**Predicting Crude Oil Prices Fluctuation**

**using**

**Times Series Forecasting and Text Mining in R**

April 26,2017

**Jaclyn Giroux**

**Xiaorong Tan**

**Shaik Imam Javid**

**Arun Mantra Kalathil**

**Mazhar Kodithodika**

**Mohammed Shoeb Ahmed Shakeb**

**Content**

1. **Problem Statement…………………………………………………………………….3**
2. **Background…………………………………………………………………………….3**
3. **Methodology and Results………………………………………………………...........5**
4. **Conclusions and recommendations…………………………………………………. .27**
5. **References…………………………………………………………………………….. 28**

* **Problem Statement**

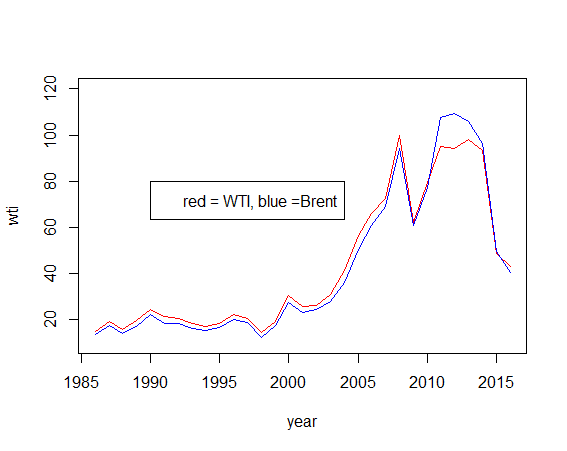
In the business environment, predictive modeling has become an important tool for gaining valuable insight into what influences particular outcomes. Whether it be a trading company in futures market, accurate prediction of Crude Oil price forecast will help to maximize its profit. Through predictive modeling, R can aid businesses across industries in discovering the business value hidden in their datasets.

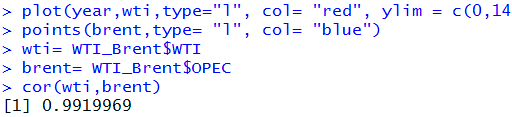
Oftentimes historical data can be a good predictor of future costs of goods and services. Times series forecasting can be used to create a good model with these predictions. However, many times price changes or outcomes are influenced by more than just historical data. The global political and economic environment can have a great impact of cost of goods and services. To account for these more qualitative factors, text mining can be integrated into the predictive modeling process. Oil prices provide an example of a commodity that is strongly influenced by global factors. In our project, we seek to illustrate how historical data and text analysis can be used to predict changes in oil prices using packages available in R.

* **Background**

Oil is a commodity, and as such, it tends to see larger fluctuations in price than more stable investments such as stocks and bonds. OPEC controls 40% of the world's supply of oil. As with any commodity, stock or bond, the laws of demand and supply cause oil prices to change. The fall in oil prices can be attributed a lower demand for oil in Europe and China, coupled with a steady supply of oil from OPEC. While demand and supply affect oil prices, it is actually oil futures that set the price of oil. Geopolitical instability also causes oil price to increase.

Major Crude Oil trade is based on Brent Crude oil and West Texas Intermediate (WTI) and their prices is influenced by various factors. Variation of Brent and WTI prices from year 1983 to 2016 were compared as plotted below.





Both the crude oil prices are highly correlated i.e. at 0.99. Hence, we selected to predict only one of the crude oil price i.e. WTI crude Oil Price.

Main influencing factors on the fluctuation of crude oil price is Supply, demand and Geopolitical instability, dollar rate fluctuation, and the Futures market.

News outlets can provide a source of data to assess the impact of global social, economic and political factors on data. In R, there are many ways to scrap data from news webpages. Many packages can be used to complete webpage scraping and text mining include some very common packages such as ‘dplyr’ and ‘ggplot.’ The main packages used to complete text mining are ‘rvest,’‘tm,’‘Snowballc,’‘tidy text,’ ‘nlp,’ ‘magrittr,’ ‘word cloud.’

* **rvest:**

**‘**rvest’ is a package that extract data from different sections of a webpage into vectors or tables.

* **tm:**

The text mining package is used to create a term document matrix to find trends among many documents. It can examine associations, cluster and run other data manipulations within the term document matrix summarizing text trends.

* **tidy text:**

The ‘tidy text’ package assists in data mining by separating out lines of text so that they have a format with one word in each line. It can run sentiment analysis

* **nlp:**

The ‘nlp’ package conducts natural language processing to help the train a model to interpret human speech

* **magrittr:**

The magrittr packages allows the use of piping to make R code easier to read.

Also for time series forecasting, in R there are many packages in which we used ‘tseries’ and ‘forecast’.

* **tseries:**

The ‘tseries’ package can be used for doing “Augmented Dickey-Fuller test “for analysing the trend of time series.

* **forecast:**

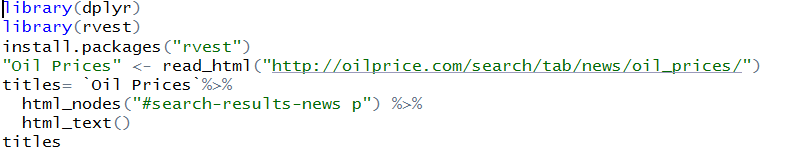
Methods and tools for displaying and analyzing univariate time series forecasts including exponential smoothing via state space models and automatic ARIMA modelling.

**Methodology**

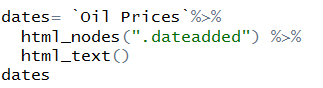
We predict the Crude oil price fluctuation using time series forecasting system and Text mining was used to validate and update the trend based on daily news generation sentiment analysis. Loughran McDonald Financial Dictionary was used to analyze daily news generated. News Data from last 100 days were scraped for our analysis.

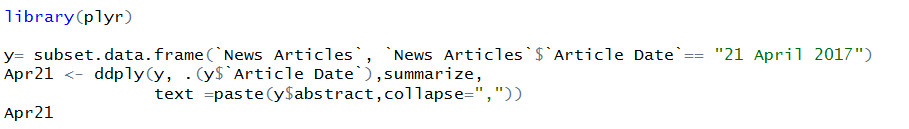
a)Text Scraping Using ***rvest*** Package

News articles from [www.oilprice.com](http://www.oilprice.com) for past 100 days were scraped using rvest package. These were sorted based on date. Codes for scraping one day’s news is shown below



Data was arranged on basis of date using the below code



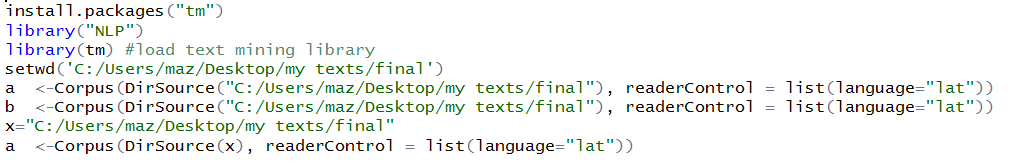


News article was than arranged based on date.

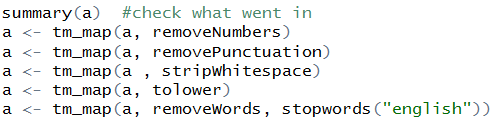
**b) Natural Language Processing**

“tm” package and “NLP” package in R is used for text mining.

