Predicting the Crude Oil Production in USA

Divya Krishnan

December 20, 2018

Contents

1.Introduction

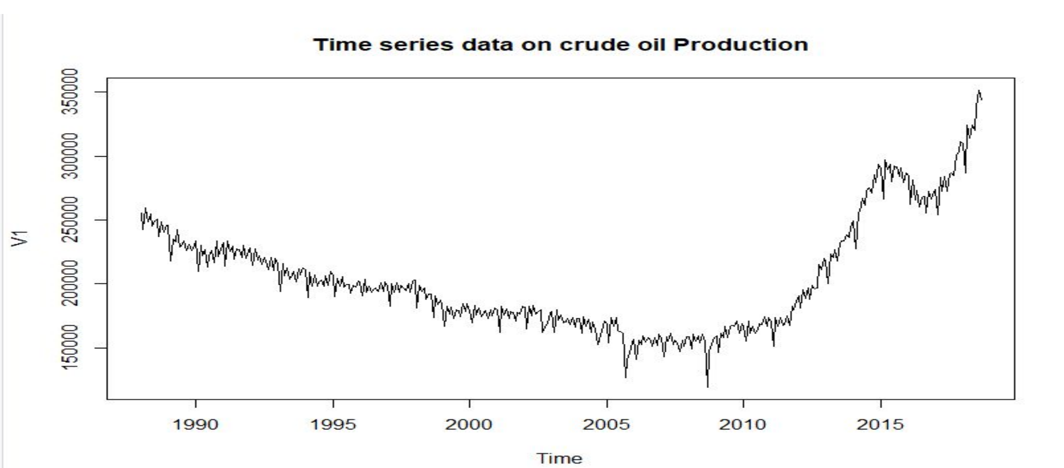
Oil is considered to be the lifeblood of industrially developed countries. Oil has become the world's most important source of energy since the mid-1950s. Its products underpin modern society, mainly supplying energy to power industry, heat homes and provide fuel for vehicles and aeroplanes to carry goods and people all over the world. The world is now pumping and consuming more oil than it ever has, with output from big producers such as the United States and Saudi Arabia at or near record levels. In August, for the first time in history, the world pumped more than 100 million barrels a day, according to a new report from the International Energy Agency (IEA). That was fueled by a continuing gusher from the U.S. shale oil patch—which has apparently turned the United States into the world's largest oil producer, with the country pumping almost 11 million barrels a day—and a rebound from OPEC, which is pumping more than it has all year.

Hence it would be interesting to see and analyze the trends in oil production in USA and forecast the near future oil production

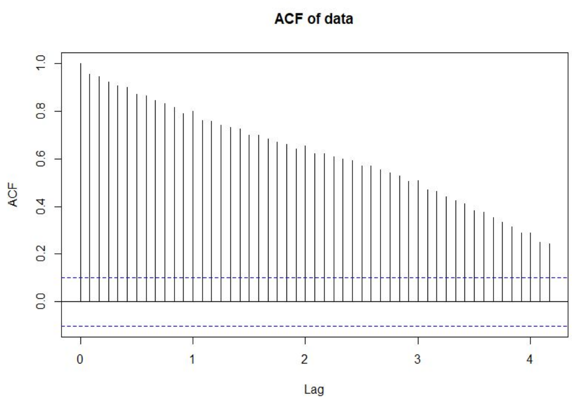
2.Model Building

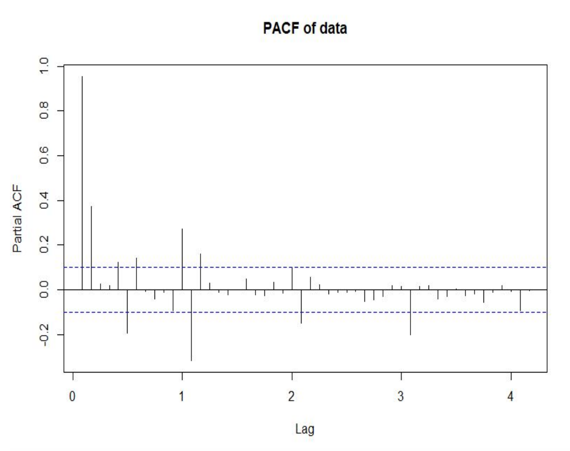
The Dataset gives details about the crude oil production in USA from 1988 to Oct 2018. The data was collected from eia.gov. It is a univariate time series data which entails the monthly production of crude oil till 2018. The oil production is measured in tons.

The trend in the oil production in the mentioned time period is given below.



The ACF and PACF of the data is given as

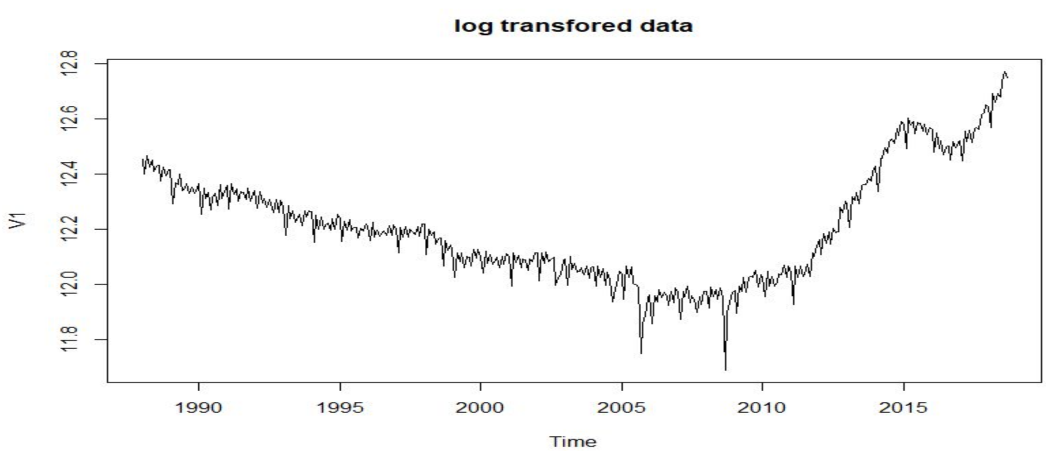




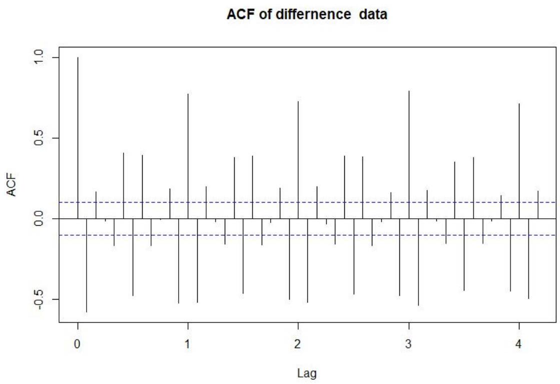
The ACF and PACF values are given as

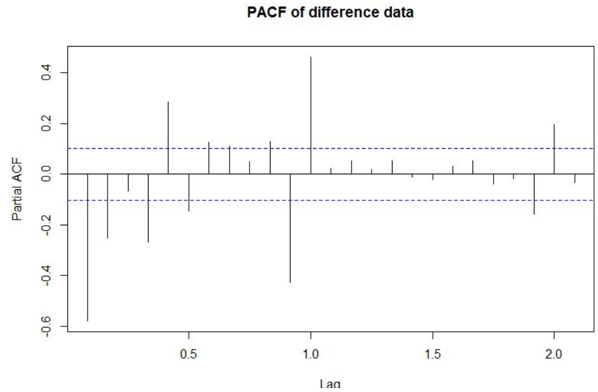
- Seasonal Arima Model

Note that the data is non stationary which is evident from the ACF and PACF plots. We also notice a few outliers in the data and do log transformation mainly to deal with outliers. After log transformation the data is given as below:

