   
## 1.000 0.571 0.404 0.330 0.313 0.328 0.431 0.483 0.388 0.283 0.273 0.266   
## 1.714 1.857 2.000 2.143 2.286 2.429 2.571 2.714 2.857 3.000 3.143 3.286   
## 0.267 0.352 0.408 0.339 0.267 0.220 0.190 0.175 0.235 0.296 0.236 0.200   
## 3.429 3.571 3.714 3.857 4.000 4.143 4.286   
## 0.208 0.199 0.195 0.221 0.265 0.221 0.179

acf(vts, plot=T, main = 'ACF on Daily Data')



pacf(vts, plot=T, main = 'PACF on Daily Data')



# Ljung Box Test

Box.test(vts, lag= 10, type = 'Ljung')

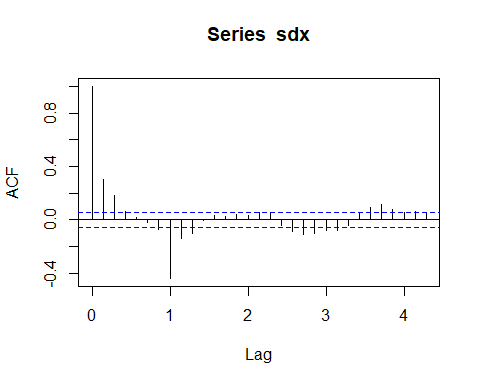
##   
## Box-Ljung test  
##   
## data: vts  
## X-squared = 1778.2, df = 10, p-value < 2.2e-16

adfTest(vts, lags = 180, type = c("c"))

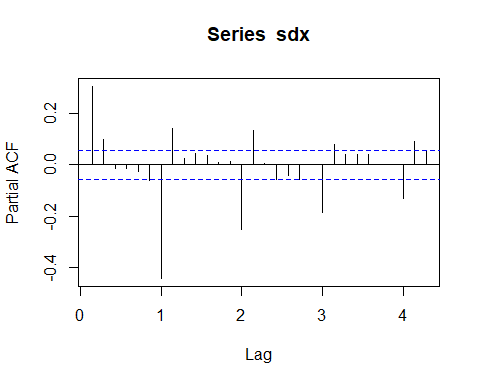
##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 180  
## STATISTIC:  
## Dickey-Fuller: -1.4601  
## P VALUE:  
## 0.5149   
##   
## Description:  
## Sat Mar 10 13:40:02 2018 by user: guy.dor

# seasonal differencing

sdx = diff(vts, 7)  
acf(sdx, plot = T)



pacf(sdx, plot = T)



adfTest(sdx, lags = 180, type = c("c"))

## Warning in adfTest(sdx, lags = 180, type = c("c")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 180  
## STATISTIC:  
## Dickey-Fuller: -3.6982  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:40:03 2018 by user: guy.dor

# Try automated selection

auto.arima(sdx, ic =c("bic"), trace=TRUE)

##   
## ARIMA(2,0,2)(1,0,1)[7] with non-zero mean : 17779.32  
## ARIMA(0,0,0) with non-zero mean : 18383.93  
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean : 18009.39  
## ARIMA(0,0,1)(0,0,1)[7] with non-zero mean : 17885.13  
## ARIMA(0,0,0) with zero mean : 18377.15  
## ARIMA(2,0,2)(0,0,1)[7] with non-zero mean : 17787.67  
## ARIMA(2,0,2)(2,0,1)[7] with non-zero mean : Inf  
## ARIMA(2,0,2)(1,0,0)[7] with non-zero mean : 18019.87  
## ARIMA(2,0,2)(1,0,2)[7] with non-zero mean : 17783.7  
## ARIMA(2,0,2) with non-zero mean : 18267.87  
## ARIMA(2,0,2)(2,0,2)[7] with non-zero mean : Inf  
## ARIMA(1,0,2)(1,0,1)[7] with non-zero mean : 17776.25  
## ARIMA(1,0,1)(1,0,1)[7] with non-zero mean : 17795.25  
## ARIMA(1,0,3)(1,0,1)[7] with non-zero mean : 17775.68  
## ARIMA(0,0,2)(1,0,1)[7] with non-zero mean : 17839.47  
## ARIMA(2,0,4)(1,0,1)[7] with non-zero mean : 17791.87  
## ARIMA(1,0,3)(1,0,1)[7] with zero mean : 17769.14  
## ARIMA(1,0,3)(0,0,1)[7] with zero mean : 17760.2  
## ARIMA(1,0,3) with zero mean : 18280.47  
## ARIMA(1,0,3)(0,0,2)[7] with zero mean : 17761.16  
## ARIMA(1,0,3)(1,0,2)[7] with zero mean : 17762.88  
## ARIMA(0,0,3)(0,0,1)[7] with zero mean : 17819.76  
## ARIMA(2,0,3)(0,0,1)[7] with zero mean : 17759.51  
## ARIMA(2,0,2)(0,0,1)[7] with zero mean : 17782.77  
## ARIMA(2,0,4)(0,0,1)[7] with zero mean : 17765.11  
## ARIMA(1,0,2)(0,0,1)[7] with zero mean : 17760.07  
## ARIMA(3,0,4)(0,0,1)[7] with zero mean : 17754.81  
## ARIMA(3,0,4)(0,0,1)[7] with non-zero mean : 17761.35  
## ARIMA(3,0,4)(1,0,1)[7] with zero mean : 17785.39  
## ARIMA(3,0,4) with zero mean : Inf  
## ARIMA(3,0,4)(0,0,2)[7] with zero mean : Inf  
## ARIMA(3,0,4)(1,0,2)[7] with zero mean : Inf  
## ARIMA(4,0,4)(0,0,1)[7] with zero mean : 17765.26  
## ARIMA(3,0,3)(0,0,1)[7] with zero mean : 17755.64  
## ARIMA(3,0,5)(0,0,1)[7] with zero mean : 17757.38  
## ARIMA(4,0,5)(0,0,1)[7] with zero mean : 17769.97  
## ARIMA(2,0,2)(1,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,0) with non-zero mean : 18383.93  
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean : 18010.5  
## ARIMA(0,0,1)(0,0,1)[7] with non-zero mean : 17881.94  
## ARIMA(0,0,0) with zero mean : 18377.15  
## ARIMA(0,0,1)(1,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,1) with non-zero mean : 18301.9  
## ARIMA(0,0,1)(0,0,2)[7] with non-zero mean : Inf  
## ARIMA(0,0,1)(1,0,2)[7] with non-zero mean : Inf  
## ARIMA(1,0,1)(0,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,0)(0,0,1)[7] with non-zero mean : 18042.76  
## ARIMA(0,0,2)(0,0,1)[7] with non-zero mean : 17832.15  
## ARIMA(1,0,3)(0,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,2)(0,0,1)[7] with zero mean : 17827.26  
## ARIMA(0,0,2)(1,0,1)[7] with zero mean : Inf  
## ARIMA(0,0,2) with zero mean : 18270.12  
## ARIMA(0,0,2)(0,0,2)[7] with zero mean : 17816.69  
## ARIMA(1,0,2)(0,0,2)[7] with zero mean : Inf  
## ARIMA(0,0,1)(0,0,2)[7] with zero mean : 17873.23  
## ARIMA(0,0,3)(0,0,2)[7] with zero mean : Inf  
## ARIMA(1,0,3)(0,0,2)[7] with zero mean : Inf  
## ARIMA(0,0,2)(0,0,2)[7] with non-zero mean : Inf  
## ARIMA(0,0,2)(1,0,2)[7] with zero mean : Inf  
##   
## Best model: ARIMA(0,0,2)(0,0,2)[7] with zero mean   
##   
##   
##   
## Best model: ARIMA(0,0,2)(0,0,2)[7] with zero mean

## Series: sdx   
## ARIMA(0,0,2)(0,0,2)[7] with zero mean   
##   
## Coefficients:  
## ma1 ma2 sma1 sma2  
## 0.3877 0.2206 -0.7642 -0.1474  
## s.e. 0.0301 0.0274 0.0288 0.0301  
##   
## sigma^2 estimated as 305850: log likelihood=-8890.73  
## AIC=17791.46 AICc=17791.51 BIC=17816.69

m1=Arima(sdx,order=c(0,0,3),seasonal=list(order=c(0,0,1),period=7), method="ML")  
m1

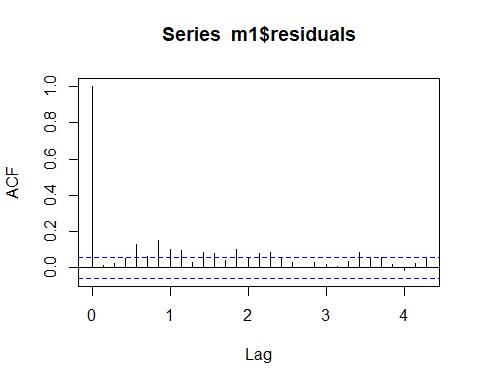
## Series: sdx   
## ARIMA(0,0,3)(0,0,1)[7] with non-zero mean   
##   
## Coefficients:  
## ma1 ma2 ma3 sma1 intercept  
## 0.4092 0.2648 0.1351 -0.8491 -7.2482  
## s.e. 0.0307 0.0302 0.0294 0.0388 4.6341  
##   
## sigma^2 estimated as 305075: log likelihood=-8887.4  
## AIC=17786.8 AICc=17786.88 BIC=17817.08

coeftest(m1)

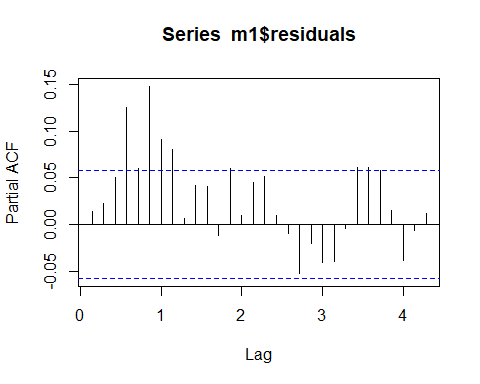
##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 0.409243 0.030733 13.3162 < 2.2e-16 \*\*\*  
## ma2 0.264780 0.030207 8.7654 < 2.2e-16 \*\*\*  
## ma3 0.135104 0.029352 4.6029 4.167e-06 \*\*\*  
## sma1 -0.849101 0.038780 -21.8952 < 2.2e-16 \*\*\*  
## intercept -7.248229 4.634150 -1.5641 0.1178   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Residual Analysis

acf(m1$residuals)



pacf(m1$residuals)

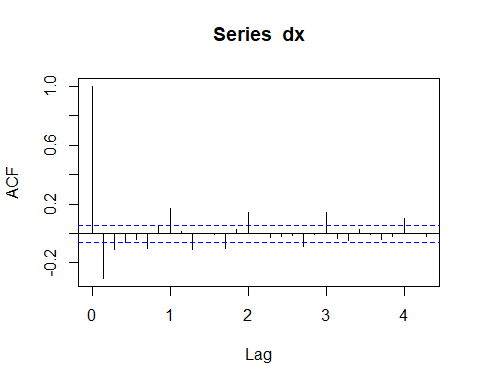


Box.test(m1$residuals, lag = 10, type = 'Ljung')

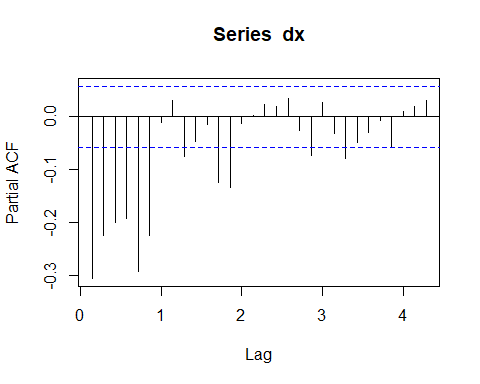
##   
## Box-Ljung test  
##   
## data: m1$residuals  
## X-squared = 86.312, df = 10, p-value = 2.875e-14

# found autocorrelation

dx = diff(vts)  
acf(dx, plot = T)



pacf(dx, plot = T)



Box.test(dx, lag= 10, type = 'Ljung')

##   
## Box-Ljung test  
##   
## data: dx  
## X-squared = 192.78, df = 10, p-value < 2.2e-16

adfTest(dx, lags = 180, type = c("c"))

## Warning in adfTest(dx, lags = 180, type = c("c")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 180  
## STATISTIC:  
## Dickey-Fuller: -3.6098  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:40:10 2018 by user: guy.dor

auto.arima(dx, max.p = 10, max.q = 10, ic =c("bic"), trace=TRUE)

##   
## ARIMA(2,0,2)(1,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,0) with non-zero mean : 18275.59  
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean : 18109.06  
## ARIMA(0,0,1)(0,0,1)[7] with non-zero mean : 17965.97  
## ARIMA(0,0,0) with zero mean : 18268.55  
## ARIMA(0,0,1)(1,0,1)[7] with non-zero mean : 17898.78  
## ARIMA(0,0,1)(1,0,0)[7] with non-zero mean : 17948.71  
## ARIMA(0,0,1)(1,0,2)[7] with non-zero mean : Inf  
## ARIMA(0,0,1) with non-zero mean : 17994.32  
## ARIMA(0,0,1)(2,0,2)[7] with non-zero mean : Inf  
## ARIMA(1,0,1)(1,0,1)[7] with non-zero mean : 17871.47  
## ARIMA(1,0,0)(1,0,1)[7] with non-zero mean : 18056.29  
## ARIMA(1,0,2)(1,0,1)[7] with non-zero mean : 17875.31  
## ARIMA(0,0,0)(1,0,1)[7] with non-zero mean : 18213.79  
## ARIMA(1,0,1)(1,0,1)[7] with zero mean : 17864.86  
## ARIMA(1,0,1)(0,0,1)[7] with zero mean : 17873.17  
## ARIMA(1,0,1)(2,0,1)[7] with zero mean : 17842.68  
## ARIMA(1,0,1)(2,0,0)[7] with zero mean : 17849.48  
## ARIMA(1,0,1)(2,0,2)[7] with zero mean : Inf  
## ARIMA(1,0,1)(1,0,0)[7] with zero mean : 17868.55  
## ARIMA(0,0,1)(2,0,1)[7] with zero mean : 17905.58  
## ARIMA(2,0,1)(2,0,1)[7] with zero mean : 17848.82  
## ARIMA(1,0,0)(2,0,1)[7] with zero mean : 18068.35  
## ARIMA(1,0,2)(2,0,1)[7] with zero mean : 17849.24  
## ARIMA(0,0,0)(2,0,1)[7] with zero mean : 18212.95  
## ARIMA(2,0,2)(2,0,1)[7] with zero mean : 17856.96  
## ARIMA(1,0,1)(2,0,1)[7] with non-zero mean : 17849.64  
## ARIMA(2,0,2)(1,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,0) with non-zero mean : 18275.59  
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean : 18111.69  
## ARIMA(0,0,1)(0,0,1)[7] with non-zero mean : 17963.2  
## ARIMA(0,0,0) with zero mean : 18268.55  
## ARIMA(0,0,1)(1,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,1) with non-zero mean : 17991.79  
## ARIMA(0,0,1)(0,0,2)[7] with non-zero mean : 17940.65  
## ARIMA(1,0,1)(0,0,2)[7] with non-zero mean : 17864.84  
## ARIMA(1,0,0)(0,0,2)[7] with non-zero mean : 18110.69  
## ARIMA(1,0,2)(0,0,2)[7] with non-zero mean : 17871.7  
## ARIMA(0,0,0)(0,0,2)[7] with non-zero mean : 18247.97  
## ARIMA(2,0,2)(0,0,2)[7] with non-zero mean : 17875.49  
## ARIMA(1,0,1)(0,0,2)[7] with zero mean : 17858.27  
## ARIMA(1,0,1)(1,0,2)[7] with zero mean : Inf  
## ARIMA(1,0,1)(0,0,1)[7] with zero mean : 17872.79  
## ARIMA(0,0,1)(0,0,2)[7] with zero mean : 17933.74  
## ARIMA(2,0,1)(0,0,2)[7] with zero mean : 17865.09  
## ARIMA(1,0,0)(0,0,2)[7] with zero mean : 18103.65  
## ARIMA(1,0,2)(0,0,2)[7] with zero mean : 17865.14  
## ARIMA(0,0,0)(0,0,2)[7] with zero mean : 18240.93  
## ARIMA(2,0,2)(0,0,2)[7] with zero mean : 17868.92  
##   
## Best model: ARIMA(1,0,1)(0,0,2)[7] with zero mean   
##   
##   
##   
## Best model: ARIMA(1,0,1)(0,0,2)[7] with zero mean