Scraped data was placed into a corpus folder and it was made source directory as shown below.



Data was cleaned to remove numbers, stop words, whitespaces and was made into all lower cases to generate uniformity.



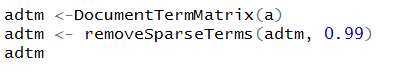
To remove the stemmed form of words and make it into uniform word root stemming mapping was done as shown in the below

**a <- tm\_map(a, stemDocument)**

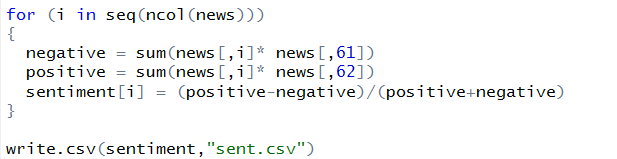
Loughran McDonald Financial Dictionary was placed into the corpus directory, the transform according to them.

Corpus <- tm\_map(Corpus, content\_transformer(stemCompletion), dictionary = positive)

Corpus <- tm\_map(Corpus, content\_transformer(stemCompletion), dictionary = negative)

4) Transformed into term document matrix and sparse words were removed.

5) Sentiment for daily news were generated using a for loop.

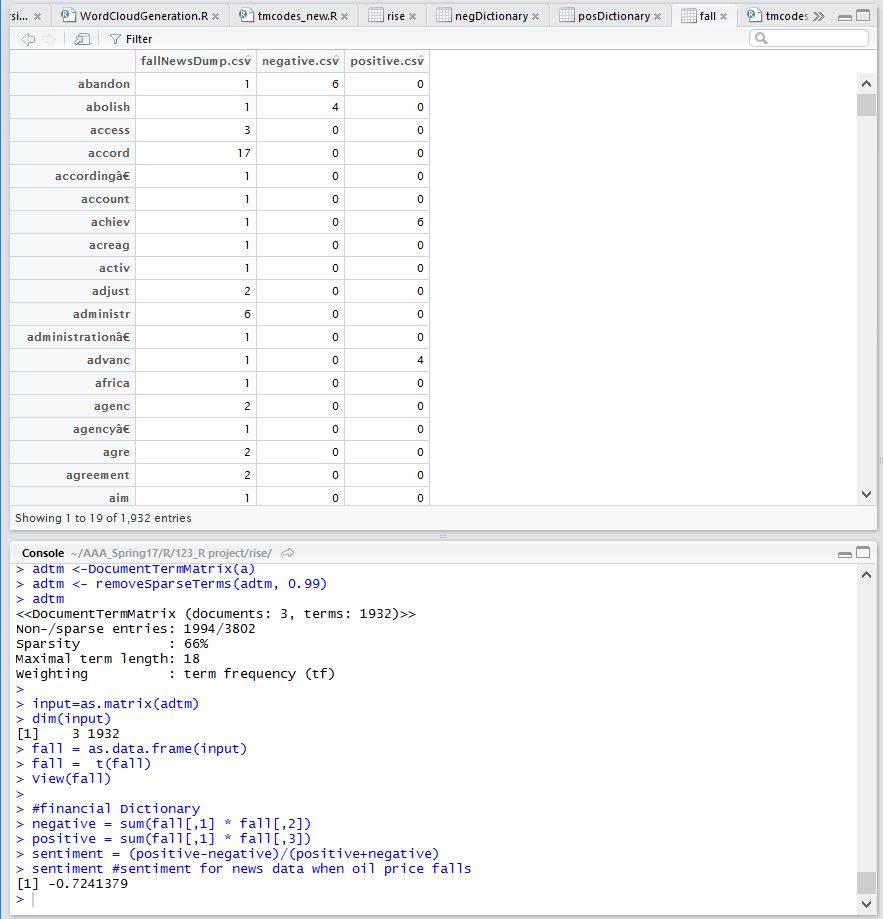


Sentiments time series for daily news was generated and saved into file named “sents.csv”.

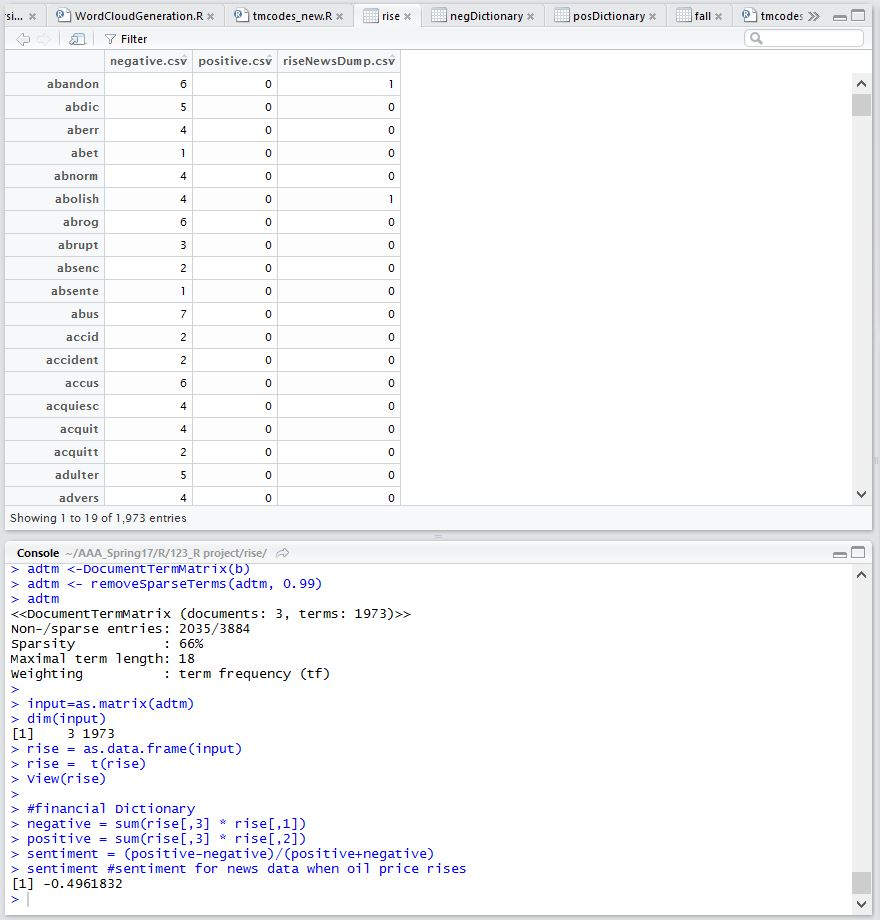
Correlation of the future price fluctuation and daily news sentiment time series a found out to less. This is suggestive that we must scrape from multiple sources and larger time period will give a more affirmative conclusive result.

Hence, we classified the news when the oil price increased and oil price decreased categories. Sentiment for these two categories was analyzed and the results is shown

**Sentiment for News data when oil price falls:**



**Sentiment for News data when oil price rises:**



News articles when oil price was falling and News article when oil price was rising was segregated and sentiment analysis was done. When oil price was falling the sentiment was more negative (-0.7241379) and when the oil price was rising sentiment was less negative (-0.4961832).