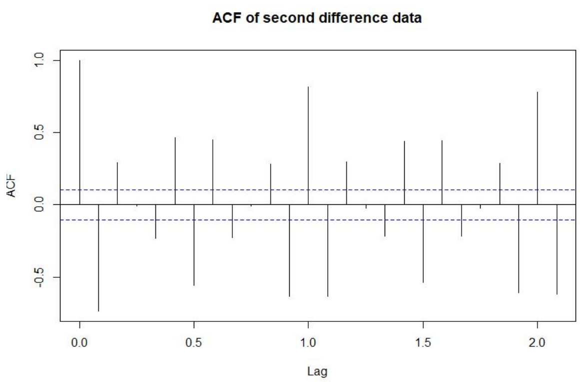


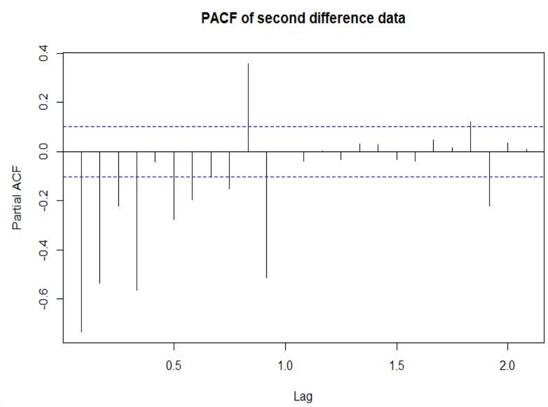
Inorder to make the data stationary we take the difference of the log transformed data. After differencing the ACF and PACF are as follows:



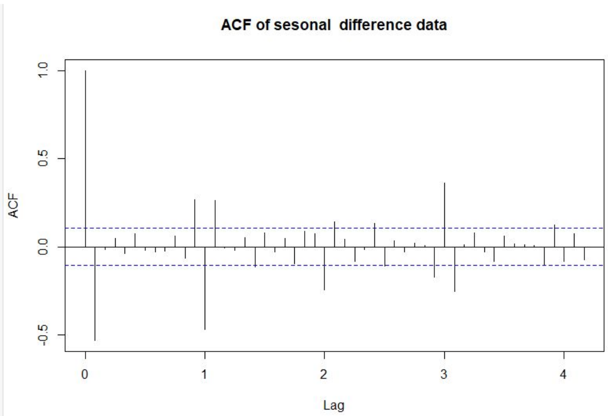


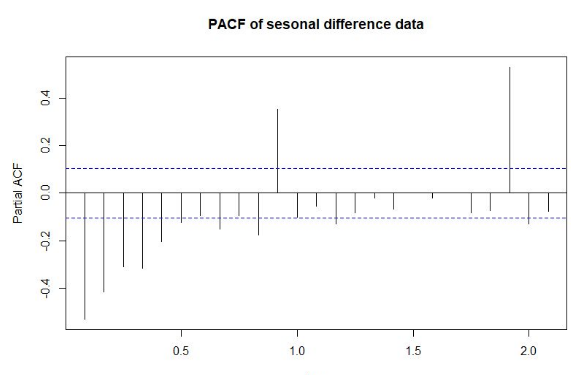
From the plot we can see that the data is still not stationary and hence take the second difference of the data. Below are the ACF and PACF obtained after the second difference.





We note from the ACF that the data has seasonality after every 12 months. Hence taking seasonal difference on the data would be a good idea. The ACF and PACF after taking seasonal difference is given below.





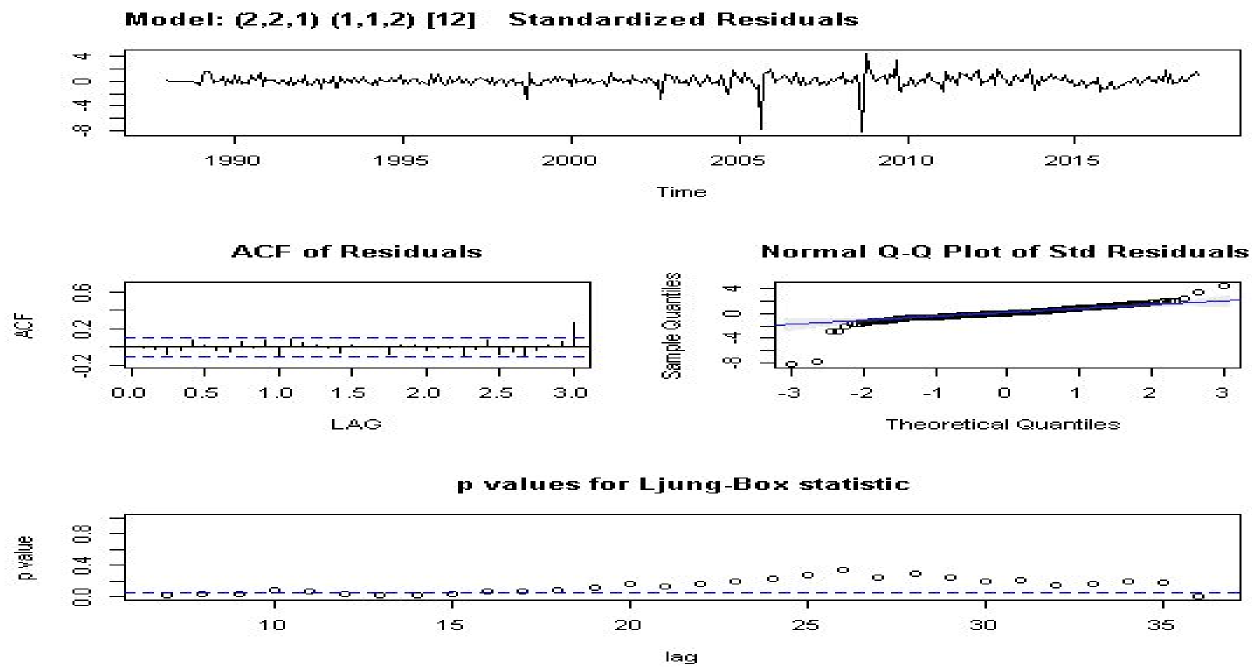
After the seasonal differencing note that the data has finally become stationary and ARIMA modeling can be done. From the ACF and PACF plot, we can say that it's an ARIMA(p,d,q) X (P,D,Q) model. We try different values of p,q and P and Q and compare the BIC of each model.

ARIMA(p,d,q)X(P,D,Q)	BIC Error
(2,2,1)X(0,1,1)	233.5391
(1,2,1)X(0,1,1)	230.0926
(2,2,1)X(1,1,1)	239.4236
(1,2,2)X(0,1,1)	229.0369.
(2,2,1)X(1,1,2)	228.0266

From the BIC error we can see that ARIMA (2,2,1)x(1,1,2) has the least BIC error and we decide that ARIMA(2,1,3) fits the model the best to give us better forecast.
The model is given as :

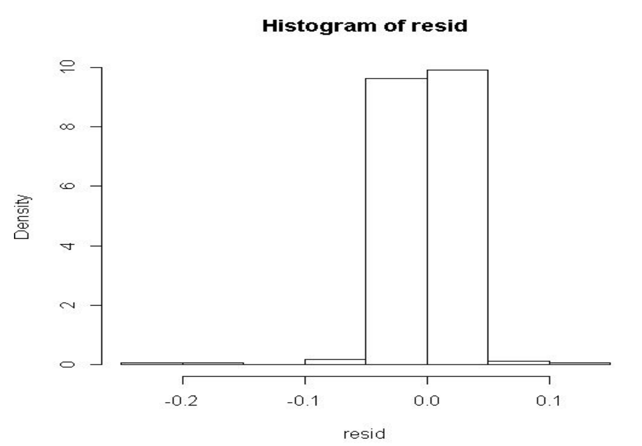
$$(1-\varphi 1B^{12})(1-\phi 1B-\phi 2B^{2})(1-B^{12})(1-B)xt=(1+\Theta 1^{12}+\Theta 2^{24})(1-\theta B)\omega t$$

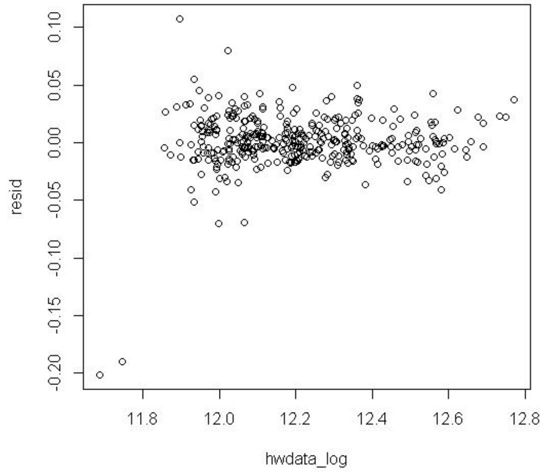
After fitting the model, on diagnostic checking we get the following plot



We note that the p values do not give us a desirable result as a lot of them have value below zero. The tails of the Q-Q plot also shows some deviation indicating presence of outliers.