## Series: dx   
## ARIMA(1,0,1)(0,0,2)[7] with zero mean   
##   
## Coefficients:  
## ar1 ma1 sma1 sma2  
## 0.3199 -0.9298 0.1735 0.1284  
## s.e. 0.0330 0.0142 0.0303 0.0274  
##   
## sigma^2 estimated as 295213: log likelihood=-8911.51  
## AIC=17833.01 AICc=17833.06 BIC=17858.27

m2=Arima(dx,order=c(1,0,1),seasonal=list(order=c(2,0,0),period=7), method="ML")  
m2

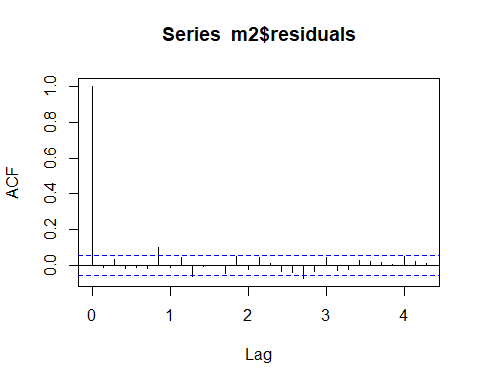
## Series: dx   
## ARIMA(1,0,1)(2,0,0)[7] with non-zero mean   
##   
## Coefficients:  
## ar1 ma1 sar1 sar2 intercept  
## 0.3271 -0.9365 0.1880 0.1366 -1.5180  
## s.e. 0.0333 0.0149 0.0303 0.0294 2.2376  
##   
## sigma^2 estimated as 292379: log likelihood=-8905.52  
## AIC=17823.04 AICc=17823.12 BIC=17853.36

coeftest(m2)

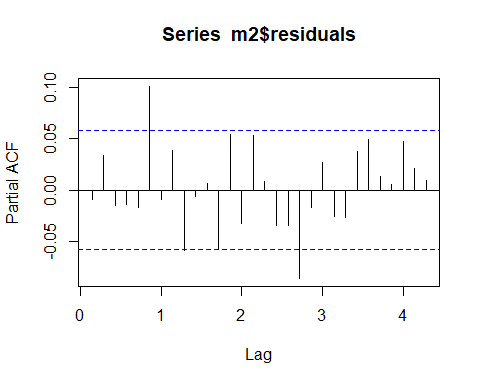
##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 0.327098 0.033330 9.8139 < 2.2e-16 \*\*\*  
## ma1 -0.936516 0.014941 -62.6789 < 2.2e-16 \*\*\*  
## sar1 0.188023 0.030335 6.1983 5.706e-10 \*\*\*  
## sar2 0.136584 0.029396 4.6464 3.378e-06 \*\*\*  
## intercept -1.518027 2.237586 -0.6784 0.4975   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Residual analysis

acf(m2$residuals)



pacf(m2$residuals)



Box.test(m2$residuals, lag = 6, type = 'Ljung', fitdf=4)

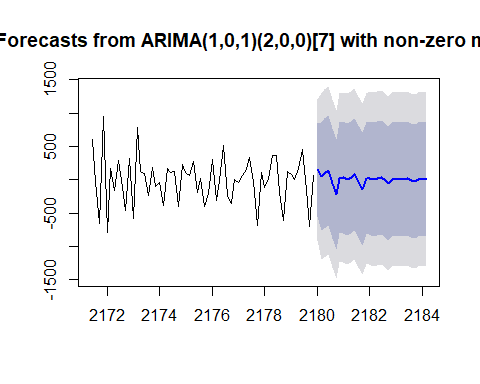
##   
## Box-Ljung test  
##   
## data: m2$residuals  
## X-squared = 13.87, df = 2, p-value = 0.000973

# Forecast and Backtesting

source("backtest.R")  
backtest(m2, dx, h=1, orig=length(vts)\*0.8)

## [1] "RMSE of out-of-sample forecasts"  
## [1] 401.4255  
## [1] "Mean absolute error of out-of-sample forecasts"  
## [1] 301.0132  
## [1] "Mean Absolute Percentage error"  
## [1] 2.240109  
## [1] "Symmetric Mean Absolute Percentage error"  
## [1] 1.215878

m2f=forecast(m2,h=30)  
plot(forecast(m2f, h = 30), include = 60)



m5=Arima(dx,order=c(1,0,7),seasonal=list(order=c(2,0,0),period=7), method="ML")  
m5

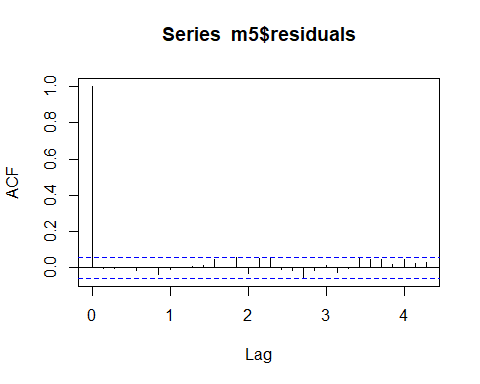
## Series: dx   
## ARIMA(1,0,7)(2,0,0)[7] with non-zero mean   
##   
## Coefficients:  
## ar1 ma1 ma2 ma3 ma4 ma5 ma6  
## -0.4972 -0.1185 -0.4753 -0.1869 -0.0693 -0.0444 0.1390  
## s.e. 0.0684 0.0677 0.0453 0.0360 0.0320 0.0329 0.0328  
## ma7 sar1 sar2 intercept  
## -0.2444 0.4031 0.1386 -0.8093  
## s.e. 0.0381 0.0379 0.0298 0.1986  
##   
## sigma^2 estimated as 283174: log likelihood=-8885.98  
## AIC=17795.95 AICc=17796.23 BIC=17856.57

coeftest(m5)

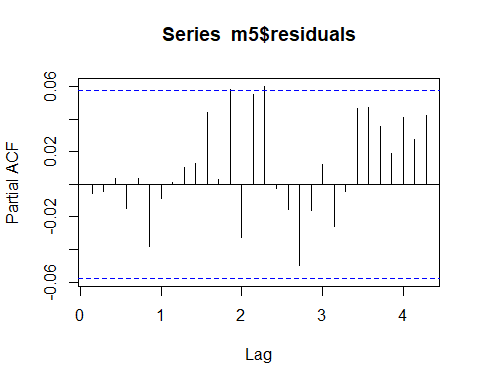
##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.497182 0.068392 -7.2696 3.605e-13 \*\*\*  
## ma1 -0.118541 0.067682 -1.7515 0.07987 .   
## ma2 -0.475337 0.045250 -10.5046 < 2.2e-16 \*\*\*  
## ma3 -0.186937 0.035973 -5.1966 2.030e-07 \*\*\*  
## ma4 -0.069298 0.031968 -2.1678 0.03018 \*   
## ma5 -0.044442 0.032867 -1.3522 0.17633   
## ma6 0.138972 0.032753 4.2430 2.205e-05 \*\*\*  
## ma7 -0.244413 0.038059 -6.4220 1.345e-10 \*\*\*  
## sar1 0.403103 0.037857 10.6480 < 2.2e-16 \*\*\*  
## sar2 0.138587 0.029764 4.6561 3.222e-06 \*\*\*  
## intercept -0.809326 0.198640 -4.0743 4.615e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Residual Analysis

acf(m5$residuals)



pacf(m5$residuals)



Box.test(m5$residuals, lag = 22, type = 'Ljung', fitdf=8)

##   
## Box-Ljung test  
##   
## data: m5$residuals  
## X-squared = 22.282, df = 14, p-value = 0.07301

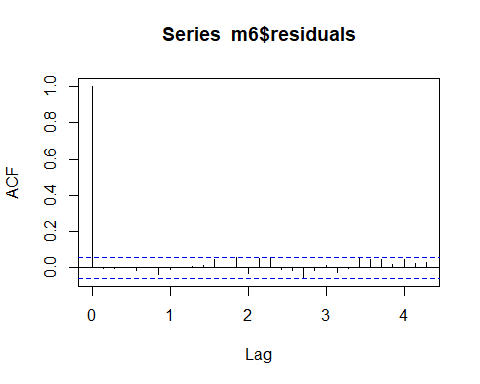
m6=Arima(vts,order=c(1,1,7),seasonal=list(order=c(2,0,0),period=7), method="ML", include.constant = TRUE)  
m6

## Series: vts   
## ARIMA(1,1,7)(2,0,0)[7] with drift   
##   
## Coefficients:  
## ar1 ma1 ma2 ma3 ma4 ma5 ma6 ma7  
## -0.4972 -0.1185 -0.4754 -0.187 -0.0693 -0.0444 0.1390 -0.2444  
## s.e. 0.0684 0.0677 0.0452 0.036 0.0320 0.0329 0.0328 0.0381  
## sar1 sar2 drift  
## 0.4031 0.1386 -0.8093  
## s.e. 0.0379 0.0298 0.1986  
##   
## sigma^2 estimated as 283174: log likelihood=-8885.98  
## AIC=17795.95 AICc=17796.23 BIC=17856.57

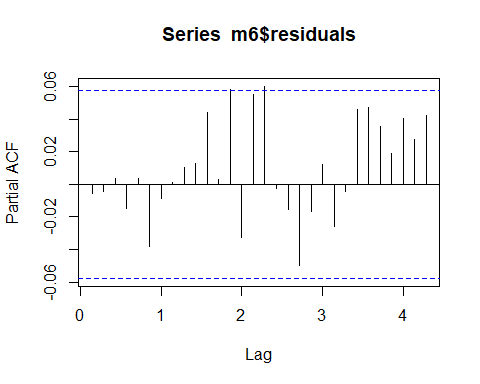
coeftest(m6)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.497231 0.068383 -7.2712 3.562e-13 \*\*\*  
## ma1 -0.118481 0.067674 -1.7507 0.07999 .   
## ma2 -0.475354 0.045247 -10.5058 < 2.2e-16 \*\*\*  
## ma3 -0.186977 0.035972 -5.1979 2.016e-07 \*\*\*  
## ma4 -0.069305 0.031968 -2.1679 0.03016 \*   
## ma5 -0.044449 0.032867 -1.3524 0.17625   
## ma6 0.138977 0.032753 4.2431 2.204e-05 \*\*\*  
## ma7 -0.244410 0.038060 -6.4218 1.347e-10 \*\*\*  
## sar1 0.403083 0.037856 10.6477 < 2.2e-16 \*\*\*  
## sar2 0.138600 0.029764 4.6566 3.215e-06 \*\*\*  
## drift -0.809335 0.198637 -4.0744 4.613e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

acf(m6$residuals)



pacf(m6$residuals)



Box.test(m6$residuals, lag = 22, type = 'Ljung', fitdf=8)

##   
## Box-Ljung test  
##   
## data: m6$residuals  
## X-squared = 22.298, df = 14, p-value = 0.0727

# Forecast and Backtesting

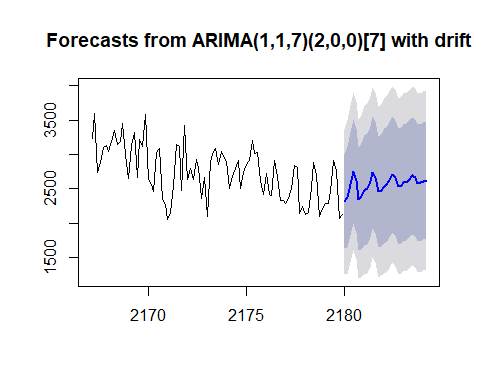
source("backtest.R")  
backtest(m6, vts, h=1, orig=length(dx)\*0.8, inc.drift=TRUE)