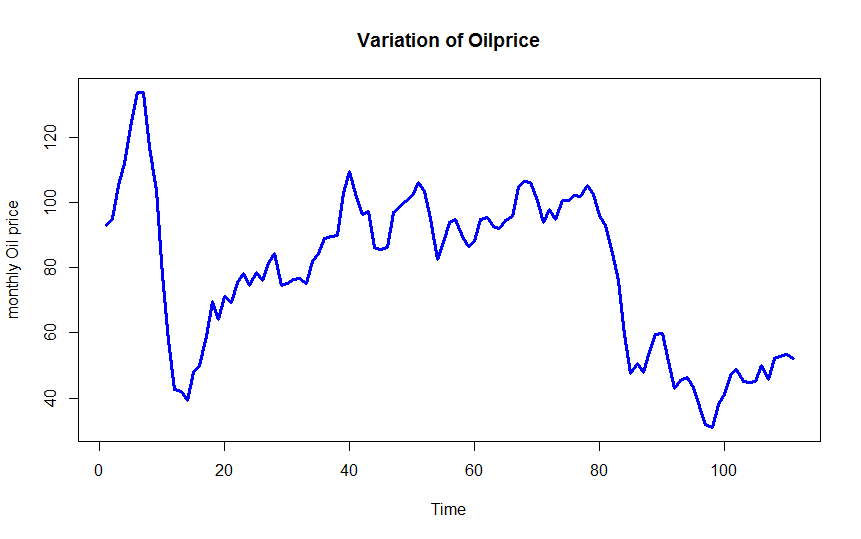
The trend was consisted when normal dictionary was used instead of financial dictionary.

We also tried combination of the dictionary to and got the similar result.

**Time series Forecasting for Predicting Oil Price**

1. Plot of crude oil price (WTI)

plot(y, type = 'l',xlab='Time',ylab='monthly Oil price',col="blue",main="Variation of Oilprice",lwd=3)

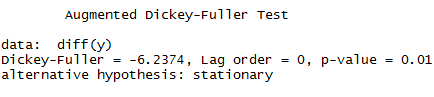


2) **Checking Trend and seasonality**

install.packages("tseries")

library(tseries)

adf.test(diff(y), alternative="stationary", k=0)



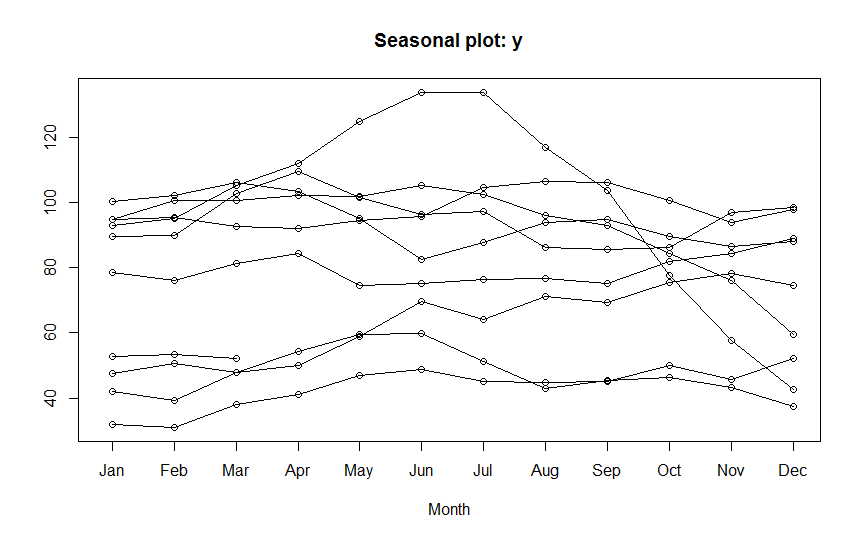
Test results shows there is trend in the time series.

install.packages("forecast")

library(forecast)

seasonplot(y,12,test="ocsb")

nsdiffs(y,12,test="ocsb")





From the plot, it’s not so easy to analyse the seasonality. But the test result of ocsb shows there is no seasonality in the time series.

3)TIme series forecasting using Exponential Model.

fit <- ets(y)

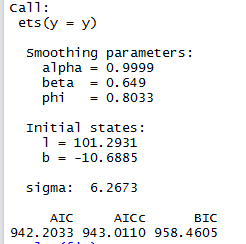
fit

plot(fit)

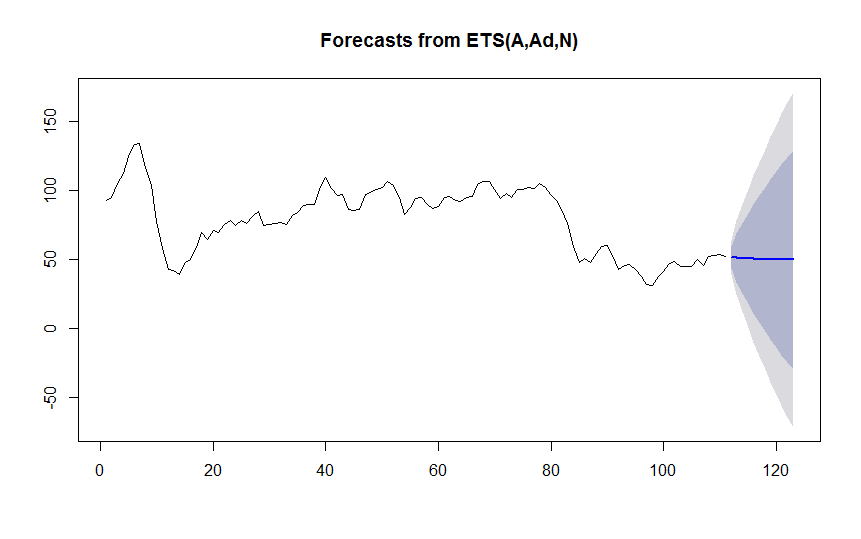
accuracy(fit)

forecast(fit, 12)

plot(forecast(fit, 12))







**4) Autocorrelation and Partial autocorrelation plots.**

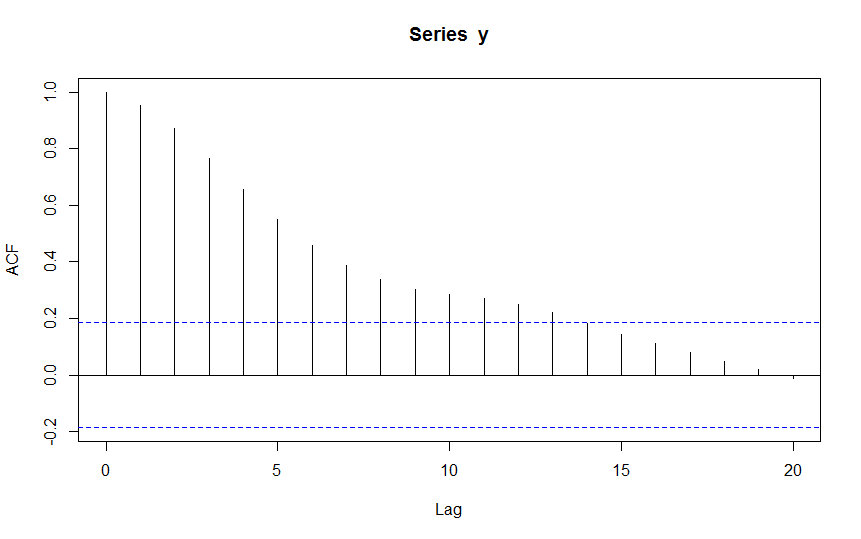
acf(y)