Let's further explore the pattern shown by the residue by looking at the histogram plot of the residuals

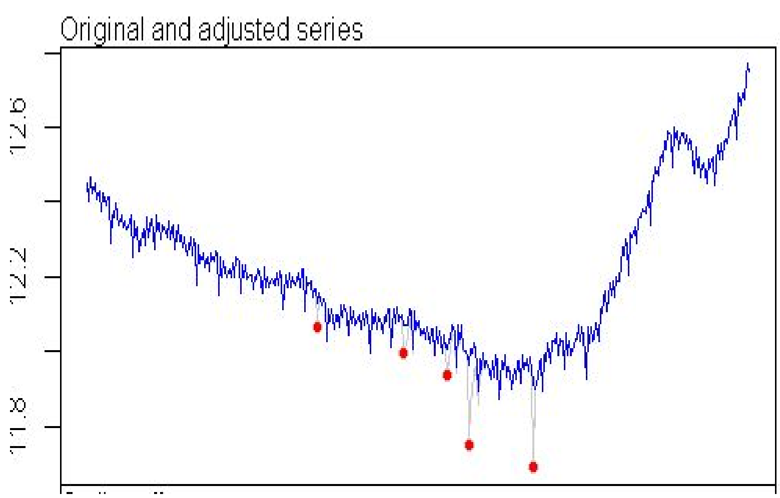




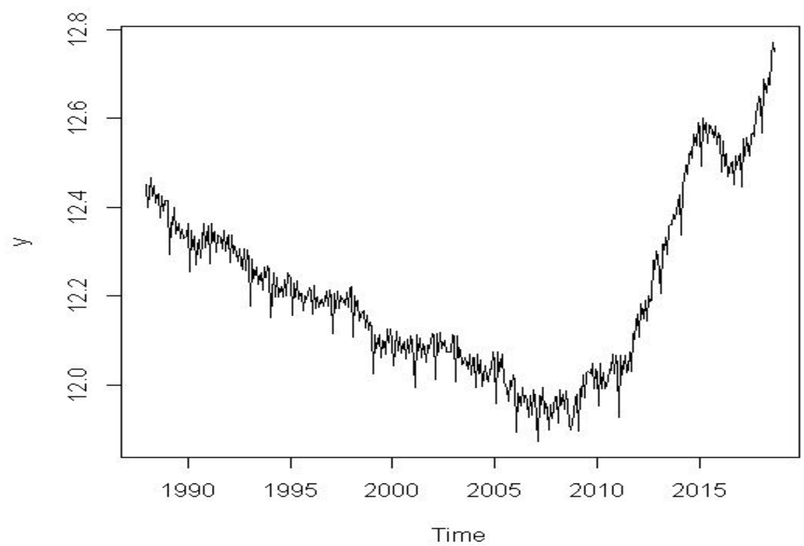
The histogram clearly does not show a gaussian pattern and and the residue vs data plot shows presence of clear outlier. Hence simple ARIMA model won't be a good approach for forecasting the oil production and we must proceed to fix the outliers for better predictions.

- Outlier Detection

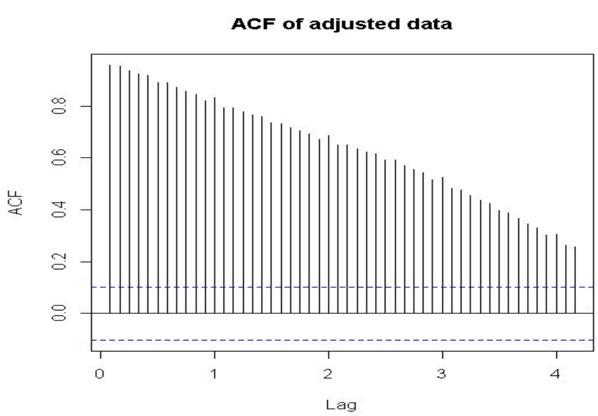
The seasonal ARIMA model does not give us a good diagnostic result hence we decide to do outlier detection and fitting ARIMA model to the data.
The outliers are dealt using TSO package in R. The original and adjusted series is given below

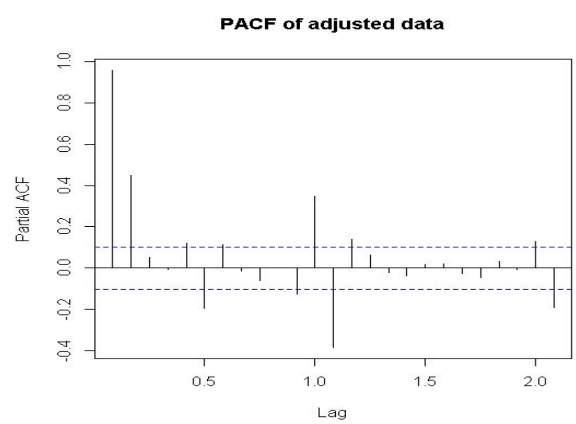


The new adjusted time series data :

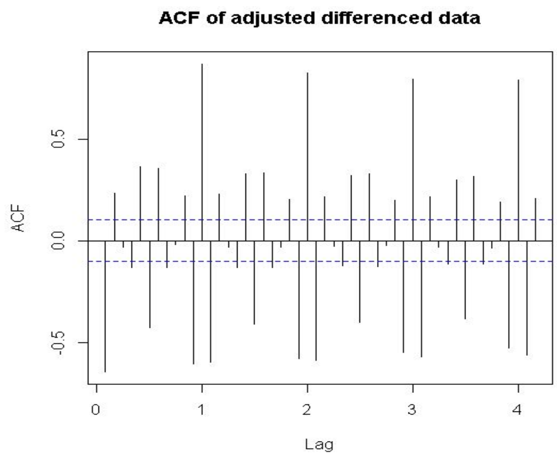


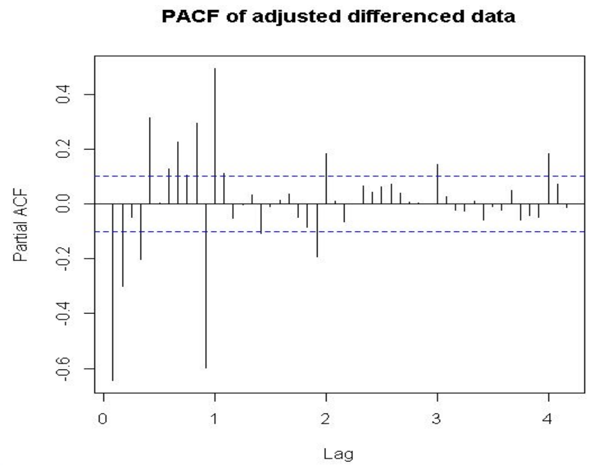
The ACF and PACF of the data is given below :