## [1] "RMSE of out-of-sample forecasts"  
## [1] 411.2558  
## [1] "Mean absolute error of out-of-sample forecasts"  
## [1] 311.0098  
## [1] "Mean Absolute Percentage error"  
## [1] 0.163914  
## [1] "Symmetric Mean Absolute Percentage error"  
## [1] 0.1206297

f1=forecast(m6,h=30)  
f1

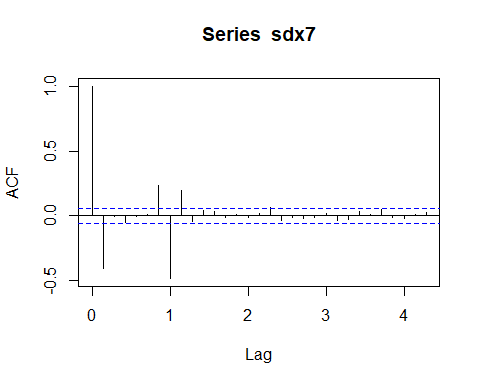
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 2180.000 2314.270 1632.015 2996.525 1270.851 3357.689  
## 2180.143 2380.032 1648.927 3111.137 1261.904 3498.161  
## 2180.286 2571.023 1825.179 3316.867 1430.353 3711.692  
## 2180.429 2755.064 2005.205 3504.923 1608.254 3901.874  
## 2180.571 2627.375 1874.690 3380.060 1476.242 3778.507  
## 2180.714 2353.710 1599.907 3107.514 1200.867 3506.553  
## 2180.857 2388.051 1619.871 3156.232 1213.220 3562.882  
## 2181.000 2477.747 1683.149 3272.345 1262.515 3692.979  
## 2181.143 2501.932 1694.507 3309.358 1267.081 3736.783  
## 2181.286 2603.079 1794.533 3411.626 1366.514 3839.645  
## 2181.429 2738.987 1929.384 3548.589 1500.806 3977.167  
## 2181.571 2667.532 1857.605 3477.460 1428.856 3906.209  
## 2181.714 2459.176 1649.005 3269.346 1220.126 3698.225  
## 2181.857 2481.291 1668.926 3293.656 1238.886 3723.696  
## 2182.000 2543.144 1711.471 3374.816 1271.211 3815.077  
## 2182.143 2561.649 1724.559 3398.738 1281.431 3841.866  
## 2182.286 2628.513 1790.522 3466.504 1346.917 3910.110  
## 2182.429 2708.435 1869.962 3546.909 1426.100 3990.770  
## 2182.571 2661.563 1822.865 3500.261 1378.886 3944.241  
## 2182.714 2539.278 1700.451 3378.106 1256.402 3822.154  
## 2182.857 2552.581 1712.510 3392.652 1267.804 3837.358  
## 2183.000 2589.574 1743.439 3435.708 1295.523 3883.625  
## 2183.143 2600.014 1751.802 3448.225 1302.786 3897.242  
## 2183.286 2640.614 1792.100 3489.128 1342.924 3938.304  
## 2183.429 2691.295 1842.584 3540.005 1393.304 3989.285  
## 2183.571 2662.127 1813.332 3510.922 1364.006 3960.248  
## 2183.714 2583.587 1734.736 3432.437 1285.382 3881.792  
## 2183.857 2591.643 1742.338 3440.948 1292.743 3890.543  
## 2184.000 2614.756 1762.845 3466.667 1311.870 3917.642  
## 2184.143 2621.158 1768.391 3473.926 1316.963 3925.353

plot(f1, include =90)

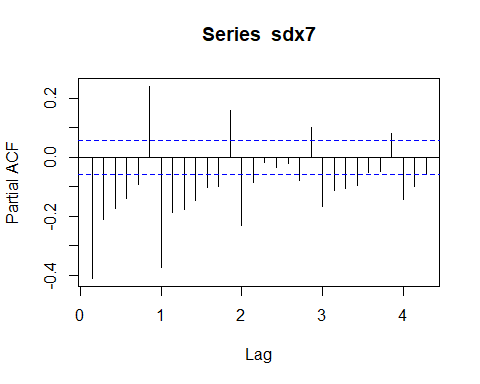


# Seasonal differencing

sdx7 = diff(dx,7)  
acf(sdx7, plot = T)



pacf(sdx7, plot = T)



adfTest(sdx7, lags = 180, type = c("c"))

## Warning in adfTest(sdx7, lags = 180, type = c("c")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 180  
## STATISTIC:  
## Dickey-Fuller: -4.1536  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:44:58 2018 by user: guy.dor

# Automated selection

auto.arima(sdx7, max.p = 10, max.q = 10, ic =c("bic"), trace=TRUE)

##   
## ARIMA(2,0,2)(1,0,1)[7] with non-zero mean : 17781.8  
## ARIMA(0,0,0) with non-zero mean : 18744.09  
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean : 18258.46  
## ARIMA(0,0,1)(0,0,1)[7] with non-zero mean : 17812.63  
## ARIMA(0,0,0) with zero mean : 18737.04  
## ARIMA(2,0,2)(0,0,1)[7] with non-zero mean : 17745.36  
## ARIMA(2,0,2) with non-zero mean : Inf  
## ARIMA(2,0,2)(0,0,2)[7] with non-zero mean : Inf  
## ARIMA(2,0,2)(1,0,2)[7] with non-zero mean : 17772.27  
## ARIMA(1,0,2)(0,0,1)[7] with non-zero mean : 17795.11  
## ARIMA(3,0,2)(0,0,1)[7] with non-zero mean : 17756.68  
## ARIMA(2,0,1)(0,0,1)[7] with non-zero mean : 17783.82  
## ARIMA(2,0,3)(0,0,1)[7] with non-zero mean : 17752.28  
## ARIMA(1,0,1)(0,0,1)[7] with non-zero mean : 17811.67  
## ARIMA(3,0,3)(0,0,1)[7] with non-zero mean : 17763.44  
## ARIMA(2,0,2)(0,0,1)[7] with zero mean : 17738.35  
## ARIMA(2,0,2)(1,0,1)[7] with zero mean : 17770.57  
## ARIMA(2,0,2) with zero mean : Inf  
## ARIMA(2,0,2)(0,0,2)[7] with zero mean : Inf  
## ARIMA(2,0,2)(1,0,2)[7] with zero mean : 17765.23  
## ARIMA(1,0,2)(0,0,1)[7] with zero mean : 17788.55  
## ARIMA(3,0,2)(0,0,1)[7] with zero mean : 17749.84  
## ARIMA(2,0,1)(0,0,1)[7] with zero mean : 17777.23  
## ARIMA(2,0,3)(0,0,1)[7] with zero mean : 17745.28  
## ARIMA(1,0,1)(0,0,1)[7] with zero mean : 17805.39  
## ARIMA(3,0,3)(0,0,1)[7] with zero mean : 17751.49  
## ARIMA(2,0,2)(1,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,0) with non-zero mean : 18744.09  
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean : 18258.84  
## ARIMA(0,0,1)(0,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,0) with zero mean : 18737.04  
## ARIMA(1,0,0) with non-zero mean : 18539.69  
## ARIMA(1,0,0)(2,0,0)[7] with non-zero mean : 18161.53  
## ARIMA(1,0,0)(2,0,1)[7] with non-zero mean : Inf  
## ARIMA(0,0,0)(2,0,0)[7] with non-zero mean : 18323.04  
## ARIMA(2,0,0)(2,0,0)[7] with non-zero mean : 18120.61  
## ARIMA(2,0,1)(2,0,0)[7] with non-zero mean : Inf  
## ARIMA(3,0,1)(2,0,0)[7] with non-zero mean : Inf  
## ARIMA(2,0,0)(2,0,0)[7] with zero mean : 18113.56  
## ARIMA(2,0,0)(1,0,0)[7] with zero mean : 18201.06  
## ARIMA(2,0,0)(2,0,1)[7] with zero mean : Inf  
## ARIMA(1,0,0)(2,0,0)[7] with zero mean : 18154.48  
## ARIMA(3,0,0)(2,0,0)[7] with zero mean : 18078.61  
## ARIMA(3,0,1)(2,0,0)[7] with zero mean : Inf  
## ARIMA(4,0,1)(2,0,0)[7] with zero mean : Inf  
## ARIMA(3,0,0)(2,0,0)[7] with non-zero mean : 18085.66  
## ARIMA(3,0,0)(1,0,0)[7] with zero mean : 18161.84  
## ARIMA(3,0,0)(2,0,1)[7] with zero mean : Inf  
## ARIMA(4,0,0)(2,0,0)[7] with zero mean : 18050.42  
## ARIMA(5,0,1)(2,0,0)[7] with zero mean : Inf  
## ARIMA(4,0,0)(2,0,0)[7] with non-zero mean : 18057.47  
## ARIMA(4,0,0)(1,0,0)[7] with zero mean : 18129.34  
## ARIMA(4,0,0)(2,0,1)[7] with zero mean : Inf  
## ARIMA(5,0,0)(2,0,0)[7] with zero mean : 18015.73  
## ARIMA(6,0,1)(2,0,0)[7] with zero mean : Inf  
## ARIMA(5,0,0)(2,0,0)[7] with non-zero mean : 18022.78  
## ARIMA(5,0,0)(1,0,0)[7] with zero mean : 18107.22  
## ARIMA(5,0,0)(2,0,1)[7] with zero mean : Inf  
## ARIMA(6,0,0)(2,0,0)[7] with zero mean : 18017.41  
##   
## Best model: ARIMA(5,0,0)(2,0,0)[7] with zero mean   
##   
##   
##   
## Best model: ARIMA(5,0,0)(2,0,0)[7] with zero mean

## Series: sdx7   
## ARIMA(5,0,0)(2,0,0)[7] with zero mean   
##   
## Coefficients:  
## ar1 ar2 ar3 ar4 ar5 sar1 sar2  
## -0.5507 -0.3941 -0.3432 -0.2761 -0.1920 -0.6650 -0.2919  
## s.e. 0.0292 0.0326 0.0331 0.0326 0.0294 0.0288 0.0287  
##   
## sigma^2 estimated as 365435: log likelihood=-8979.68  
## AIC=17975.37 AICc=17975.49 BIC=18015.73

m3=Arima(sdx7,order=c(1,0,0),seasonal=list(order=c(1,0,0),period=7), method="ML")  
m3

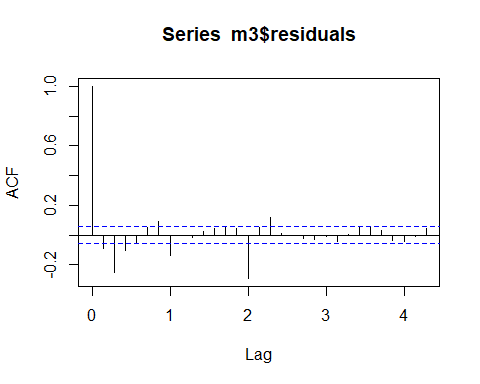
## Series: sdx7   
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean   
##   
## Coefficients:  
## ar1 sar1 intercept  
## -0.3928 -0.4709 -0.0644  
## s.e. 0.0272 0.0260 9.8014  
##   
## sigma^2 estimated as 462049: log likelihood=-9115.33  
## AIC=18238.66 AICc=18238.69 BIC=18258.84

coeftest(m3)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.392816 0.027183 -14.4507 <2e-16 \*\*\*  
## sar1 -0.470890 0.025986 -18.1210 <2e-16 \*\*\*  
## intercept -0.064401 9.801387 -0.0066 0.9948   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Residual Analysis

acf(m3$residuals)



Box.test(m3$residuals, lag = 12, type = 'Ljung', fitdf = 2)

##   
## Box-Ljung test  
##   
## data: m3$residuals  
## X-squared = 140.14, df = 10, p-value < 2.2e-16

m4=Arima(sdx7,order=c(8,0,0), method="ML")  
m4

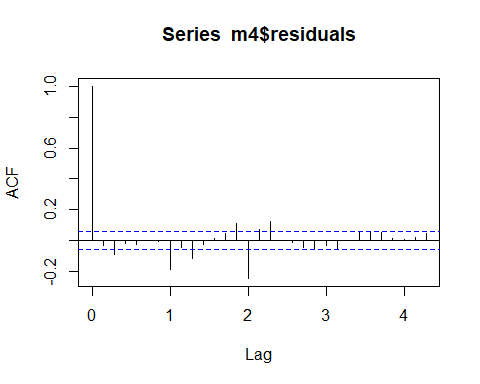
## Series: sdx7   
## ARIMA(8,0,0) with non-zero mean   
##   
## Coefficients:  
## ar1 ar2 ar3 ar4 ar5 ar6 ar7  
## -0.5267 -0.2902 -0.2647 -0.2164 -0.1216 -0.0187 -0.4578  
## s.e. 0.0290 0.0300 0.0311 0.0319 0.0319 0.0311 0.0299  
## ar8 intercept  
## -0.1862 -0.0324  
## s.e. 0.0290 6.2048  
##   
## sigma^2 estimated as 421241: log likelihood=-9059.36  
## AIC=18138.72 AICc=18138.92 BIC=18189.18

coeftest(m4)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.526711 0.029005 -18.1590 < 2.2e-16 \*\*\*  
## ar2 -0.290206 0.029963 -9.6855 < 2.2e-16 \*\*\*  
## ar3 -0.264671 0.031139 -8.4997 < 2.2e-16 \*\*\*  
## ar4 -0.216350 0.031944 -6.7728 1.263e-11 \*\*\*  
## ar5 -0.121599 0.031903 -3.8115 0.0001381 \*\*\*  
## ar6 -0.018723 0.031123 -0.6016 0.5474594   
## ar7 -0.457751 0.029894 -15.3126 < 2.2e-16 \*\*\*  
## ar8 -0.186223 0.028981 -6.4257 1.313e-10 \*\*\*  
## intercept -0.032447 6.204799 -0.0052 0.9958276   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Residual analysis

acf(m4$residuals)



Box.test(m4$residuals,lag=12,fitdf=8, type='Ljung')

##   
## Box-Ljung test  
##   
## data: m4$residuals  
## X-squared = 74.842, df = 4, p-value = 2.109e-15

# WEEKLY DATA

scv=read.table("WeeklyViolations.csv",header=T, sep='\t')   
  
head(scv)

## Week Violations  
## 1 6/29/2014 20969  
## 2 7/6/2014 29312  
## 3 7/13/2014 30143  
## 4 7/20/2014 28810  
## 5 7/27/2014 24745  
## 6 8/3/2014 22033

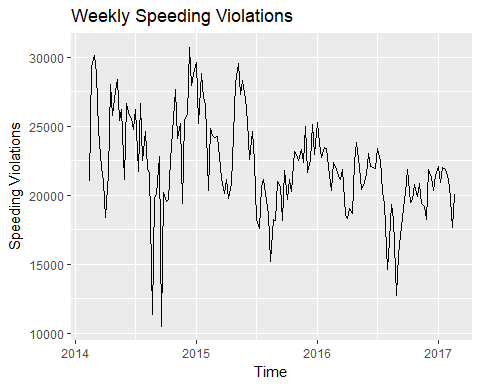
x=scv$Violations  
head(x)

## [1] 20969 29312 30143 28810 24745 22033

xts=ts(x,frequency=52,start=c(2014,7), end = c(2017,8,20))

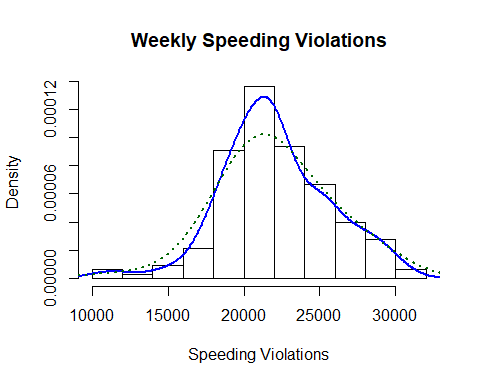
# Create time plot

autoplot(xts)+ylab("Speeding Violations")+ggtitle("Weekly Speeding Violations")