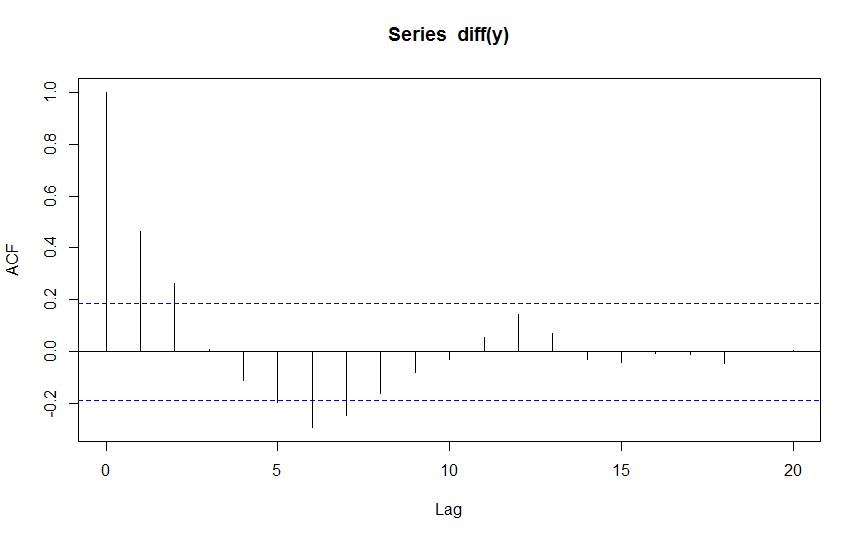
acf(diff(y))

pacf(y)

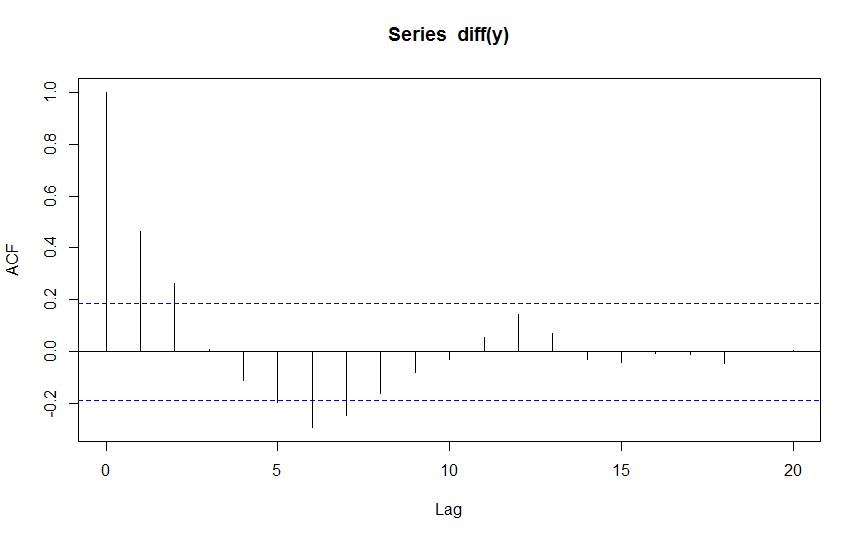
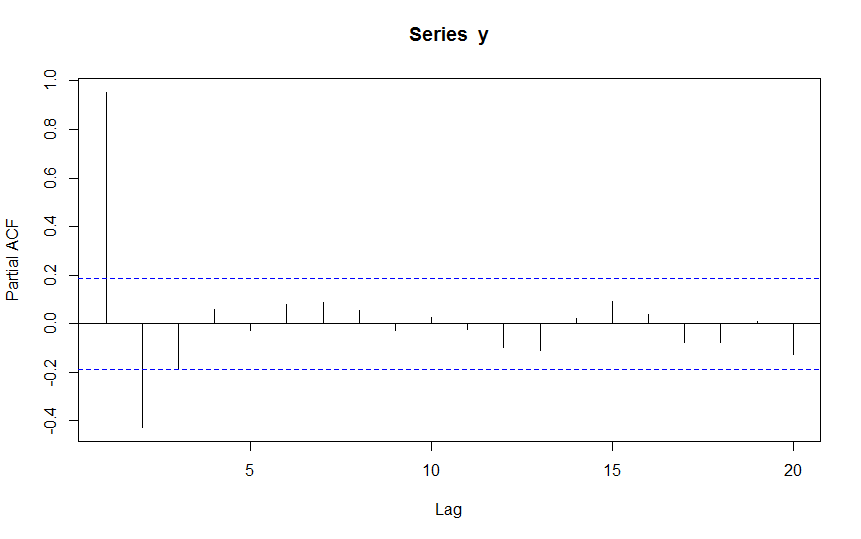
pacf(diff(y))

Autocorrelation plots with and without differencing.





Partial autocorrelation plots with and without differencing.



From the plots, we found the Values for p, d and q as 2,1,2 respectively.

**5) Time Series Forecasting using ARIMA Model.**

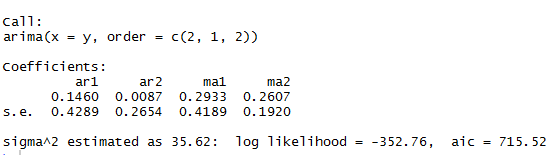
fit <- arima(y, order=c(2, 1, 2))

fit

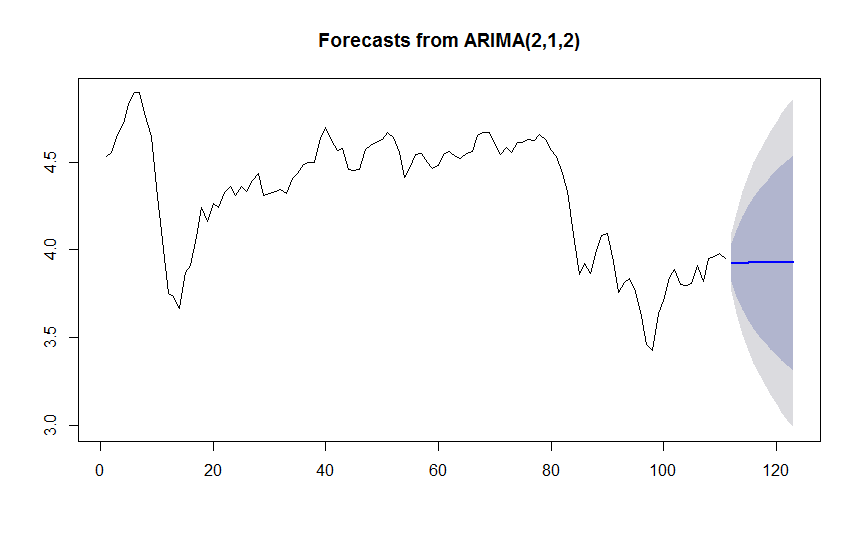
accuracy(fit)

forecast(fit,12)

plot(forecast(fit,12))