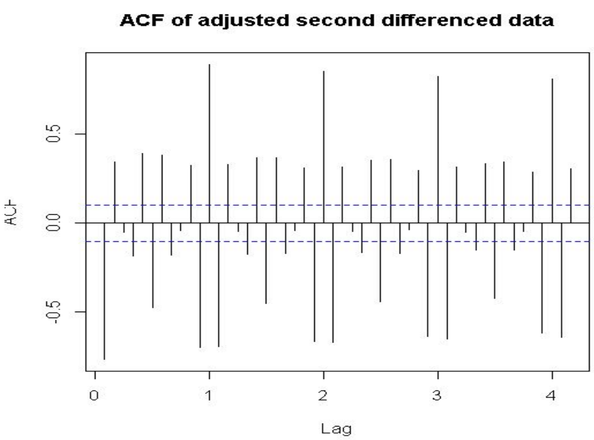


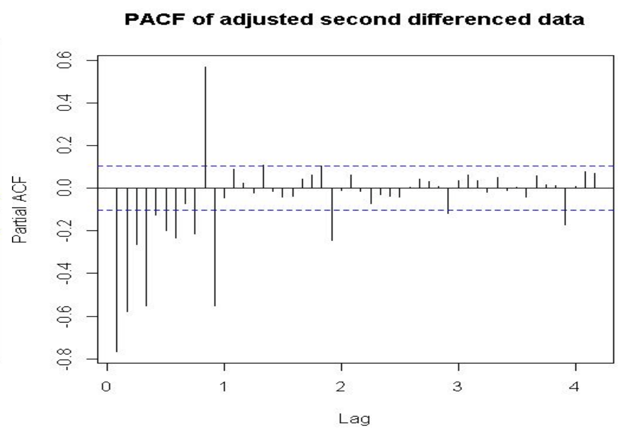
We again take the first and second difference on the data to make the data stationary. The ACF and PACF after taking the difference is given below:



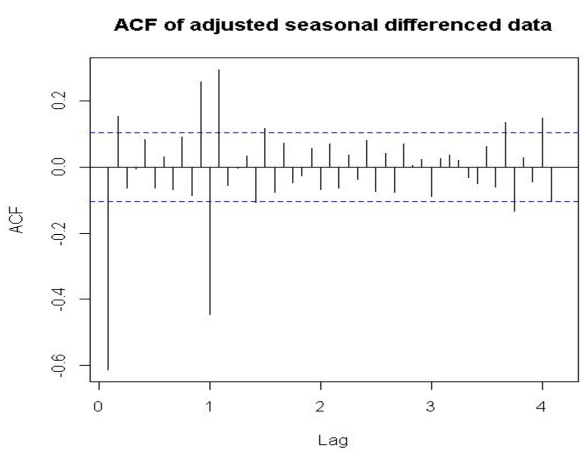


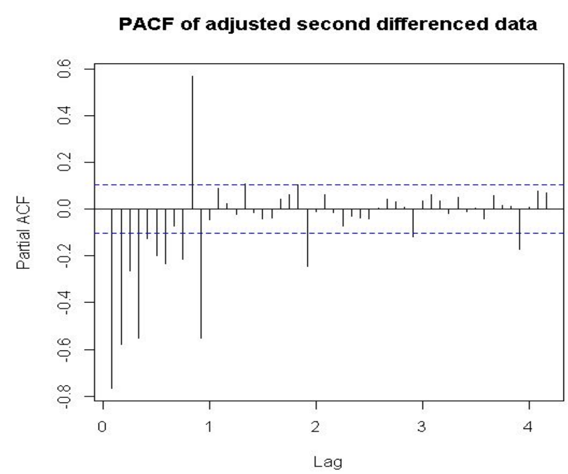
ACF and PACF after taking second difference:



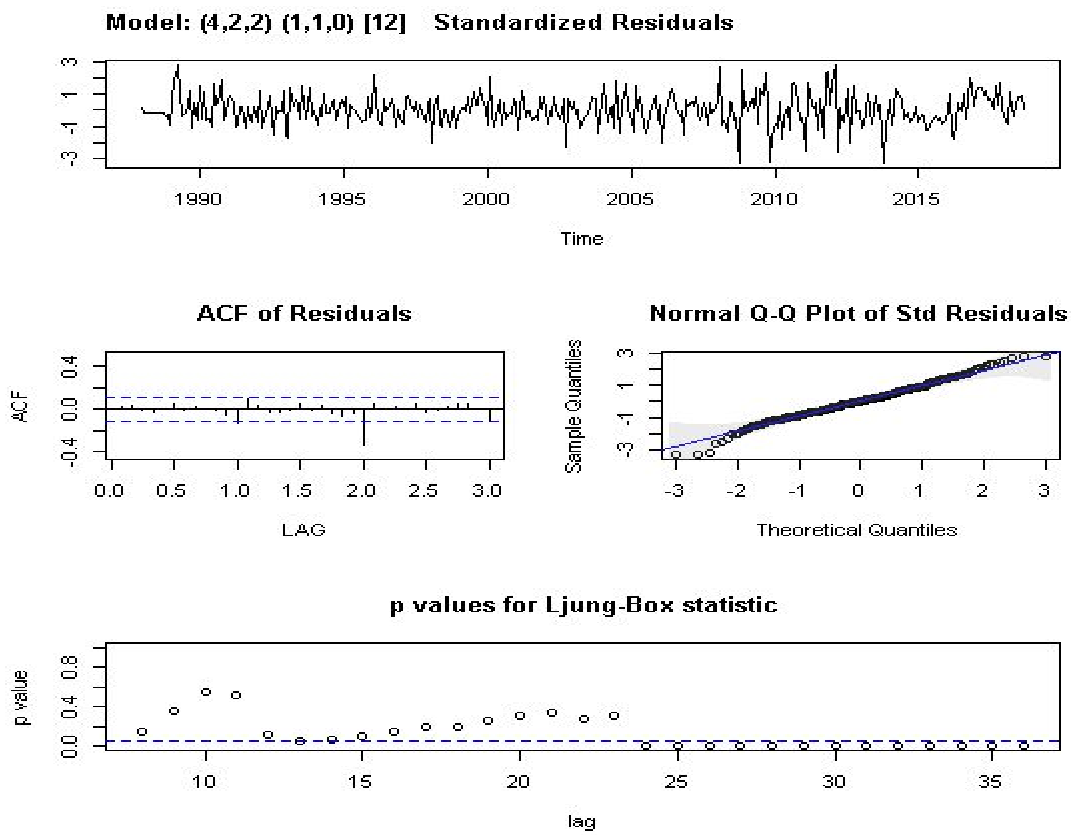


We see seasonality at lag 12 and hence take seasonal difference which results in the following ACF and PACF.