# Check Normality

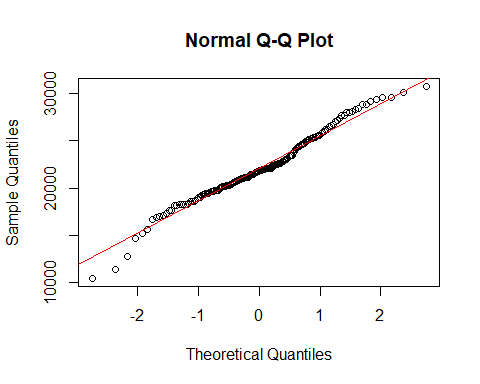
hist(x,main="Weekly Speeding Violations" , probability = TRUE, xlab = "Speeding Violations")  
lines(density(x), col="blue", lwd=2)  
lines(density(x, adjust=2), lty="dotted", col="darkgreen", lwd=2)



normalTest(x,method=c("jb"))

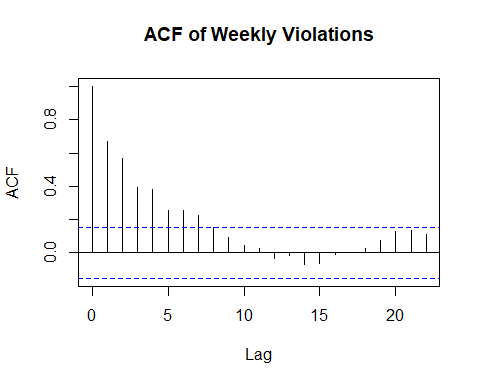
##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 1.8323  
## P VALUE:  
## Asymptotic p Value: 0.4001   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

qqnorm(x)  
qqline(x, col = 2)

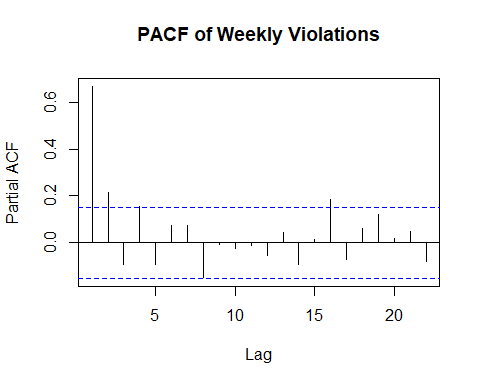


# ACF and PACF of Data

acf(x, main="ACF of Weekly Violations")



pacf(x,main="PACF of Weekly Violations")



# Dickey Fuller Tests

## Tests for AR model with time trend  
  
adfTest(xts, lags=1, type=c("ct"))

## Warning in adfTest(xts, lags = 1, type = c("ct")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 1  
## STATISTIC:  
## Dickey-Fuller: -4.6553  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(xts, lags=2, type=c("ct"))

## Warning in adfTest(xts, lags = 2, type = c("ct")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -4.9291  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(xts, lags=4, type=c("ct"))

## Warning in adfTest(xts, lags = 4, type = c("ct")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 4  
## STATISTIC:  
## Dickey-Fuller: -4.2688  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(xts, lags=17, type=c("ct"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 17  
## STATISTIC:  
## Dickey-Fuller: -3.3682  
## P VALUE:  
## 0.06236   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(xts, lags=52, type=c("ct"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 52  
## STATISTIC:  
## Dickey-Fuller: -2.7452  
## P VALUE:  
## 0.2656   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

## Tests for AR model with no time trend  
  
adfTest(x, lags=1, type=c("c"))

## Warning in adfTest(x, lags = 1, type = c("c")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 1  
## STATISTIC:  
## Dickey-Fuller: -4.0021  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(x, lags=2, type=c("c"))

## Warning in adfTest(x, lags = 2, type = c("c")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -4.2967  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(x, lags=4, type=c("c"))

## Warning in adfTest(x, lags = 4, type = c("c")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 4  
## STATISTIC:  
## Dickey-Fuller: -3.5281  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(x, lags=17, type=c("c"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 17  
## STATISTIC:  
## Dickey-Fuller: -2.2146  
## P VALUE:  
## 0.234   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(x, lags=52, type=c("c"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 52  
## STATISTIC:  
## Dickey-Fuller: 0.1664  
## P VALUE:  
## 0.968   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

# Ljung Box Tests

Box.test(x,lag = 1, type="Ljung-Box", fitdf=1)

##   
## Box-Ljung test  
##   
## data: x  
## X-squared = 75.132, df = 0, p-value < 2.2e-16

Box.test(x,lag = 3, type="Ljung-Box", fitdf=2)

##   
## Box-Ljung test  
##   
## data: x  
## X-squared = 155.23, df = 1, p-value < 2.2e-16

Box.test(x,lag = 15, type="Ljung-Box", fitdf=2)

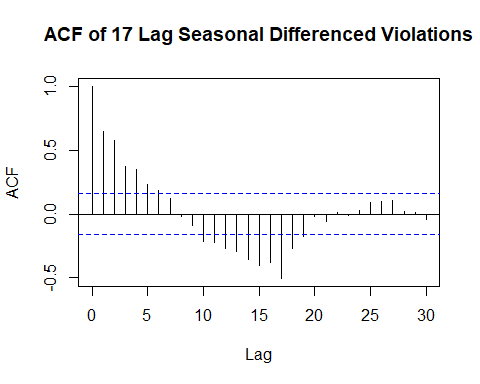
##   
## Box-Ljung test  
##   
## data: x  
## X-squared = 219.84, df = 13, p-value < 2.2e-16

# Applying Differencing to Data (s=17)

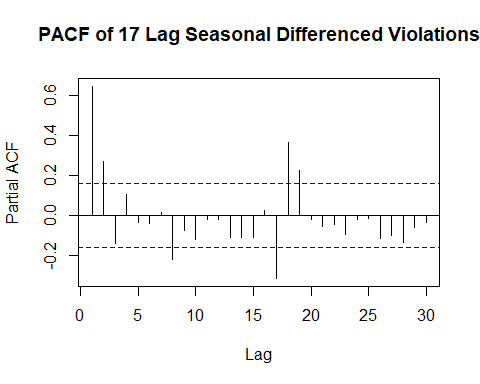
# compute seasonal difference for weeklydata (s=17)  
sd17x=diff(x,17)

# Create acf plot

acf(as.vector(sd17x),lag.max=30, main="ACF of 17 Lag Seasonal Differenced Violations")



pacf(as.vector(sd17x),lag.max=30, main="PACF of 17 Lag Seasonal Differenced Violations")



# Ljung Box Tests

Box.test(sd17x,lag = 1, type="Ljung-Box", fitdf=1)

##   
## Box-Ljung test  
##   
## data: sd17x  
## X-squared = 63.129, df = 0, p-value < 2.2e-16

Box.test(sd17x,lag = 3, type="Ljung-Box", fitdf=2)

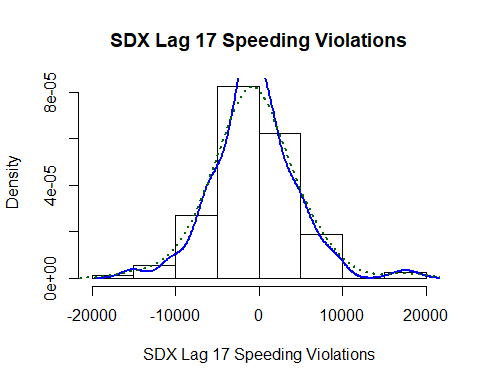
##   
## Box-Ljung test  
##   
## data: sd17x  
## X-squared = 134.46, df = 1, p-value < 2.2e-16

Box.test(sd17x,lag = 15, type="Ljung-Box", fitdf=2)

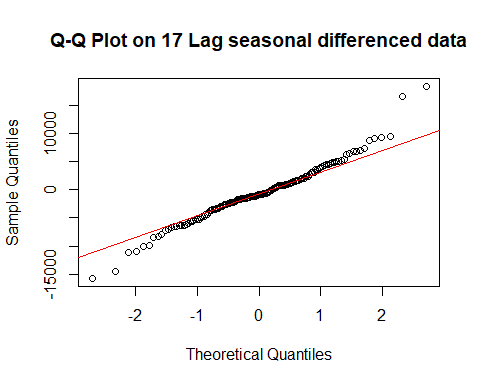
##   
## Box-Ljung test  
##   
## data: sd17x  
## X-squared = 259.88, df = 13, p-value < 2.2e-16

# Check Normality

hist(sd17x,main="SDX Lag 17 Speeding Violations" , probability = TRUE, xlab = "SDX Lag 17 Speeding Violations")  
lines(density(sd17x), col="blue", lwd=2)  
lines(density(sd17x, adjust=2), lty="dotted", col="darkgreen", lwd=2)



qqnorm(sd17x, main="Q-Q Plot on 17 Lag seasonal differenced data")  
qqline(sd17x, col = 2)

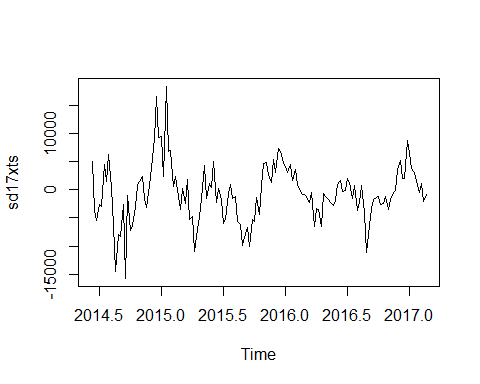


normalTest(sd17x,method=c("jb"))

##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 25.9945  
## P VALUE:  
## Asymptotic p Value: 2.267e-06   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

# Unit-root tests on Seasonal difference

sd17xts=diff(xts,17)  
plot(sd17xts)



adfTest(coredata(sd17xts), lags=2, type=c("c"))

## Warning in adfTest(coredata(sd17xts), lags = 2, type = c("c")): p-value  
## smaller than printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -4.0435  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(sd17xts, lags=2, type=c("c"))

## Warning in adfTest(sd17xts, lags = 2, type = c("c")): p-value smaller than  
## printed p-value

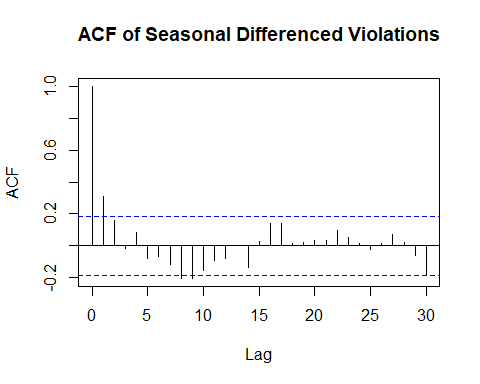
##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -4.0435  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

# Applying Differencing to Data (s=52)

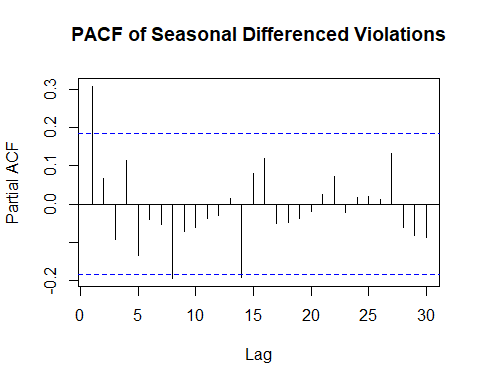
sdx=diff(x,52)

# Create acf plot

acf(as.vector(sdx),lag.max=30, main="ACF of Seasonal Differenced Violations")



pacf(as.vector(sdx),lag.max=30, main="PACF of Seasonal Differenced Violations")



# Ljung Box Tests

Box.test(sdx,lag = 1, type="Ljung-Box", fitdf=1)

##   
## Box-Ljung test  
##   
## data: sdx  
## X-squared = 10.969, df = 0, p-value < 2.2e-16

Box.test(sdx,lag = 3, type="Ljung-Box", fitdf=2)

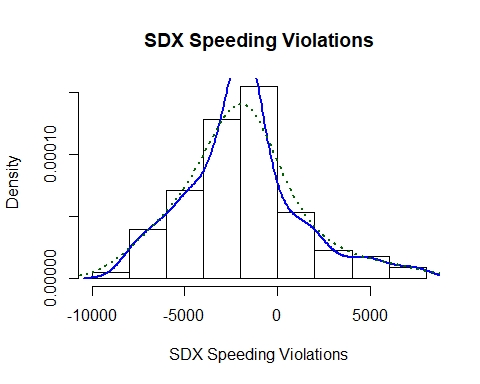
##   
## Box-Ljung test  
##   
## data: sdx  
## X-squared = 13.839, df = 1, p-value = 0.0001991

Box.test(sdx,lag = 15, type="Ljung-Box", fitdf=2)

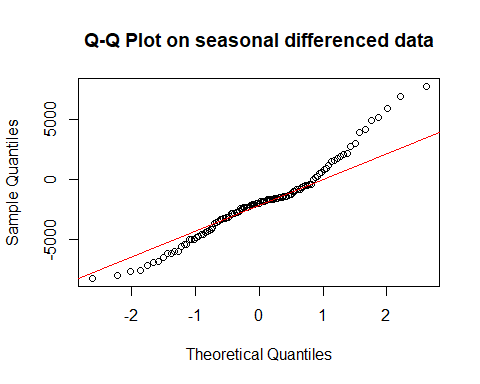
##   
## Box-Ljung test  
##   
## data: sdx  
## X-squared = 35.483, df = 13, p-value = 0.0007133

# Check Normality

hist(sdx,main="SDX Speeding Violations" , probability = TRUE, xlab = "SDX Speeding Violations")  
lines(density(sdx), col="blue", lwd=2)  
lines(density(sdx, adjust=2), lty="dotted", col="darkgreen", lwd=2)



qqnorm(sdx, main="Q-Q Plot on seasonal differenced data")  
qqline(sdx, col = 2)



basicStats(sdx)