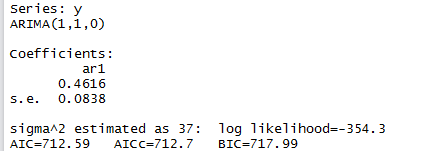
**6) Time Series Forecasting using ARIMA Model- Using Auto Arima function.**

model\_arima <- auto.arima(y)

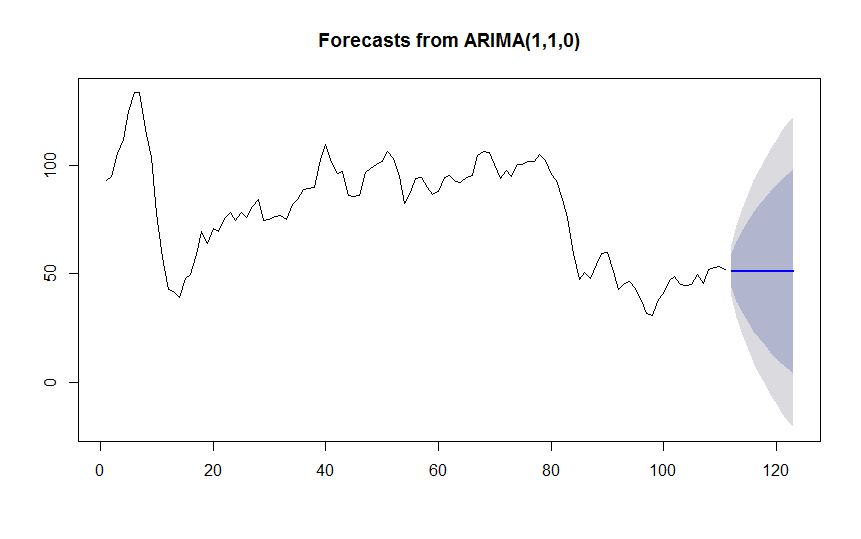
model\_arima

fcast\_arima <- forecast(model\_arima, h = 12)

plot(fcast\_arima)







**7) Adding Repressors to the Arima Model to accommodate Oil demand in Future.**

We assume oil consumption is proportional population increase and electricity use increase overtime. Thus, this were ared added as regressors to the models. We acquired data for these indicators from

Here by adding three different regressors ( Population, oil Eq Per Capita and KWH per capita) for building the ARIMA model , we analyzed if the model is showing any further improvements. Here we tried all the combinations of Regressors. Here pred1, pred2 and pred3 are the predicted values of next 12 months of the respective regressors. We predicted those values using auto ARIMA function.

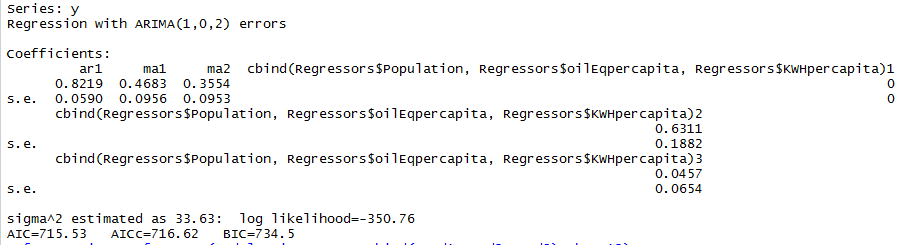
a)With Regressors- Population, oil Eq per Capita and KWH per capita.

model\_arima <- auto.arima(y,xreg = cbind(Regressors$Population,Regressors$oilEqpercapita,Regressors$KWHpercapita))

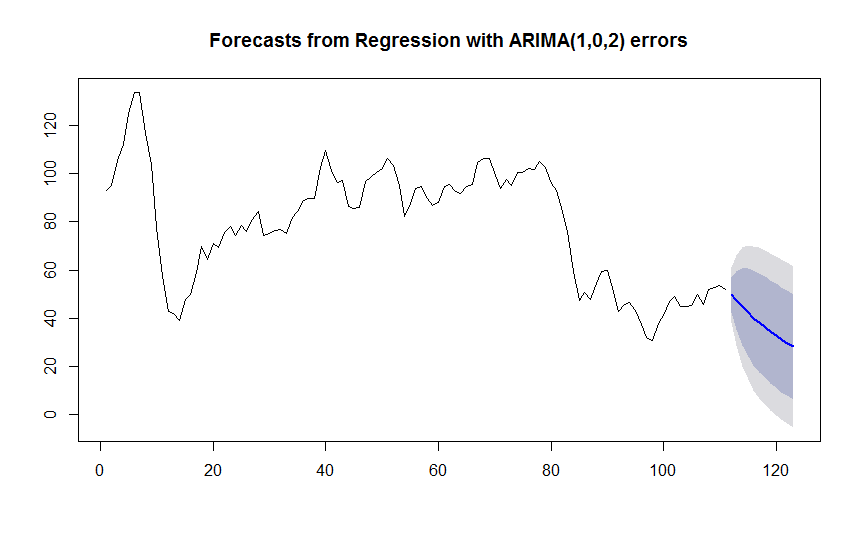
model\_arima

fcast\_arima <- forecast(model\_arima,xreg = cbind(pred1,pred2,pred3), h = 12)

plot(fcast\_arima)







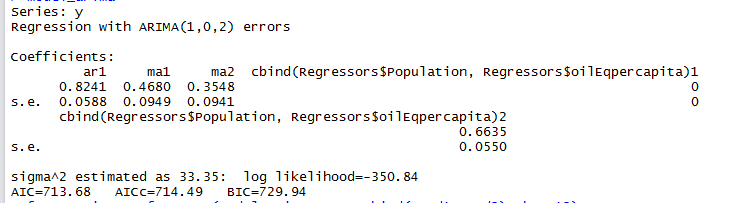
**b)With Regressors-Population, oil Eq per Capita .**

model\_arima <- auto.arima(y,xreg = cbind(Regressors$Population,Regressors$oilEqpercapita))

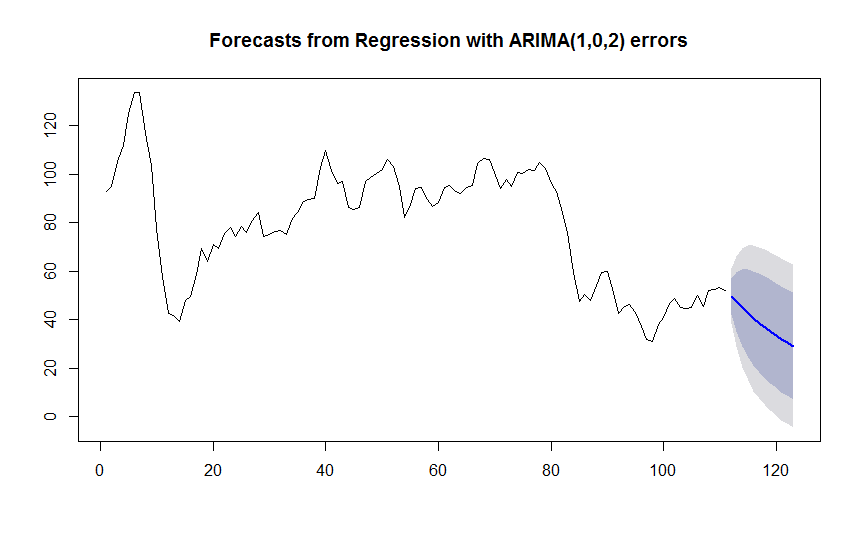
model\_arima

fcast\_arima <- forecast(model\_arima,xreg =cbind(pred1,pred2), h = 12)

plot(fcast\_arima)







As per the above diagram oil price is predicted to decrease in future. This is as result of adding electricity as usage as the regressor due to reduced electricity usage increase due sustainable practices .