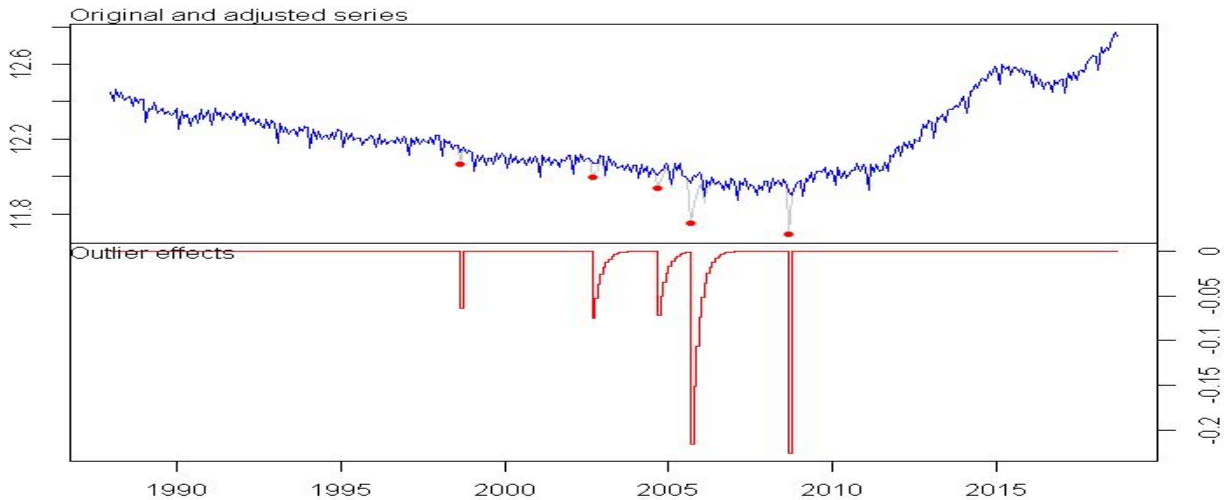
Note that the data has now become stationary and we fit seasonal ARIMA model to this adjusted data and further do diagnostic checking before making a forecast.
The diagnostic plot is given below:



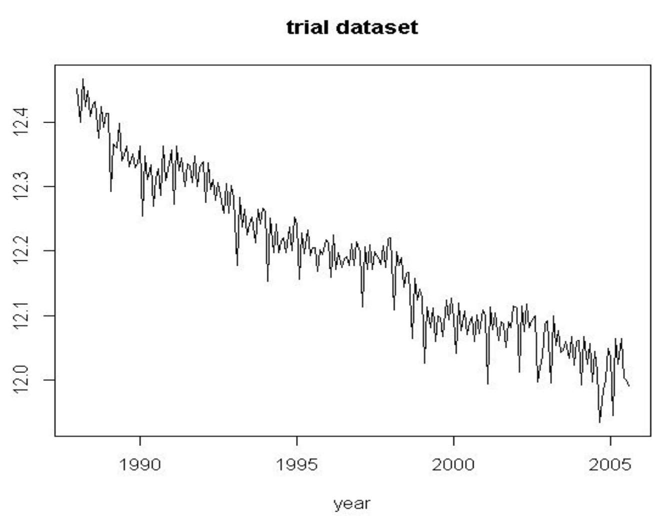
From the diagnostic plot we can see that ACF of the residuals show no correlation. The normal Q-Q plot shows some tailing at the lower end but significantly lesser than the previously fit ARIMA model. However the p values from the Ljung-Box statistics shows that p values are below zero as the lag increases.
Thus outlier detection won't be the best choice for the given the data. Hence intervention analysis might be a better fit on the data.

● Intervention analysis

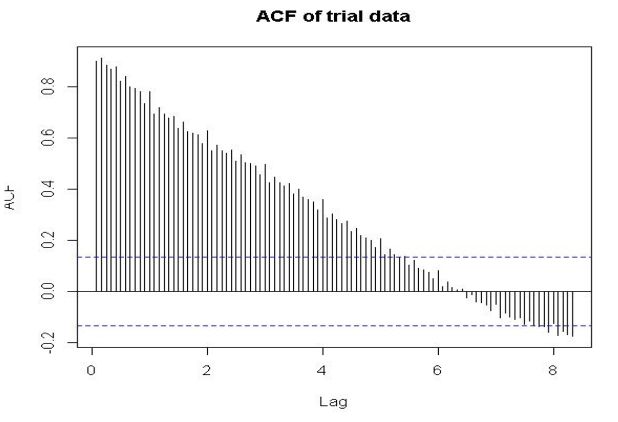
The outlier effect on the data is given as follows.

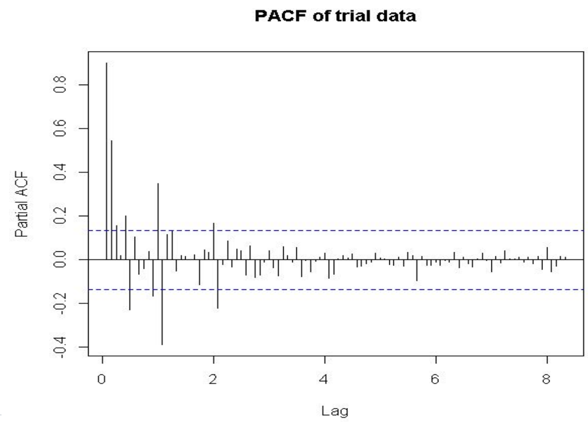


Note that most of the outlier data points have a pulse like pattern. The data point in 2008 however does not show a pulse but is rather an outlier. Hence we fix the 2008 point by the point given by tsoutlier and do intervention analysis the data. We do intervention analysis on the September 2005 data. To do intervention analysis we use the epoch prior to the intervention event. The dataset before Oct 2005 is given as follows

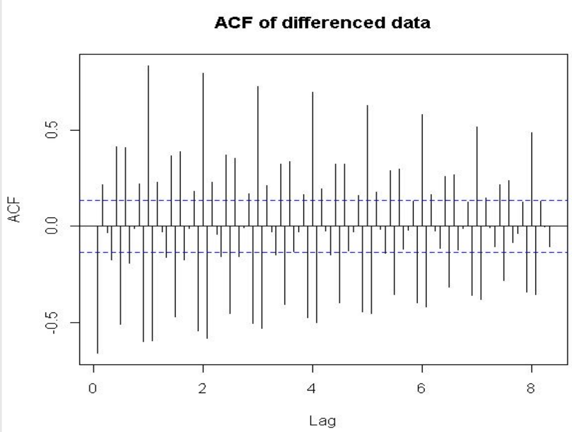


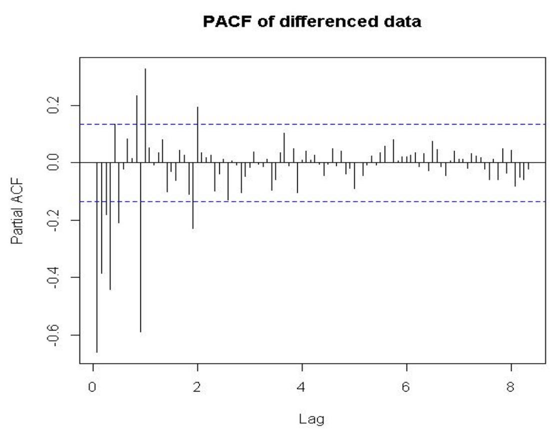
The ACF and PACF of the data is given as:

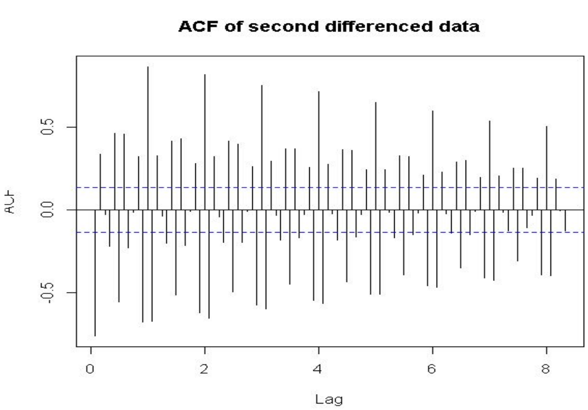


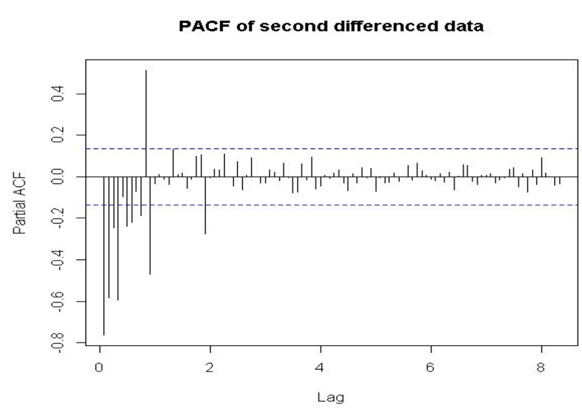


After taking the first and second difference of the data, we get the ACF and PACF of the data as follows:

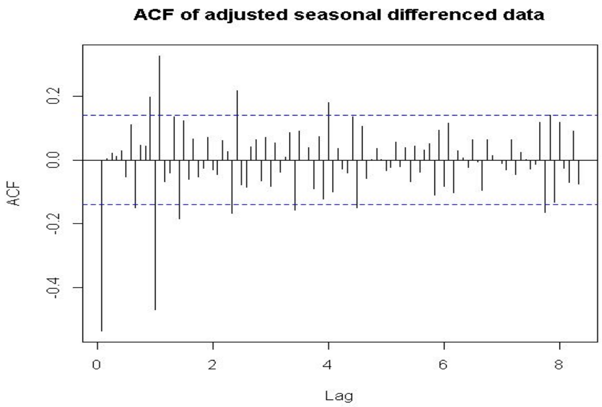


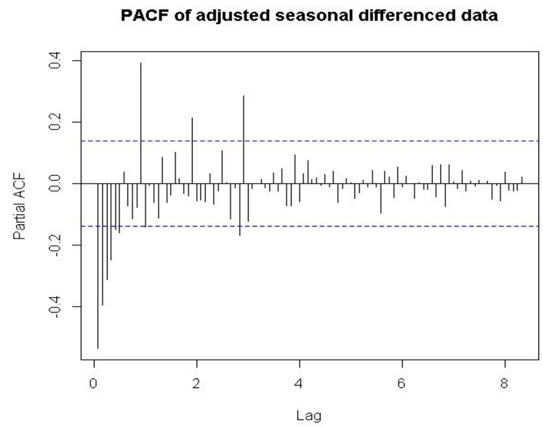




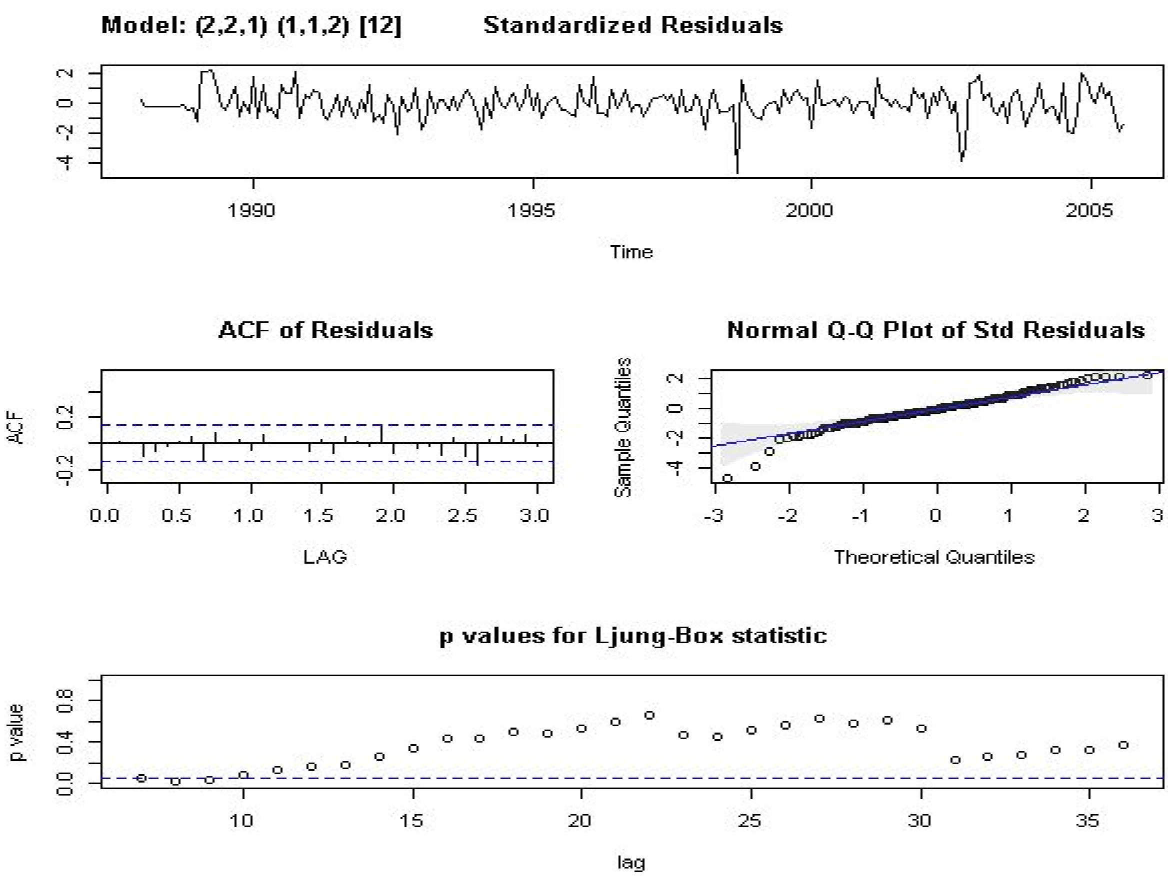


ACF and PACF of the seasonal difference of the data is given as :



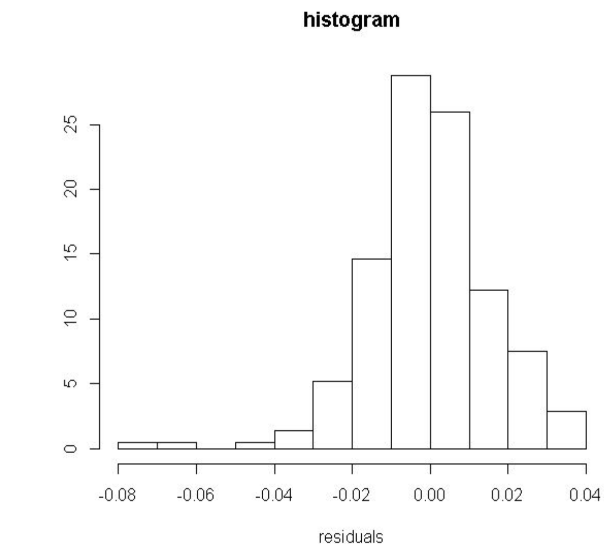


The ARIMA (2,2,1)x(1,1,2) seems to be a better fit on the model and the diagnostics show.

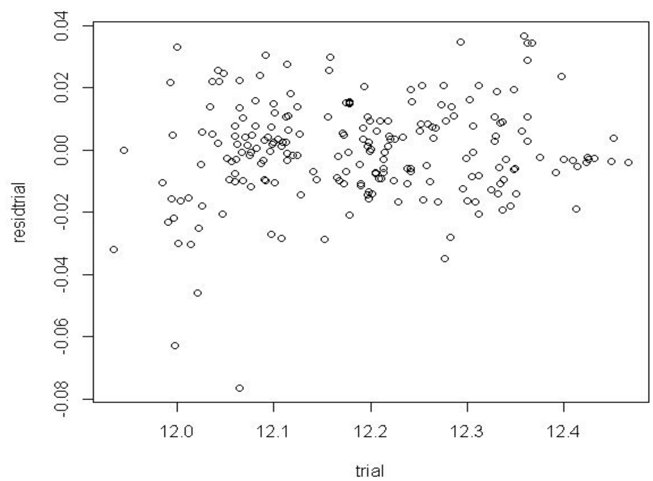


Note that ACF of the residuals show no correlation and the normal Q-Q plot shows the almost linear pattern and the p values are all above 0 hence the given model fits the data well.

The histogram of the residues have a somewhat gaussian curve.