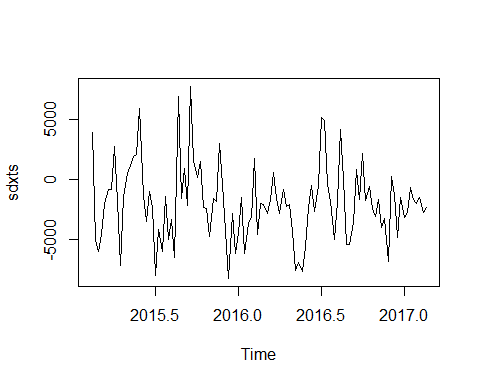
## sdx  
## nobs 1.130000e+02  
## NAs 0.000000e+00  
## Minimum -8.194000e+03  
## Maximum 7.737000e+03  
## 1. Quartile -3.584000e+03  
## 3. Quartile -6.910000e+02  
## Mean -1.908230e+03  
## Median -1.947000e+03  
## Sum -2.156300e+05  
## SE Mean 2.884426e+02  
## LCL Mean -2.479742e+03  
## UCL Mean -1.336718e+03  
## Variance 9.401502e+06  
## Stdev 3.066187e+03  
## Skewness 5.695570e-01  
## Kurtosis 7.877570e-01

normalTest(sdx,method=c("jb"))

##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 9.7221  
## P VALUE:  
## Asymptotic p Value: 0.007742   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

# Unit-root tests on Seasonal difference

sdxts=diff(xts,52)  
plot(sdxts)



adfTest(coredata(sdxts), lags=2, type=c("c"))

## Warning in adfTest(coredata(sdxts), lags = 2, type = c("c")): p-value  
## smaller than printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -5.2784  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:11 2018 by user: guy.dor

adfTest(sdxts, lags=2, type=c("c"))

## Warning in adfTest(sdxts, lags = 2, type = c("c")): p-value smaller than  
## printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -5.2784  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:45:12 2018 by user: guy.dor

# MODEL EVALUATION OF WEEKLY DATA

# try automated order selection

#Original data  
auto.arima(xts)

## Series: xts   
## ARIMA(3,1,2)(1,0,0)[52]   
##   
## Coefficients:  
## ar1 ar2 ar3 ma1 ma2 sar1  
## -0.086 0.4997 -0.0416 -0.3880 -0.5583 0.2289  
## s.e. 0.219 0.1190 0.1040 0.2027 0.1982 0.1124  
##   
## sigma^2 estimated as 6615352: log likelihood=-1454.66  
## AIC=2923.33 AICc=2924.08 BIC=2944.72

auto.arima(xts,stepwise=FALSE,approx=FALSE)

## Series: xts   
## ARIMA(2,1,2)(1,0,0)[52]   
##   
## Coefficients:  
## ar1 ar2 ma1 ma2 sar1  
## -0.1465 0.5234 -0.3415 -0.6058 0.2287  
## s.e. 0.1413 0.0880 0.1417 0.1342 0.1124  
##   
## sigma^2 estimated as 6579490: log likelihood=-1454.74  
## AIC=2921.47 AICc=2922.03 BIC=2939.81

#Differnced data(s=52)  
auto.arima(sdxts)

## Series: sdxts   
## ARIMA(2,0,0) with non-zero mean   
##   
## Coefficients:  
## ar1 ar2 intercept  
## 0.2973 0.0603 -1879.7917  
## s.e. 0.0988 0.0989 447.6106  
##   
## sigma^2 estimated as 9125576: log likelihood=-998.35  
## AIC=2004.7 AICc=2005.1 BIC=2015.36

auto.arima(sdxts,stepwise=FALSE,approx=FALSE)

## Series: sdxts   
## ARIMA(1,0,2) with non-zero mean   
##   
## Coefficients:  
## ar1 ma1 ma2 intercept  
## -0.6496 0.9994 0.4017 -1877.6381  
## s.e. 0.1933 0.1885 0.0991 410.8637  
##   
## sigma^2 estimated as 8832707: log likelihood=-996.2  
## AIC=2002.4 AICc=2003 BIC=2015.72

#Differnced data(s=17)  
auto.arima(sd17xts)

## Series: sd17xts   
## ARIMA(3,0,1) with zero mean   
##   
## Coefficients:  
## ar1 ar2 ar3 ma1  
## 0.1433 0.5289 -0.0645 0.3899  
## s.e. 0.2675 0.1298 0.1214 0.2577  
##   
## sigma^2 estimated as 13589413: log likelihood=-1356.39  
## AIC=2722.79 AICc=2723.23 BIC=2737.53

auto.arima(sd17xts,stepwise=FALSE,approx=FALSE)

## Series: sd17xts   
## ARIMA(2,0,1) with zero mean   
##   
## Coefficients:  
## ar1 ar2 ma1  
## 0.0320 0.5657 0.4839  
## s.e. 0.1414 0.0959 0.1570  
##   
## sigma^2 estimated as 13517998: log likelihood=-1356.53  
## AIC=2721.06 AICc=2721.35 BIC=2732.85

m0=Arima(xts, order=c(3,1,2),seasonal=list(order=c(1,0,0),period=52), method="ML")  
m1=Arima(xts, order=c(2,1,2),seasonal=list(order=c(1,0,0),period=52), method="ML")

m00=Arima(sdxts, order=c(2,0,0), method="ML")  
m11=Arima(sdxts, order=c(1,0,2), method="ML")

# m11 can be translated to m12 for the original data

m12=Arima(xts, order=c(1,0,2),seasonal=list(order=c(0,1,0),period=52), include.constant=TRUE, method="ML")

m000=Arima(sd17xts, order=c(3,0,1), method="ML")  
m111=Arima(sd17xts, order=c(2,0,1), method="ML")

m0

## Series: xts   
## ARIMA(3,1,2)(1,0,0)[52]   
##   
## Coefficients:  
## ar1 ar2 ar3 ma1 ma2 sar1  
## -0.0969 0.5034 -0.0383 -0.3780 -0.5682 0.2296  
## s.e. 0.1681 0.0976 0.0953 0.1544 0.1506 0.1106  
##   
## sigma^2 estimated as 6614425: log likelihood=-1454.66  
## AIC=2923.32 AICc=2924.08 BIC=2944.72

m1

## Series: xts   
## ARIMA(2,1,2)(1,0,0)[52]   
##   
## Coefficients:  
## ar1 ar2 ma1 ma2 sar1  
## -0.1451 0.5229 -0.3429 -0.6045 0.2287  
## s.e. 0.1418 0.0882 0.1423 0.1347 0.1123  
##   
## sigma^2 estimated as 6579504: log likelihood=-1454.74  
## AIC=2921.47 AICc=2922.03 BIC=2939.81

m00

## Series: sdxts   
## ARIMA(2,0,0) with non-zero mean   
##   
## Coefficients:  
## ar1 ar2 intercept  
## 0.2973 0.0603 -1879.433  
## s.e. 0.0988 0.0989 447.624  
##   
## sigma^2 estimated as 9125576: log likelihood=-998.35  
## AIC=2004.7 AICc=2005.1 BIC=2015.36

m11

## Series: sdxts   
## ARIMA(1,0,2) with non-zero mean   
##   
## Coefficients:  
## ar1 ma1 ma2 intercept  
## -0.6496 0.9994 0.4017 -1877.8586  
## s.e. 0.1933 0.1885 0.0991 410.8673  
##   
## sigma^2 estimated as 8832706: log likelihood=-996.2  
## AIC=2002.4 AICc=2003 BIC=2015.72

m12

## Series: xts   
## ARIMA(1,0,2)(0,1,0)[52] with drift   
##   
## Coefficients:  
## ar1 ma1 ma2 drift  
## -0.6496 0.9994 0.4017 -36.1084  
## s.e. 0.1933 0.1885 0.0991 7.9012  
##   
## sigma^2 estimated as 8833034: log likelihood=-996.2  
## AIC=2002.4 AICc=2003 BIC=2015.72

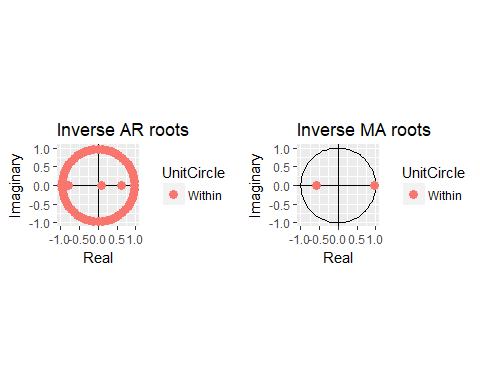
m000

## Series: sd17xts   
## ARIMA(3,0,1) with non-zero mean   
##   
## Coefficients:  
## ar1 ar2 ar3 ma1 intercept  
## 0.1466 0.5249 -0.0675 0.3851 -462.450  
## s.e. 0.2687 0.1303 0.1210 0.2593 1049.511  
##   
## sigma^2 estimated as 13671345: log likelihood=-1356.3  
## AIC=2724.6 AICc=2725.22 BIC=2742.29

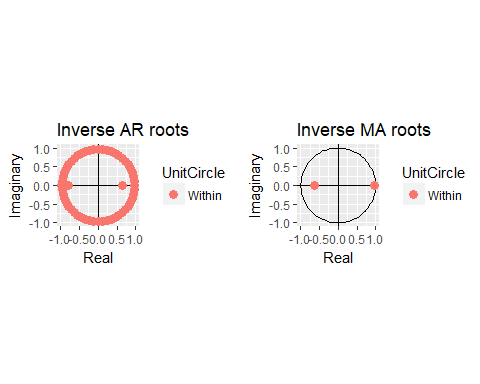
m111

## Series: sd17xts   
## ARIMA(2,0,1) with non-zero mean   
##   
## Coefficients:  
## ar1 ar2 ma1 intercept  
## 0.0298 0.5635 0.4837 -460.0259  
## s.e. 0.1419 0.0959 0.1573 1092.7025  
##   
## sigma^2 estimated as 13600230: log likelihood=-1356.44  
## AIC=2722.88 AICc=2723.32 BIC=2737.62

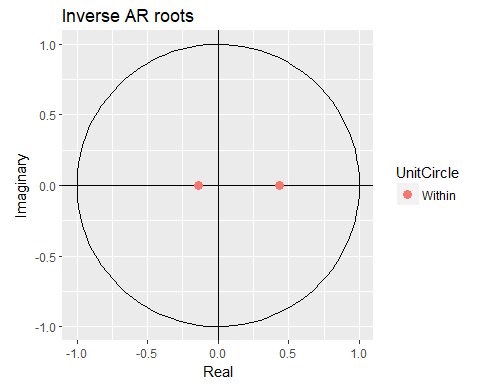
autoplot(m0)



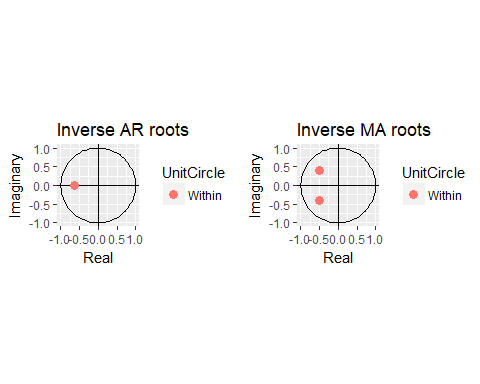
autoplot(m1)



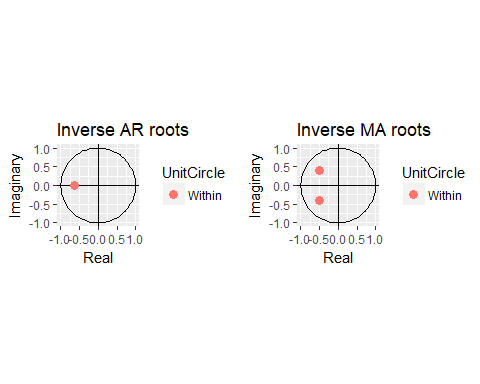
autoplot(m00)



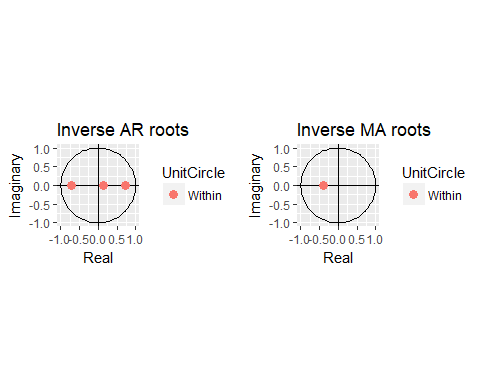
autoplot(m11)



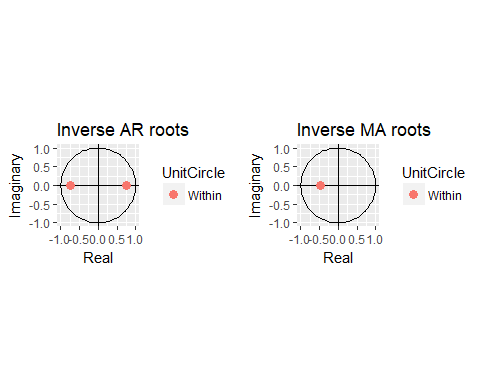
autoplot(m12)



autoplot(m000)



autoplot(m111)



# Coefficient tests

coeftest(m0)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.096882 0.168069 -0.5764 0.5643166   
## ar2 0.503396 0.097576 5.1590 2.483e-07 \*\*\*  
## ar3 -0.038338 0.095266 -0.4024 0.6873710   
## ma1 -0.377974 0.154409 -2.4479 0.0143697 \*   
## ma2 -0.568167 0.150600 -3.7727 0.0001615 \*\*\*  
## sar1 0.229607 0.110587 2.0762 0.0378711 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(m1)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.145111 0.141791 -1.0234 0.30611   
## ar2 0.522904 0.088216 5.9276 3.074e-09 \*\*\*  
## ma1 -0.342894 0.142275 -2.4101 0.01595 \*   
## ma2 -0.604491 0.134743 -4.4863 7.248e-06 \*\*\*  
## sar1 0.228663 0.112328 2.0357 0.04178 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(m00)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 2.9728e-01 9.8825e-02 3.0081 0.002628 \*\*   
## ar2 6.0310e-02 9.8937e-02 0.6096 0.542138   
## intercept -1.8794e+03 4.4762e+02 -4.1987 2.685e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(m11)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -6.4960e-01 1.9331e-01 -3.3604 0.0007782 \*\*\*  
## ma1 9.9943e-01 1.8851e-01 5.3017 1.147e-07 \*\*\*  
## ma2 4.0170e-01 9.9143e-02 4.0517 5.085e-05 \*\*\*  
## intercept -1.8779e+03 4.1087e+02 -4.5705 4.866e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(m12)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.649604 0.193325 -3.3602 0.0007789 \*\*\*  
## ma1 0.999422 0.188526 5.3012 1.150e-07 \*\*\*  
## ma2 0.401690 0.099145 4.0516 5.088e-05 \*\*\*  
## drift -36.108429 7.901227 -4.5700 4.878e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(m000)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 0.146600 0.268701 0.5456 0.5854   
## ar2 0.524904 0.130312 4.0280 5.624e-05 \*\*\*  
## ar3 -0.067466 0.121033 -0.5574 0.5772   
## ma1 0.385086 0.259327 1.4849 0.1376   
## intercept -462.450031 1049.510800 -0.4406 0.6595   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(m111)