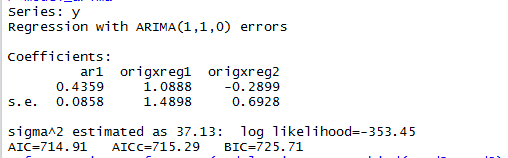
**c)With Regressors- oil Eq per Capita and KWH per capita.**

model\_arima <- auto.arima(y,xreg = cbind(Regressors$oilEqpercapita,Regressors$KWHpercapita))

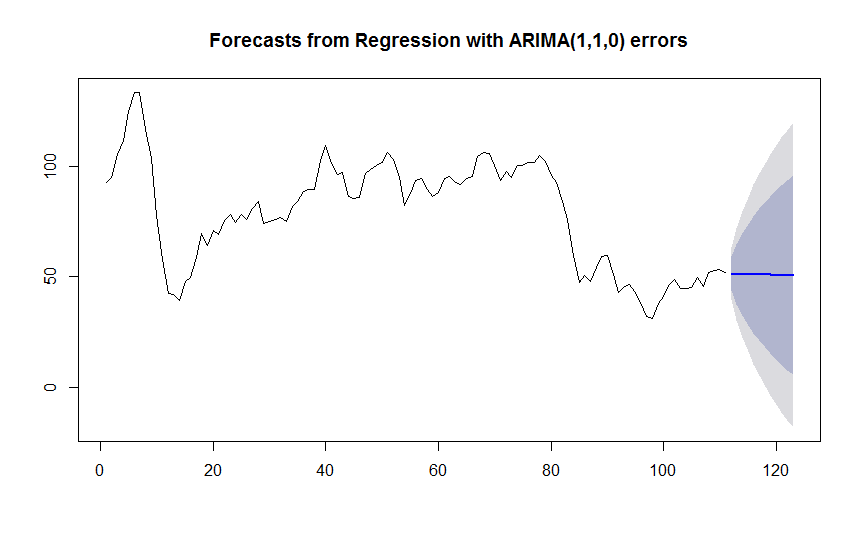
model\_arima

fcast\_arima <- forecast(model\_arima,xreg =cbind(pred2,pred3), h = 12)

plot(fcast\_arima)







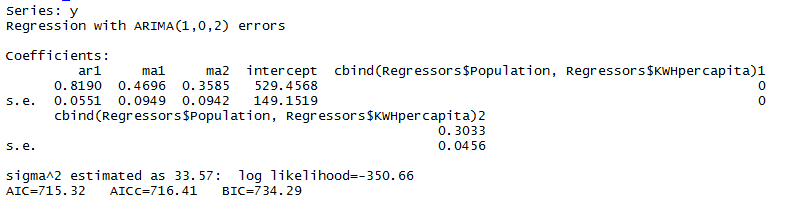
d)With Regressors -Population and KWH per capita.

model\_arima <- auto.arima(y,xreg = cbind(Regressors$Population,Regressors$KWHpercapita))

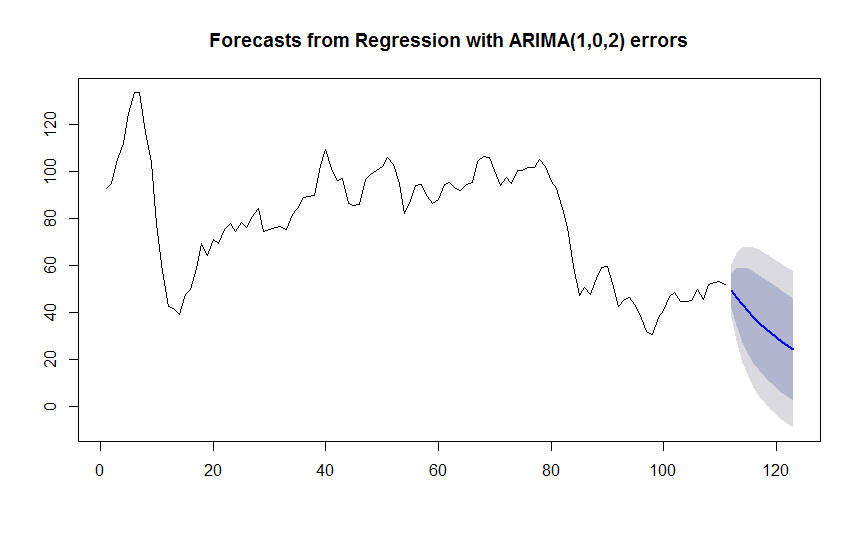
model\_arima

fcast\_arima <- forecast(model\_arima,xreg =cbind(pred1,pred3), h = 12)

plot(fcast\_arima)







As per the above diagram oil price is predicted to decrease in future. This is as result of adding electricity as usage as the regressor due to reduced electricity usage increase due sustainable practices

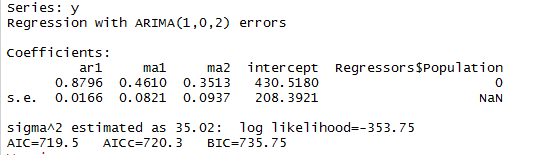
e)With Regressors Population.

model\_arima <- auto.arima(y,xreg = Regressors$Population)

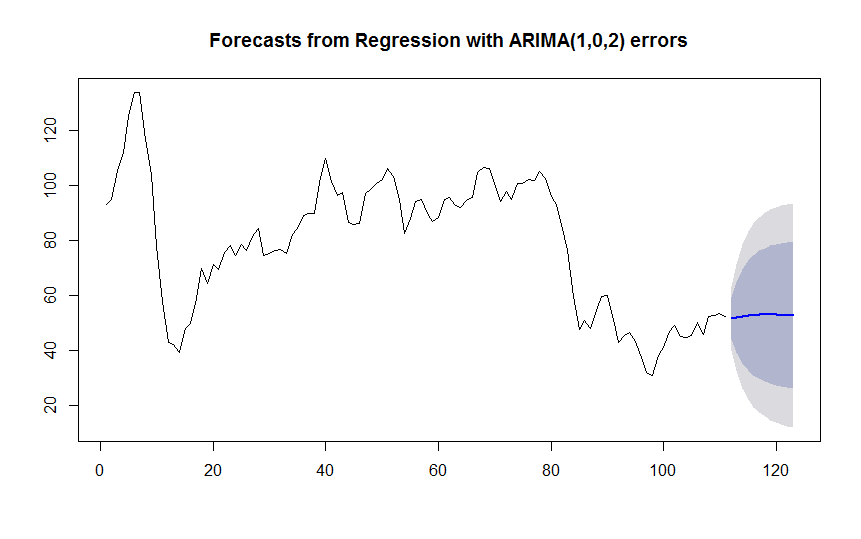
model\_arima

fcast\_arima <- forecast(model\_arima,xreg = pred1, h = 12)

plot(fcast\_arima)