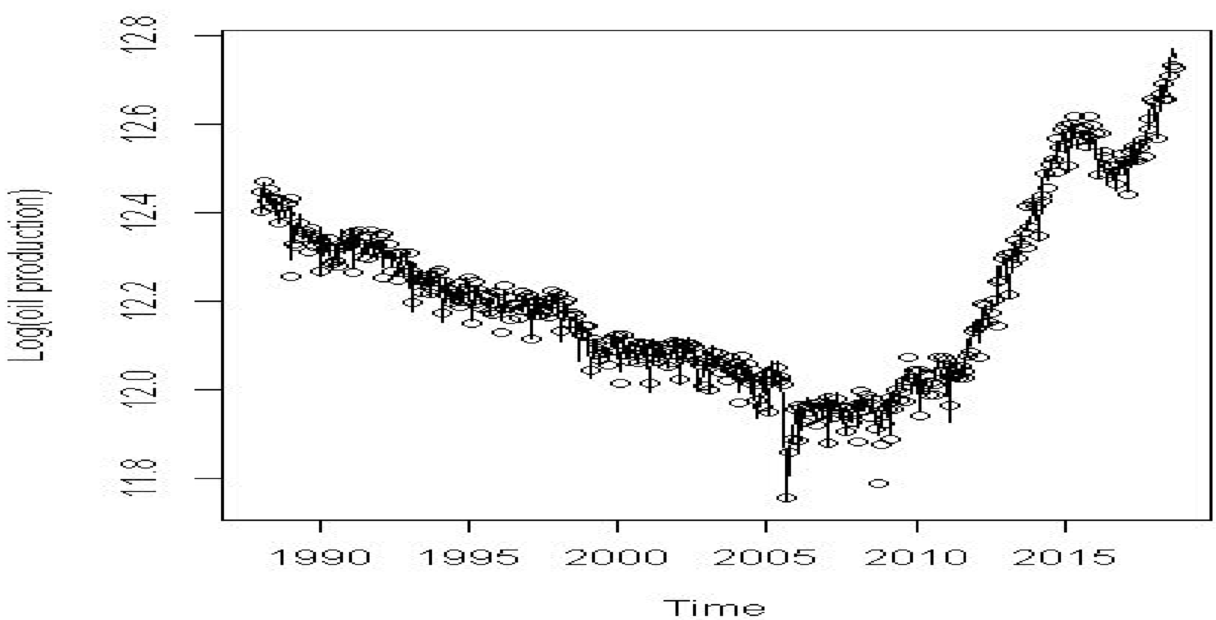
The residual vs the data is a random cloud and it shows a random cloud with a few outliers.



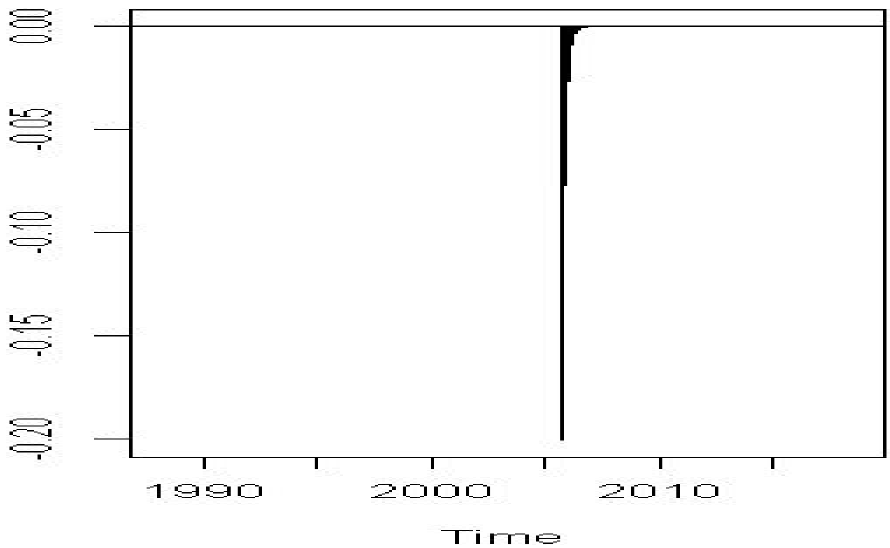
We then do pulse intervention analysis on the data after Oct 2005.
The model can be written as :
$mt = \omega 0\ P\,t(T) + \omega 1/(1 - \omega 2B)P\,t(T)$
Using the airmax function we estimate the parameters. The parameters along with their SE is given as :

coef	double [9]	-0.227 -0.198 -0.974 -0.378 -0.608 -0.388 ...
ar1	double [1]	-0.2269221
ar2	double [1]	-0.1979141
ma1	double [1]	-0.9738904
sar1	double [1]	-0.3781901
sma1	double [1]	-0.6084206
sma2	double [1]	-0.3882251
I911-MA0	double [1]	0.0264696
I911.1-AR1	double [1]	0.5985822
I911.1-MA0	double [1]	-0.2272369
sigma2	double [1]	0.0004807931

The fitted data is given below:

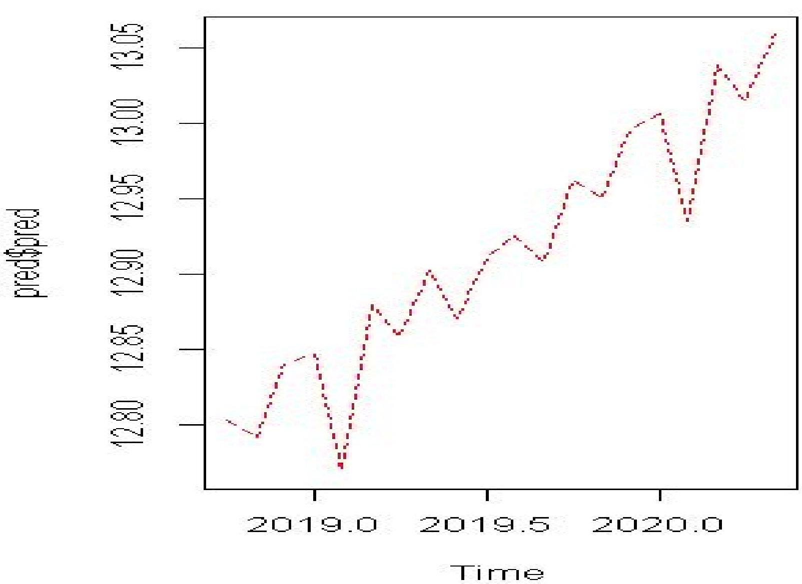


The intervention effects is given by :



3.Forecasting

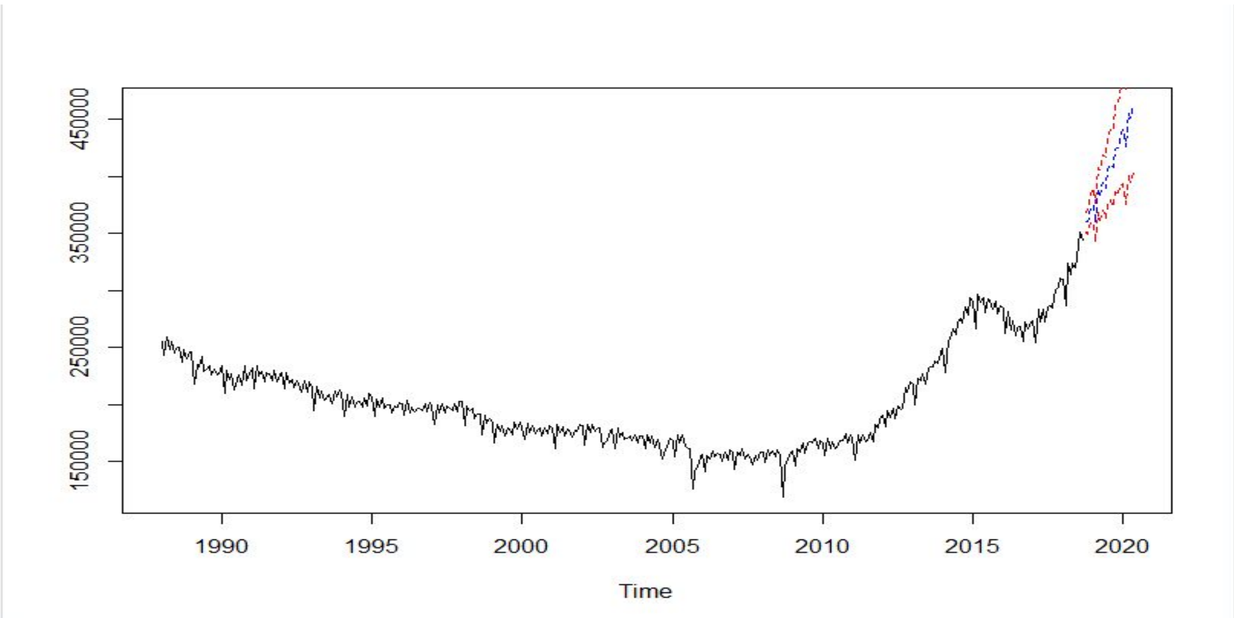
The forecasted time series is given as:



Thus after fitting the ARMA(2,2,1)x(1,1,2) on the data I predict the oil production in USA for the next 20 months. The trend can be seen on figure below and the predicted values can be seen on table.

```
$`pred`
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct
2018                                                                                   12.80370
2019 12.84916 12.77041 12.88114 12.86020 12.90482 12.87290 12.91357 12.92793 12.91103 12.96506
2020 13.00958 12.93902 13.04209 13.01835 13.06236
          Nov      Dec
2018 12.79272 12.84182
2019 12.95265 12.99854
2020

$se
            Jan        Feb        Mar        Apr        May        Jun        Jul        Aug
2018
2019 0.03149596 0.03580946 0.04005343 0.04428064 0.04850770 0.05275097 0.05702215 0.06132917
2020 0.08392451 0.08861832 0.09336861 0.09817631 0.10304186
            Sep        Oct        Nov        Dec
2018            0.01750964 0.02260161 0.02701366
2019 0.06567793 0.07018782 0.07471686 0.07929399
2020
```



4.Conclusion

From the graph we can see that the oil production in USA shows a general upward trend with a few falls for the next 20 months. This trend agrees with the current government policies that aim to increase the crude oil production in USA further.

Bibliography

- https://www.eia.gov/todayinenergy/detail.php?id=37053
- https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=WCRFPUS2&f=W
- https://foreignpolicy.com/2018/09/13/oil-production-record-levels-why-are-oil-prices-heading-higher-opec-iea-venezuela-iran/
- https://money.cnn.com/2018/09/12/investing/us-oil-production-russia-saudi-arabia/index.html

Appendix

Code

```
hwdata = read.table("C:\\Users\\d.krishnan\\Desktop\\data1.csv",  sep=" ")
hwdata = ts(hwdata, frequency =12, start = c(1988,1))
plot(hwdata)
a=acf(hwdata)
b=pacf(hwdata)
hwdata_log=log(hwdata)
plot(hwdata_log)
diff1 = diff(hwdata_log)
c=acf(diff1,50)
d= pacf(diff1)
diff2 =diff(diff1)
acf(diff2)
pacf(diff2)
diff3 = (diff(diff2, 12))
acf(diff3,50)
pacf(diff3)
mdl1 = arima(co2, order = c(2,2,1),seasonal = list(order=c(0,1,1),period=12))
BIC(mdl1)
mdl2 = arima(co2, order = c(1,2,1),seasonal = list(order=c(0,1,1),period=12))
BIC(mdl2)
mdl3 = arima(co2, order = c(2,2,1),seasonal = list(order=c(1,1,1),period=12))
BIC(mdl3)
mdl4 = arima(co2, order = c(1,2,2),seasonal = list(order=c(0,1,1),period=12))
BIC(mdl4)
mdl5 = arima(co2, order = c(2,2,1),seasonal = list(order=c(1,1,2),period=12))
BIC(mdl4)

mdl =  sarima(hwdata_log,2,2,1,1,1,2,12)
resid = residuals(mdl$fit)
hist(resid,probability = T)
plot(hwdata_log,resid)

##outliers

lines(tsclean(hwdata),col='red')
outliers = tso(hwdata_log)
plot(outliers)
outliers[["yadj"]]
plot(outliers[["yadj"]])
acf(outliers[["yadj"]],50, main='ACF of adjusted data')
pacf(outliers[["yadj"]],,50, main='PACF of adjusted data')
```

```
diff0 = diff(outliers[["yadj"]])
acf(diff0,50, main='ACF of adjusted differenced data')
pacf(diff0,50, main='PACF of adjusted differenced data')
diff01 = diff(diff0)
acf(diff01,50, main='ACF of adjusted second differenced data')
pacf(diff01,50, main='PACF of adjusted second differenced data')
diff012 = diff(diff01,12)
acf(diff012,50, main='ACF of adjusted seasonal differenced data')
pacf(diff012,50, main='PACF of adjusted seasonal differenced data')

mdl_out =  sarima(outliers[["yadj"]],4,2,2,1,1,0,12)

# intervention analysis

hwdata[249]=155873
trial = log(hwdata[1:212])
trial = ts(trial, frequency = 12,start=c(1988,1))
plot(trial,main = 'trial dataset', xlab = 'year', ylab='')
acf(trial,100,main='ACF of trial data')
pacf(trial,100, main='PACF of trial data')

diffi1 = diff(trial)
plot(diffi1)
acf(diffi1,100, main='ACF of differenced data')
pacf(diffi1,100,main='PACF of differenced data')

diffi2 = diff(diffi1)
plot(diffi2)
acf(diffi2,100,main='ACF of second differenced data')
pacf(diffi2,100,main='PACF of second differenced data')

sdiffi = diff(diffi2,12)
plot(sdiffi)
acf(sdiffi,100,main='ACF of adjusted seasonal differenced data')
pacf(sdiffi,100,main='PACF of adjusted seasonal differenced data')

mdl_trial = sarima(trial,2,2,1,1,1,2,12)
residtrial = residuals((mdl_trial$fit))
hist(residtrial,probability = T, main = 'histogram',xlab='residuals',ylab='')
plot(trial,residtrial)
oil.m1=arimax(hwdata_log,order=c(2,2,1),
        seasonal=list(order=c(1,1,2),period=12),
        xtransf=data.frame(I911=1*(seq(hwdata)==213),
                 I911=1*(seq(hwdata)==213)),
        transfer=list(c(0,0),c(1,0)),
        method='ML')

plot(log(hwdata),ylab='Log(oil production)')
```

```
points(fitted(oil.m1))
Nine11p=1*(seq(log(hwdata))==213)
plot(ts(Nine11p*(0.0265) + filter(Nine11p,filter=.583,method='recursive', side=1)*
        (-0.2272),frequency=12,start=1988),ylab='oil production',
    type='h'); abline(h=0)

library(stats)

tf =  1*(seq(1:(length(hwdata)+20))==213)*(0.0265) +
filter(1*(seq(1:(length(hwdata)+20))==213),filter=0.583,method='recursive',side=1)*(-0.2272)
forecast.arima = arima(log(hwdata),order=c(2,2,1), seasonal = c(1,1,2), xreg=tf[(1:(length(tf)-20))])
forecast.arima
pred=predict(forecast.arima,n.ahead = 20, newxreg=tf[370:length(tf)])
plot(pred$pred, type ='l', col='red',lty=2)
plot(cbind(hwdata, pred$pred), plot.type = "single", ylab = "", type = "n")
lines(hwdata)
lines(pred$pred, col='blue',lty=2)
lines(pred$pred + 1.96 * pred$se, type = "l", col = "red", lty = 2)
lines(pred$pred - 1.96 * pred$se, type = "l", col = "red", lty = 2)

#plot(pred)
#tc = forecast(forecast.arima,n.ahead = 20, xreg=tf[370:length(tf)])
#plot(tc)

tt = auto.arima(trial)
plot(forecast(tt))
```