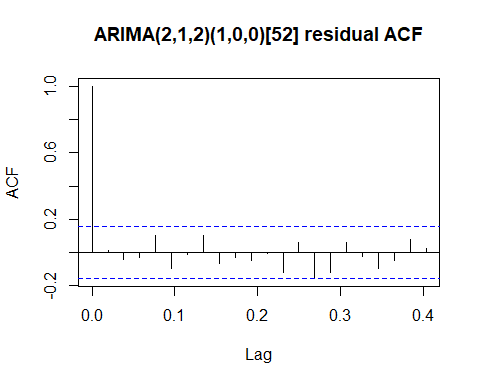
##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 0.029761 0.141869 0.2098 0.833842   
## ar2 0.563459 0.095853 5.8784 4.143e-09 \*\*\*  
## ma1 0.483734 0.157266 3.0759 0.002099 \*\*   
## intercept -460.025917 1092.702550 -0.4210 0.673756   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# M1 and (M11) M12 Pass Coef Tests

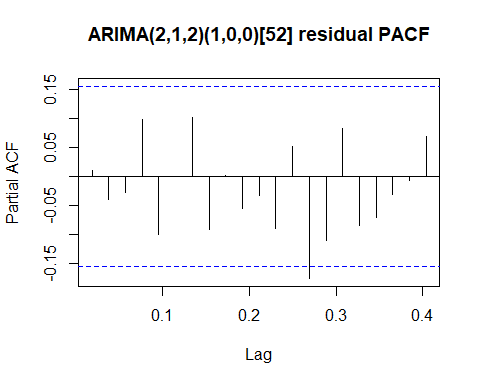
# Residual Analysis

## M1

acf(m1$resid, main="ARIMA(2,1,2)(1,0,0)[52] residual ACF")



pacf(m1$resid, main="ARIMA(2,1,2)(1,0,0)[52] residual PACF")



# ljung box test on residuals  
Box.test(m1$residuals, 26, "Ljung-Box", fitdf=5)

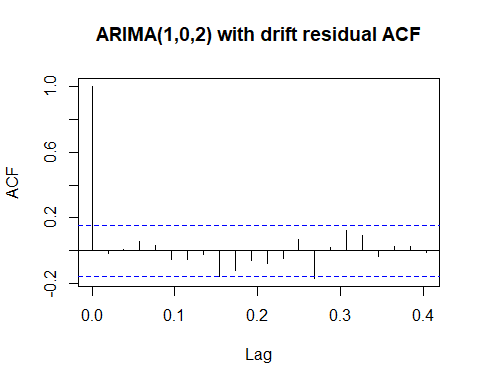
##   
## Box-Ljung test  
##   
## data: m1$residuals  
## X-squared = 21.114, df = 21, p-value = 0.452

Box.test(m1$residuals, 52, "Ljung-Box", fitdf=5)

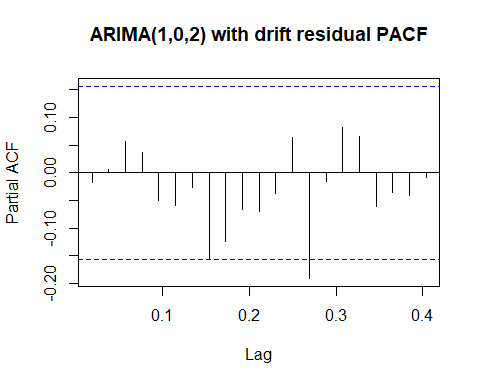
##   
## Box-Ljung test  
##   
## data: m1$residuals  
## X-squared = 46.362, df = 47, p-value = 0.4989

## M12

acf(m12$resid, main="ARIMA(1,0,2) with drift residual ACF")



pacf(m12$resid, main="ARIMA(1,0,2) with drift residual PACF")



# ljung box test on residuals  
Box.test(m12$residuals, 26, "Ljung-Box", fitdf=4)

##   
## Box-Ljung test  
##   
## data: m12$residuals  
## X-squared = 23.533, df = 22, p-value = 0.3722

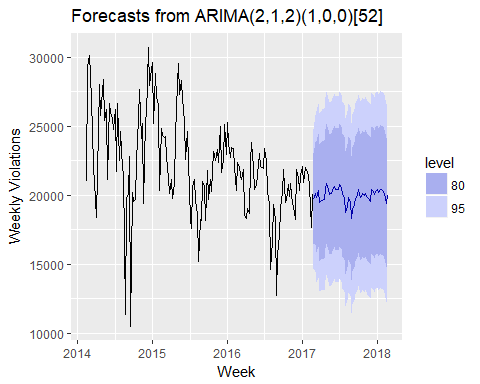
Box.test(m12$residuals, 52, "Ljung-Box", fitdf=4)

##   
## Box-Ljung test  
##   
## data: m12$residuals  
## X-squared = 61.853, df = 48, p-value = 0.08635

# FORECASTING

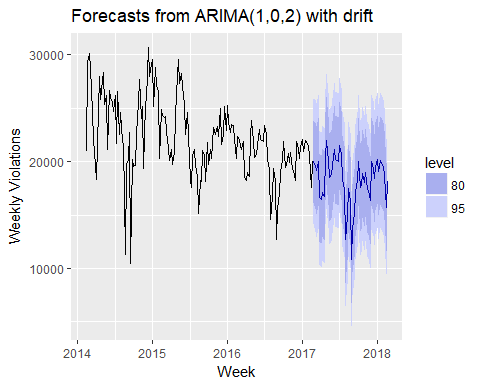
## M1

f1=forecast(m1, h=52)  
  
autoplot(f1)+ xlab("Week")+ylab("Weekly Violations")+ggtitle("Forecasts from ARIMA(2,1,2)(1,0,0)[52]")



# M12

f12=forecast(m12, h=52)  
  
autoplot(f12)+ xlab("Week")+ylab("Weekly Violations")+ggtitle("Forecasts from ARIMA(1,0,2) with drift")



# Backtesting

source("backtest.R")  
  
  
backtest(m1, xts,h=1, orig=length(xts)\*0.8)

## [1] "RMSE of out-of-sample forecasts"  
## [1] 2064.699  
## [1] "Mean absolute error of out-of-sample forecasts"  
## [1] 1611.693  
## [1] "Mean Absolute Percentage error"  
## [1] 0.09048906  
## [1] "Symmetric Mean Absolute Percentage error"  
## [1] 0.08853523

backtest(m12, xts,h=1, orig=length(xts)\*0.8)

## [1] "RMSE of out-of-sample forecasts"  
## [1] 2484.861  
## [1] "Mean absolute error of out-of-sample forecasts"  
## [1] 2007.186  
## [1] "Mean Absolute Percentage error"  
## [1] 0.1059659  
## [1] "Symmetric Mean Absolute Percentage error"  
## [1] 0.1062556

# MONTHLY DATA

mcv=read.table("MonthlyViolations.csv",header=T, sep=',')   
head(mcv)

## Month Violations  
## 1 7/1/2014 127226  
## 2 8/1/2014 92873  
## 3 9/1/2014 117935  
## 4 10/1/2014 110497  
## 5 11/1/2014 104819  
## 6 12/1/2014 106074

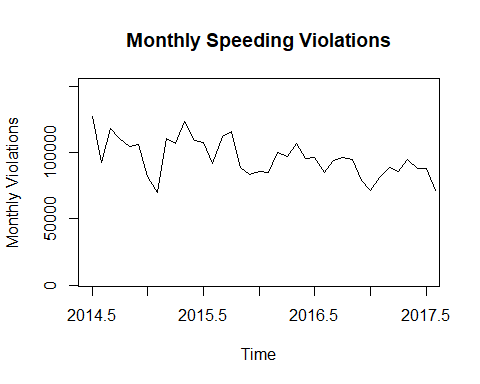
x=mcv$Violations  
head(x)

## [1] 127226 92873 117935 110497 104819 106074

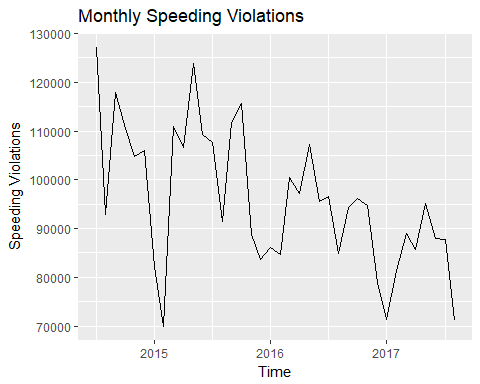
xts=ts(x,frequency=12,start=c(2014,7))

# Create time plot

plot(xts, main="Monthly Speeding Violations", xlim=c(2014.5,2017.5), ylim=c(5000,150000), ylab="Monthly Violations")

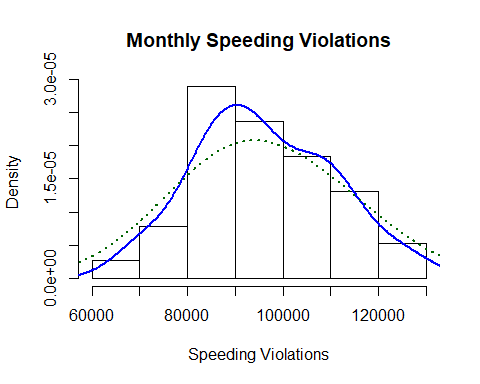


autoplot(xts)+ylab("Speeding Violations")+ggtitle("Monthly Speeding Violations")



# Check Normality

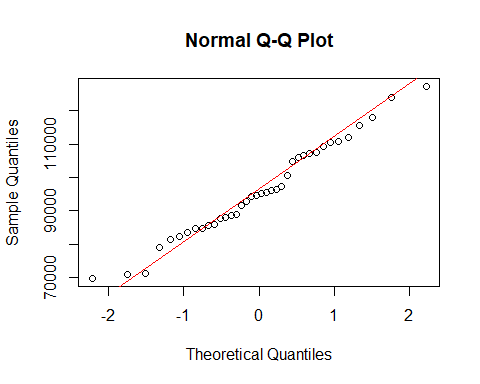
hist(x,main="Monthly Speeding Violations" , probability = TRUE, xlab = "Speeding Violations")  
lines(density(x), col="blue", lwd=2)  
lines(density(x, adjust=2), lty="dotted", col="darkgreen", lwd=2)



normalTest(x,method=c("jb"))

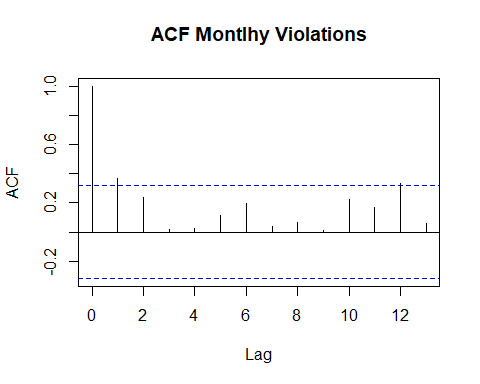
##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 0.7347  
## P VALUE:  
## Asymptotic p Value: 0.6926   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

qqnorm(x)  
qqline(x, col = 2)

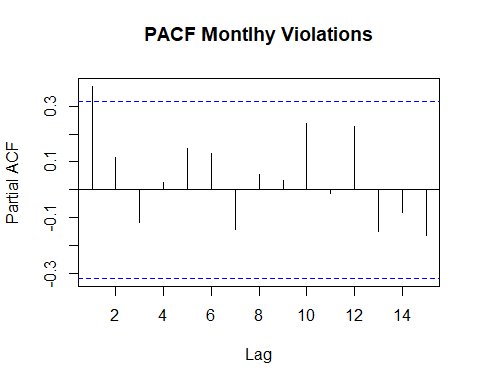


# create acf plots

acf(x,lag.max=13, main="ACF Montlhy Violations")



pacf(x, main="PACF Montlhy Violations")



# Dickey Fuller Tests

# tests for AR model with time trend  
adfTest(xts, lags=6, type=c("ct"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 6  
## STATISTIC:  
## Dickey-Fuller: -3.6189  
## P VALUE:  
## 0.04487   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

adfTest(xts, lags=12, type=c("ct"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 12  
## STATISTIC:  
## Dickey-Fuller: -2.9264  
## P VALUE:  
## 0.2113   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

# tests for AR model with no time trend  
adfTest(x, lags=6, type=c("c"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 6  
## STATISTIC:  
## Dickey-Fuller: -1.0325  
## P VALUE:  
## 0.6676   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

adfTest(x, lags=12, type=c("c"))

## Warning in adfTest(x, lags = 12, type = c("c")): p-value greater than  
## printed p-value

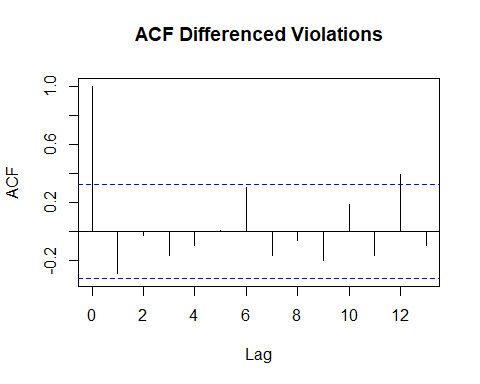
##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 12  
## STATISTIC:  
## Dickey-Fuller: 0.8308  
## P VALUE:  
## 0.99   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

# Differencing

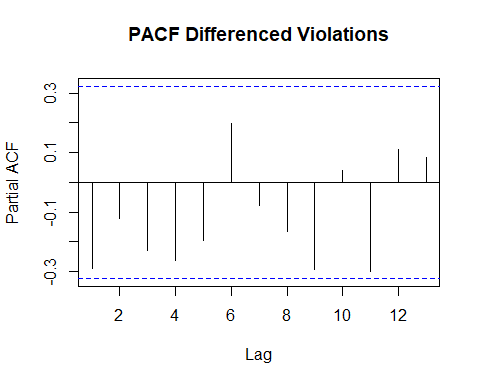
dx=diff(x)

# Create acf plots

acf(as.vector(dx),lag.max=13, main="ACF Differenced Violations")