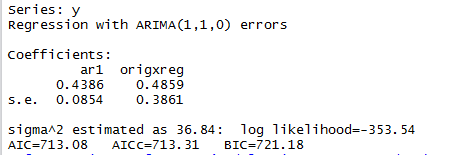
f) With Regressors -oil Eq per Capita.

model\_arima <- auto.arima(y,xreg = Regressors$oilEqpercapita)

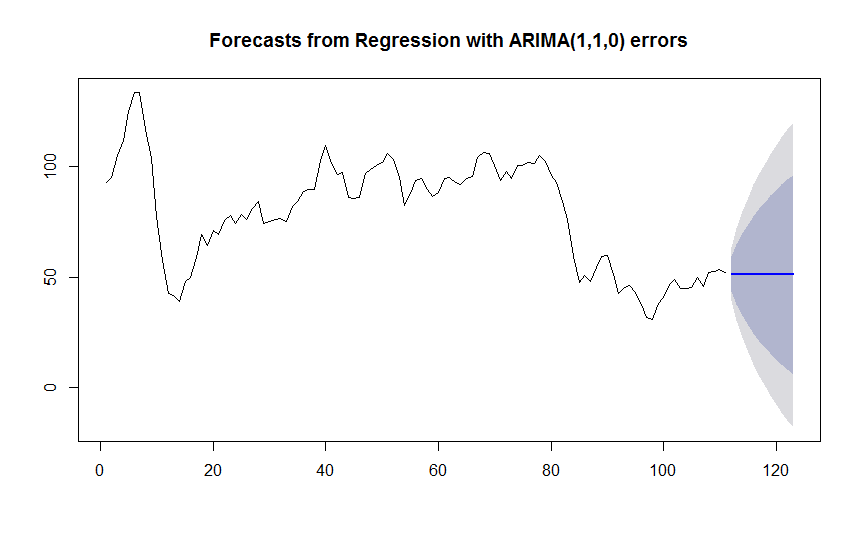
model\_arima

fcast\_arima <- forecast(model\_arima,xreg = pred2, h = 12)

plot(fcast\_arima)







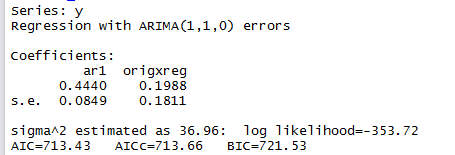
g)With Regressors - KWH per capita.

model\_arima <- auto.arima(y,xreg = Regressors$KWHpercapita)

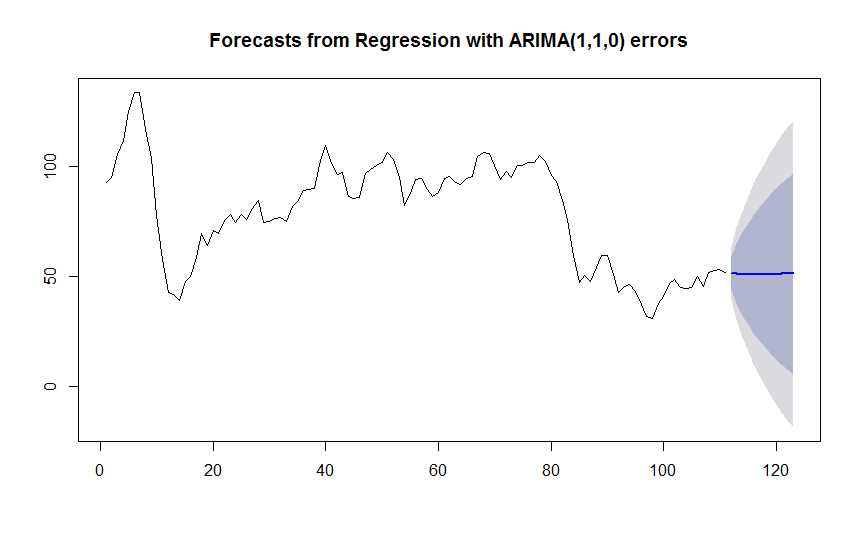
model\_arima

fcast\_arima <- forecast(model\_arima,xreg = pred3, h = 12)

plot(fcast\_arima)

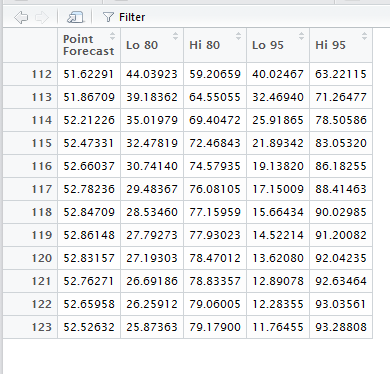






**8) Forecasted values with ARIMA model with regressor as only Population.**

Here we forecasted the values for next 12 months based on the Arima model with regressor as only population. Though we couldn’t find much difference in all models, this seems to be better model considering both statistics of fit and face validity test.



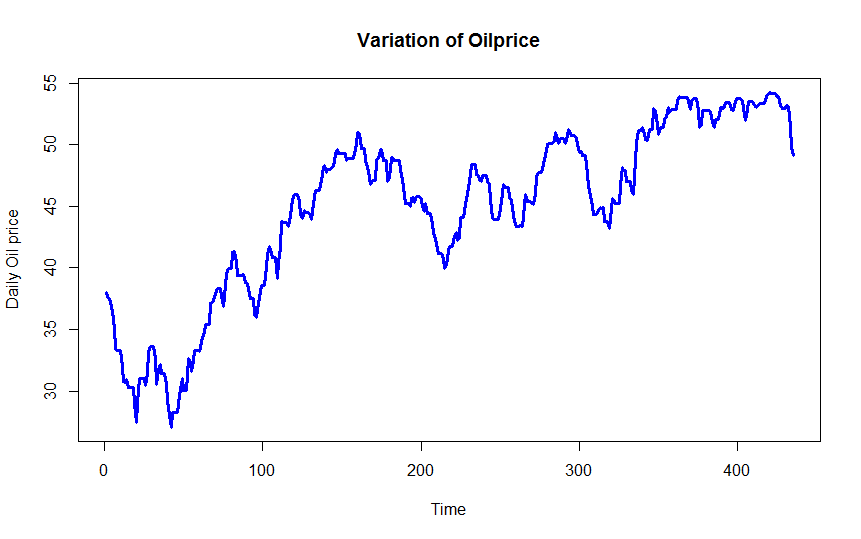
**FORECAST DAILY CRUDE OIL PRICE**

In futures market forecast of daily price accommodating the recent trend will more useful to make efficient decision combined with the monthly forecast.

**Time series forecasting for daily price using Arima Model.**

Here we did time series forecasting of daily oil price using Arima Models.

plot(y, type = 'l',xlab='Time',ylab='Daily Oil price',col="blue",main="Variation of Oilprice",lwd=3)



a) Arima Model (2,1,1) ( from acf and pacf plots)

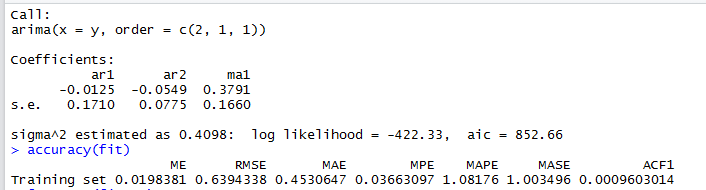
fit <- arima(y, order=c(2, 1, 1))

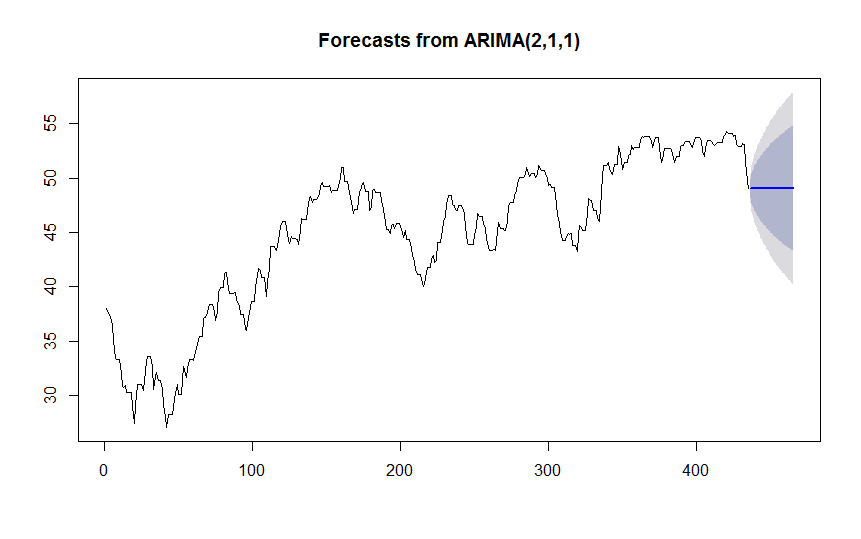
fit

accuracy(fit)

forecast(fit,30)

plot(forecast(fit,30))





b) AutoArima

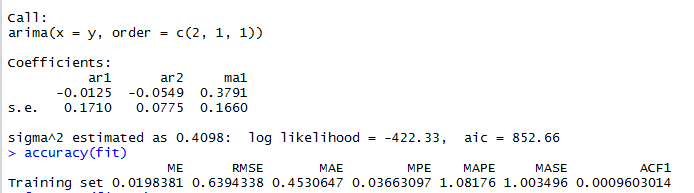
model\_arima <- auto.arima(y)

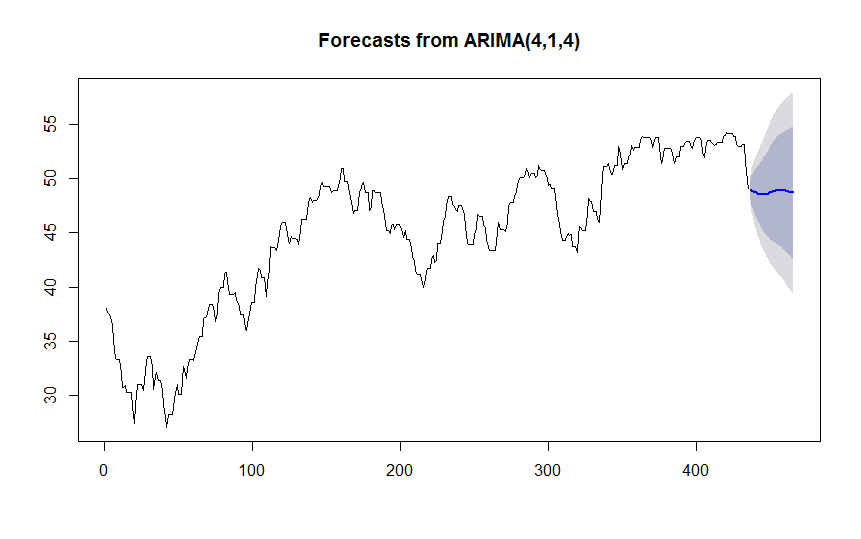
model\_arima

accuracy(model\_arima)

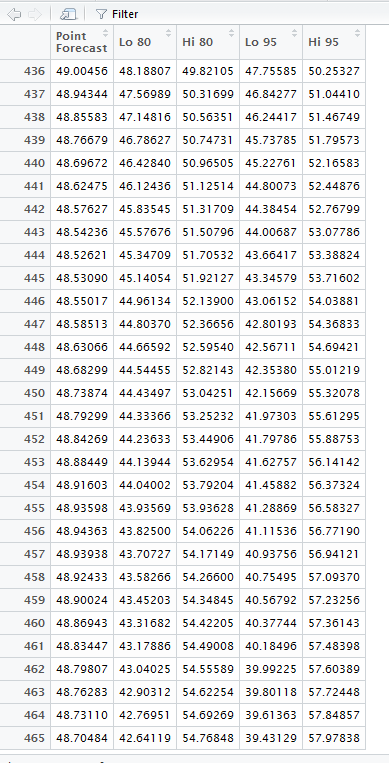
fcast\_arima <- forecast(model\_arima, h = 30)

plot(fcast\_arima)





**c) Forecasted values for next 30 days using the Arima (4,1,4)Model.**



* **Conclusions and Recommendations**
* Using time series forecasting, oil prices can be predicted up to MAPE of 1.08%
* However, real time changes can be accommodated into model using text mining.
* News article text mining can help to predict the trend of oil price fluctuation. This can be acquired by more data scraping from multiple news websites which will improve the accuracy of the prediction.
* More comprehensive oil specific training dictionary and geopolitical stability index could be made for better predictive accuracy and added as regressor into the time series.
* **References**

[**https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf**](https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf)

[**https://cran.r-project.org/web/packages/rvest/rvest.pdf**](https://cran.r-project.org/web/packages/rvest/rvest.pdf)

[**https://cran.r-project.org/web/packages**](https://cran.r-project.org/web/packages/rvest/rvest.pdf)**/tseries/tseries.pdf**

**https://cran.r-project.org/web/packages/forecast/forecast.pdf**

[**https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html**](https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html)

[**https://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html**](https://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html)

[**https://cran.r-project.org/web/packages/NLP/NLP.pdf**](https://cran.r-project.org/web/packages/NLP/NLP.pdf)

[**https://blog.rstudio.org/2014/11/24/rvest-easy-web-scraping-with-r/**](https://blog.rstudio.org/2014/11/24/rvest-easy-web-scraping-with-r/)

[**https://rstudio-pubs-static.s3.amazonaws.com/66739\_c4422a1761bd4ee0b0bb 8821d7780e12.html**](https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html)

[**https://cran.r-project.org/web/packages/tidytext/vignettes/tidying\_casting.html**](https://cran.r-project.org/web/packages/tidytext/vignettes/tidying_casting.html)

[**http://data.worldbank.org/**](http://data.worldbank.org/)

[**http://oilprice.com/**](http://oilprice.com/)

***“Analysis of the International Oil Price Fluctuations and Its Influencing Factors”* Lingyu Yan School of Earth Sciences and Resources, China University of Geosciences (Beijing), Beijing, China.**