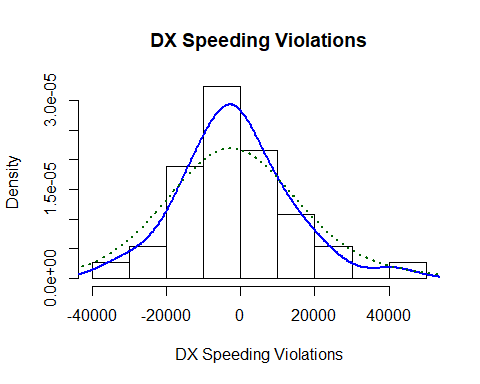


pacf(as.vector(dx),lag.max=13, main="PACF Differenced Violations")



# Check for Normality

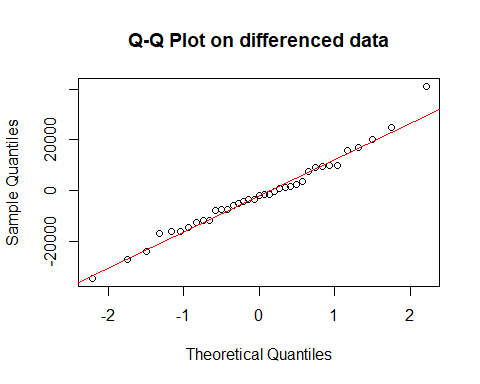
hist(dx,main="DX Speeding Violations" , probability = TRUE, xlab = "DX Speeding Violations")  
lines(density(dx), col="blue", lwd=2)  
lines(density(dx, adjust=2), lty="dotted", col="darkgreen", lwd=2)



normalTest(dx,method=c("jb"))

##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 2.1225  
## P VALUE:  
## Asymptotic p Value: 0.346   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

qqnorm(dx, main="Q-Q Plot on differenced data")  
qqline(dx, col = 2)

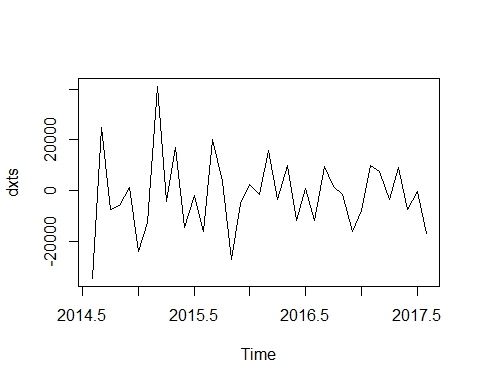


basicStats(dx)

## dx  
## nobs 3.700000e+01  
## NAs 0.000000e+00  
## Minimum -3.435300e+04  
## Maximum 4.104600e+04  
## 1. Quartile -1.159300e+04  
## 3. Quartile 7.581000e+03  
## Mean -1.517865e+03  
## Median -1.771000e+03  
## Sum -5.616100e+04  
## SE Mean 2.420203e+03  
## LCL Mean -6.426265e+03  
## UCL Mean 3.390535e+03  
## Variance 2.167232e+08  
## Stdev 1.472152e+04  
## Skewness 3.912090e-01  
## Kurtosis 6.389030e-01

# Unit-root tests on first difference

dxts=diff(xts)  
plot(dxts)



adfTest(coredata(dxts), lags=2, type=c("c"))

## Warning in adfTest(coredata(dxts), lags = 2, type = c("c")): p-value  
## smaller than printed p-value

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -4.7  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

adfTest(dxts, lags=2, type=c("c"))

## Warning in adfTest(dxts, lags = 2, type = c("c")): p-value smaller than  
## printed p-value

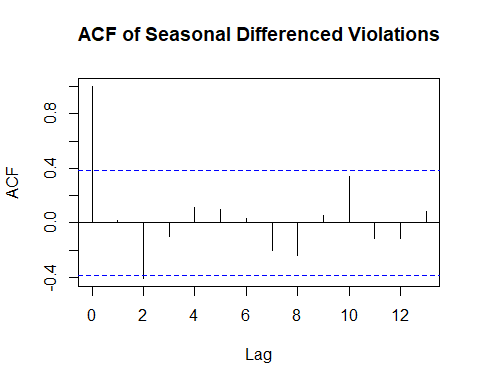
##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -4.7  
## P VALUE:  
## 0.01   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

# Seasonal Differencing

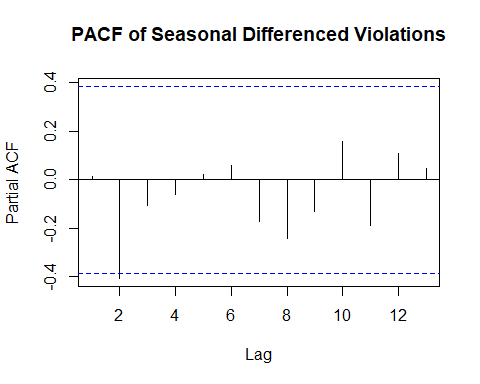
sdx=diff(x,12)

# create acf plotS

acf(as.vector(sdx),lag.max=13, main="ACF of Seasonal Differenced Violations")



pacf(as.vector(sdx),lag.max=13, main="PACF of Seasonal Differenced Violations")



# Ljung Box Tests

Box.test(sdx,lag = 1, type="Ljung-Box", fitdf=1)

##   
## Box-Ljung test  
##   
## data: sdx  
## X-squared = 0.0062061, df = 0, p-value < 2.2e-16

Box.test(sdx,lag = 3, type="Ljung-Box", fitdf=2)

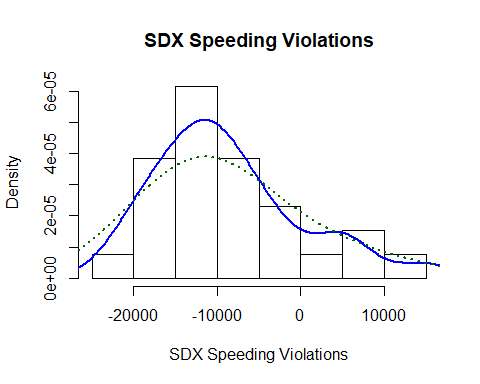
##   
## Box-Ljung test  
##   
## data: sdx  
## X-squared = 5.3449, df = 1, p-value = 0.02078

Box.test(sdx,lag = 12, type="Ljung-Box", fitdf=2)

##   
## Box-Ljung test  
##   
## data: sdx  
## X-squared = 16.562, df = 10, p-value = 0.08464

# Check Normality

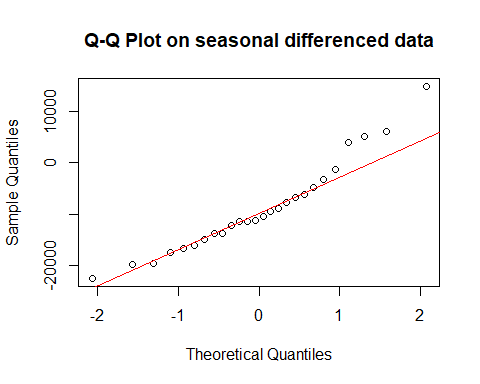
hist(sdx,main="SDX Speeding Violations" , probability = TRUE, xlab = "SDX Speeding Violations")  
lines(density(sdx), col="blue", lwd=2)  
lines(density(sdx, adjust=2), lty="dotted", col="darkgreen", lwd=2)



normalTest(sdx,method=c("jb"))

##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 3.3123  
## P VALUE:  
## Asymptotic p Value: 0.1909   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

qqnorm(sdx, main="Q-Q Plot on seasonal differenced data")  
qqline(sdx, col = 2)

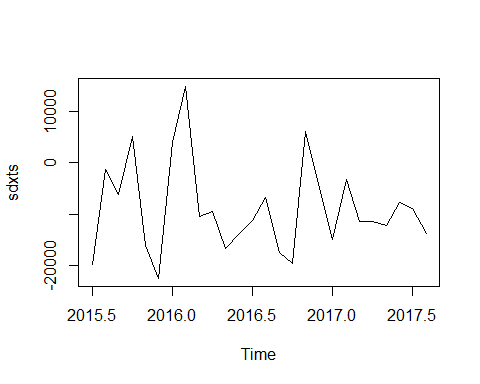


basicStats(sdx)

## sdx  
## nobs 2.600000e+01  
## NAs 0.000000e+00  
## Minimum -2.240200e+04  
## Maximum 1.478300e+04  
## 1. Quartile -1.455200e+04  
## 3. Quartile -5.048750e+03  
## Mean -8.797923e+03  
## Median -1.076650e+04  
## Sum -2.287460e+05  
## SE Mean 1.742146e+03  
## LCL Mean -1.238594e+04  
## UCL Mean -5.209906e+03  
## Variance 7.891188e+07  
## Stdev 8.883236e+03  
## Skewness 8.075100e-01  
## Kurtosis 9.861500e-02

# Unit-root tests on seasonal difference

sdxts=diff(xts,12)  
plot(sdxts)



adfTest(coredata(sdxts), lags=2, type=c("c"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -3.5071  
## P VALUE:  
## 0.01867   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

adfTest(sdxts, lags=2, type=c("c"))

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -3.5071  
## P VALUE:  
## 0.01867   
##   
## Description:  
## Sat Mar 10 13:48:53 2018 by user: guy.dor

# try automated order selection

auto.arima(xts)

## Series: xts   
## ARIMA(0,1,1)(1,0,0)[12]   
##   
## Coefficients:  
## ma1 sar1  
## -0.8527 0.7563  
## s.e. 0.0744 0.1048  
##   
## sigma^2 estimated as 84694254: log likelihood=-394.8  
## AIC=795.61 AICc=796.34 BIC=800.44

auto.arima(xts,stepwise = FALSE, approx=FALSE)

## Series: xts   
## ARIMA(0,1,1)(1,0,0)[12]   
##   
## Coefficients:  
## ma1 sar1  
## -0.8527 0.7563  
## s.e. 0.0744 0.1048  
##   
## sigma^2 estimated as 84694254: log likelihood=-394.8  
## AIC=795.61 AICc=796.34 BIC=800.44

# Force Seasonal Differencing to compare

auto.arima(xts,stepwise = FALSE,D=1, approx=FALSE)

## Series: xts   
## ARIMA(0,0,0)(0,1,0)[12] with drift   
##   
## Coefficients:  
## drift  
## -733.1603  
## s.e. 142.3598  
##   
## sigma^2 estimated as 78917804: log likelihood=-272.77  
## AIC=549.54 AICc=550.07 BIC=552.06

m3=Arima(xts, order=c(0,1,1),seasonal=list(order=c(1,0,0),period=12), method="ML")

# Force Seasonal Differencing to compare

m4=Arima(xts, order=c(0,0,0),seasonal=list(order=c(0,1,0),period=12), method="ML")

# Coefficient test

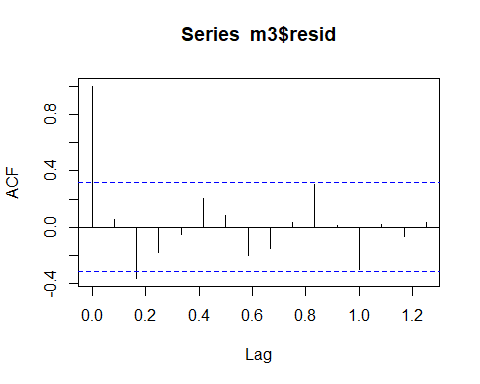
coeftest(m3)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -0.852732 0.074381 -11.4644 < 2.2e-16 \*\*\*  
## sar1 0.756302 0.104793 7.2171 5.311e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

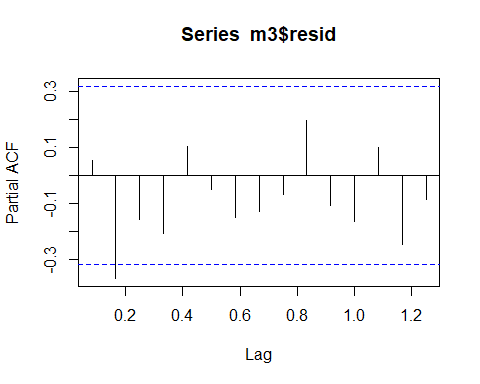
Cannot evaluate m4 coefficients, is a white noise process

# Residual Analysis

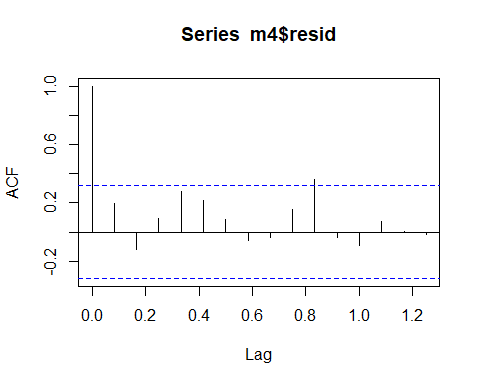
acf(m3$resid)



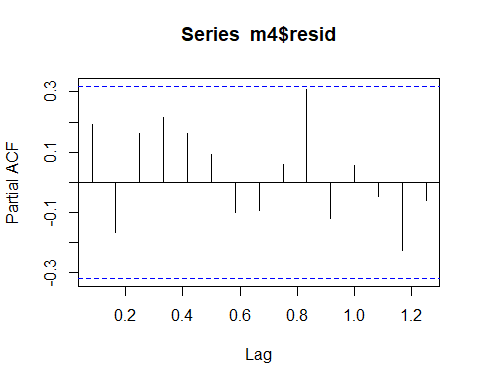
pacf(m3$resid)



acf(m4$resid)



pacf(m4$resid)

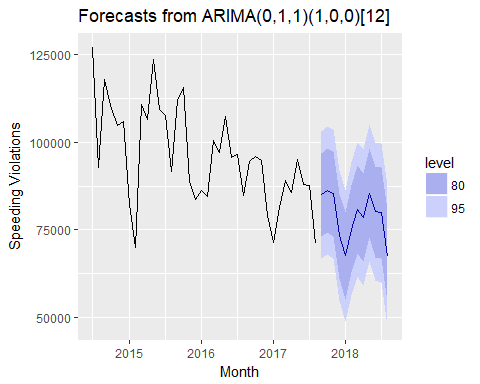


# Forecast

f3=forecast(m3, h=12)  
f3$mean

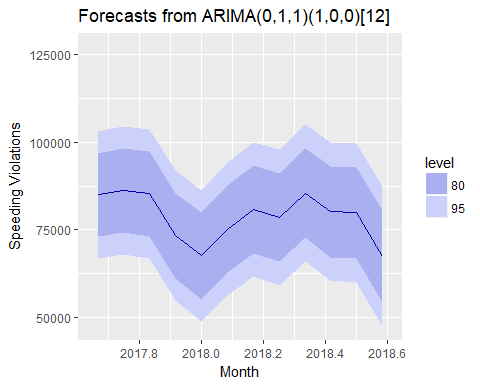
## Jan Feb Mar Apr May Jun Jul  
## 2017   
## 2018 67493.54 75153.37 80886.89 78439.50 85492.02 80085.97 79868.91  
## Aug Sep Oct Nov Dec  
## 2017 84948.99 86231.68 85234.87 73298.16  
## 2018 67319.59

autoplot(f3)+ xlab("Month")+ylab("Speeding Violations")+ggtitle("Forecasts from ARIMA(0,1,1)(1,0,0)[12]")



autoplot(f3, xlim=c(2017.65,2018.6))+ xlab("Month")+ylab("Speeding Violations")+ggtitle("Forecasts from ARIMA(0,1,1)(1,0,0)[12]")

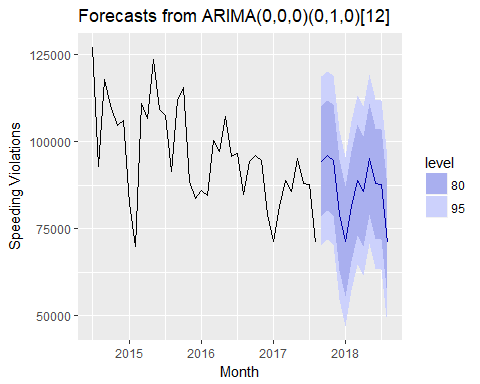
## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.



f4=forecast(m4, h=12)  
f4$mean

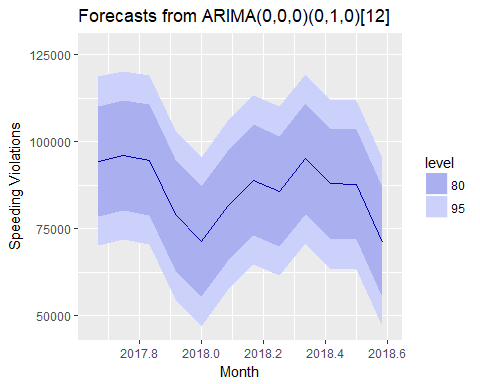
## Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov  
## 2017 94375 96071 94753  
## 2018 71295 81423 89004 85768 95093 87945 87658 71065   
## Dec  
## 2017 78970  
## 2018

autoplot(f4)+ xlab("Month")+ylab("Speeding Violations")+ggtitle("Forecasts from ARIMA(0,0,0)(0,1,0)[12]")



autoplot(f4, xlim=c(2017.65,2018.6))+ xlab("Month")+ylab("Speeding Violations")+ggtitle("Forecasts from ARIMA(0,0,0)(0,1,0)[12]")

## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.



# Backtesting

backtest(m3, xts,h=1, orig=length(xts)\*0.8)

## Warning in predict.Arima(object, n.ahead = h): MA part of model is not  
## invertible  
  
## Warning in predict.Arima(object, n.ahead = h): MA part of model is not  
## invertible

## [1] "RMSE of out-of-sample forecasts"  
## [1] 6076.338  
## [1] "Mean absolute error of out-of-sample forecasts"  
## [1] 3900.595  
## [1] "Mean Absolute Percentage error"  
## [1] 0.04977433  
## [1] "Symmetric Mean Absolute Percentage error"  
## [1] 0.04742143

backtest(m4, xts,h=1, orig=length(xts)\*0.8)

## [1] "RMSE of out-of-sample forecasts"  
## [1] 10531.4  
## [1] "Mean absolute error of out-of-sample forecasts"  
## [1] 9940.429  
## [1] "Mean Absolute Percentage error"  
## [1] 0.1178775  
## [1] "Symmetric Mean Absolute Percentage error"  
## [1] 0.1103653

# ATTEMPTED GARCH

# Violations rate of change

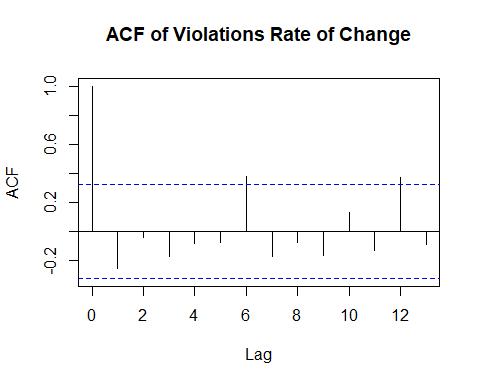
VrOc=diff(x)/x[-length(x)]  
VrOc

## [1] -0.270015563 0.269852379 -0.063068640 -0.051386010 0.011973020  
## [6] -0.224918453 -0.150518147 0.587706362 -0.038146942 0.160420788  
## [11] -0.116291095 -0.016192148 -0.149438213 0.222053473 0.033644475  
## [16] -0.232732746 -0.056717360 0.028982216 -0.017108610 0.187039138  
## [21] -0.032174571 0.102982925 -0.108111385 0.009054883 -0.120594788  
## [26] 0.112034124 0.017970861 -0.013719020 -0.166569924 -0.097188806  
## [31] 0.142057648 0.093106370 -0.036357916 0.108723533 -0.075168519  
## [36] -0.003263403 -0.189292478

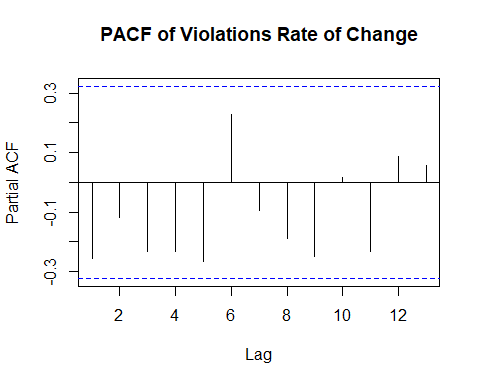
r2=VrOc^2  
absr=abs(VrOc)

# create acf plot

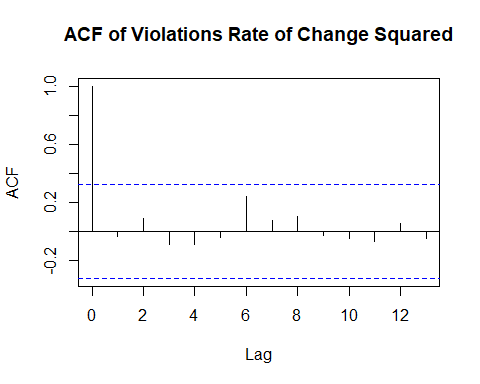
acf(as.vector(VrOc),lag.max=13, main="ACF of Violations Rate of Change")



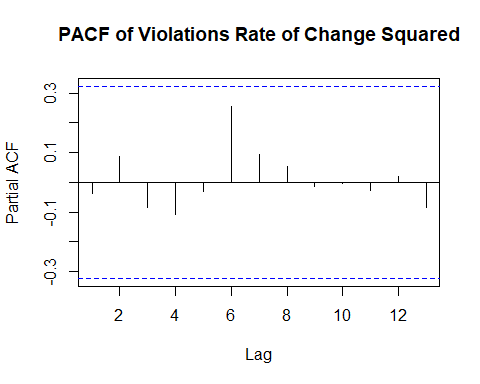
pacf(as.vector(VrOc),lag.max=13, main="PACF of Violations Rate of Change")



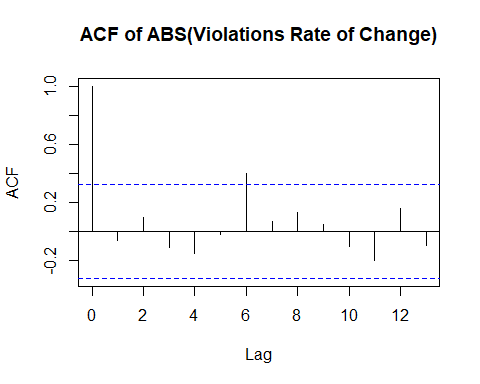
acf(as.vector(r2),lag.max=13, main="ACF of Violations Rate of Change Squared")



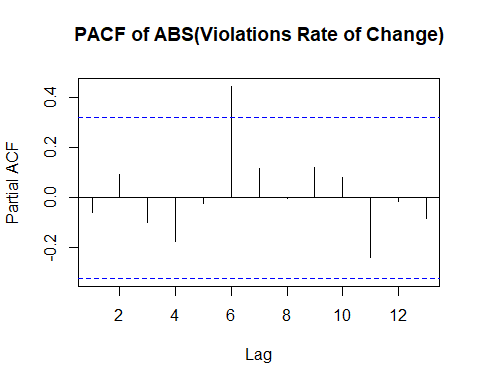
pacf(as.vector(r2),lag.max=13, main="PACF of Violations Rate of Change Squared")



acf(as.vector(absr),lag.max=13, main="ACF of ABS(Violations Rate of Change)")



pacf(as.vector(absr),lag.max=13, main="PACF of ABS(Violations Rate of Change)")



# Coefficient test

Box.test(VrOc,lag = 1, type="Ljung-Box", fitdf=1)

##   
## Box-Ljung test  
##   
## data: VrOc  
## X-squared = 2.6244, df = 0, p-value < 2.2e-16

Box.test(VrOc,lag = 3, type="Ljung-Box", fitdf=2)

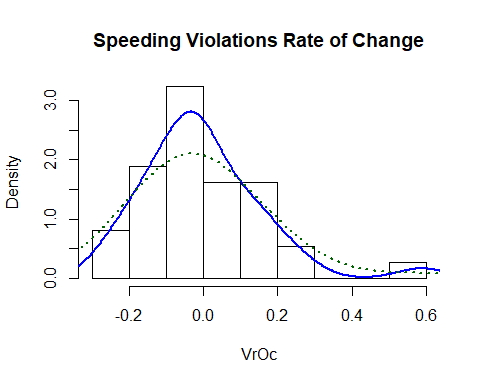
##   
## Box-Ljung test  
##   
## data: VrOc  
## X-squared = 3.9478, df = 1, p-value = 0.04693

Box.test(VrOc,lag = 12, type="Ljung-Box", fitdf=2)

##   
## Box-Ljung test  
##   
## data: VrOc  
## X-squared = 24.115, df = 10, p-value = 0.007302

# Check Normality

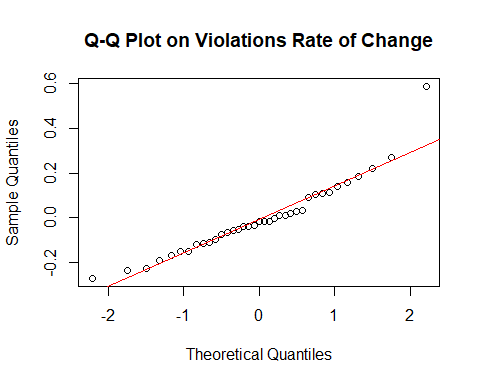
hist(VrOc,main="Speeding Violations Rate of Change" , probability = TRUE, xlab = "VrOc")  
lines(density(VrOc), col="blue", lwd=2)  
lines(density(VrOc, adjust=2), lty="dotted", col="darkgreen", lwd=2)



normalTest(VrOc,method=c("jb"))

##   
## Title:  
## Jarque - Bera Normalality Test  
##   
## Test Results:  
## STATISTIC:  
## X-squared: 24.3314  
## P VALUE:  
## Asymptotic p Value: 5.206e-06   
##   
## Description:  
## Sat Mar 10 13:49:08 2018 by user: guy.dor

qqnorm(VrOc, main="Q-Q Plot on Violations Rate of Change")  
qqline(VrOc, col = 2)



basicStats(VrOc)

## VrOc  
## nobs 37.000000  
## NAs 0.000000  
## Minimum -0.270016  
## Maximum 0.587706  
## 1. Quartile -0.108111  
## 3. Quartile 0.093106  
## Mean -0.003821  
## Median -0.017109  
## Sum -0.141373  
## SE Mean 0.026699  
## LCL Mean -0.057970  
## UCL Mean 0.050328  
## Variance 0.026376  
## Stdev 0.162406  
## Skewness 1.225327  
## Kurtosis 2.721155

# automated selection

auto.arima(VrOc)

## Series: VrOc   
## ARIMA(0,0,2) with zero mean   
##   
## Coefficients:  
## ma1 ma2  
## -0.5898 -0.3064  
## s.e. 0.1589 0.1652  
##   
## sigma^2 estimated as 0.02051: log likelihood=19.69  
## AIC=-33.38 AICc=-32.65 BIC=-28.55

m5=Arima(VrOc, order=c(0,0,2), method="ML")  
m5

## Series: VrOc   
## ARIMA(0,0,2) with non-zero mean   
##   
## Coefficients:  
## ma1 ma2 intercept  
## -0.6701 -0.3299 0.0027  
## s.e. 0.1745 0.1409 0.0026  
##   
## sigma^2 estimated as 0.01968: log likelihood=20.14  
## AIC=-32.28 AICc=-31.03 BIC=-25.84

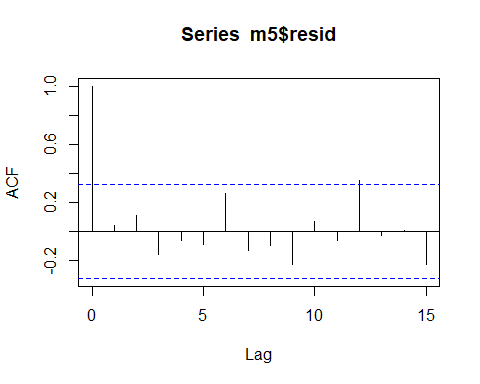
# Coefficient test

coeftest(m5)

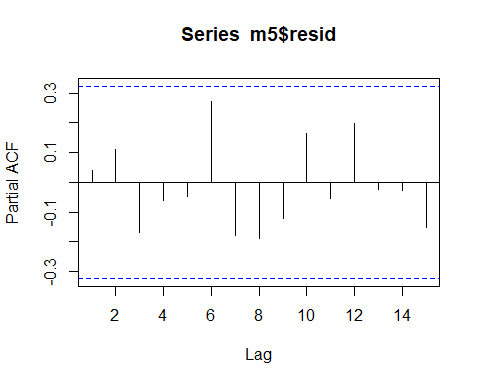
##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -0.6701440 0.1744628 -3.8412 0.0001224 \*\*\*  
## ma2 -0.3298535 0.1409436 -2.3403 0.0192671 \*   
## intercept 0.0026947 0.0026102 1.0324 0.3018940   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Residual Analysis

acf(m5$resid)



pacf(m5$resid)



# Forecast

f5=forecast(m5, h=12)  
f5$mean

## Time Series:  
## Start = 38   
## End = 49   
## Frequency = 1   
## [1] 0.129179287 0.066494190 0.002694711 0.002694711 0.002694711  
## [6] 0.002694711 0.002694711 0.002694711 0.002694711 0.002694711  
## [11] 0.002694711 0.002694711

plot(f5)

