Capstone Project Milestone Report

Urmidola Raye

An introduction to the problem: The oil market is suffering a severe downturn due to imbalance between production and demand of crude oil supply. The production has increased thousand times than the demand and as a result the oil producing countries are forced to cut down prices which resulted in layoffs of thousands of workers from their job. The United States of America also suffered heavily due to such imbalance between supply and demand of crude oil. There is practically no profitable wells in the US left to drill oil and as such they depend on import of crude oil from other countries. Countries like Algeria and Nigeria which used to export crude oil only to the US started competing for Asian markets as well. In such scenario, this study will take an attempt to forecast the export volume of crude oil between the United States and Canada, Canada being its closest neighbor.

The study will try to execute various forecasting methods especially ARIMA to predict the export volume of crude oil between US and Canada in 2015 based on historical dataset from 2000 till 2014. The forecasted values will then be compared with the actual 2015 export volume data and the accuracy of the model will be justified. If the export volume shows an increasing trend even during the recession time of oil market, then it will put pressure on the OPEC, a cartel of oil producers to reduce oil production to help firm up prices as well as mop up global market shares which in turn will improve the economic condition of the oil market.

Important field and information of the data set: The data was downloaded from http://open.canada.ca/data/. The data on origin country, destination country, number of megabarrels exported per day and number of days in a month in which export took place will be considered.

Limitations of the dataset: The dataset shows seasonality i.e. in some months the export volume shows peaks and some months show troughs but such disparity in export volume cannot be explained from our dataset.

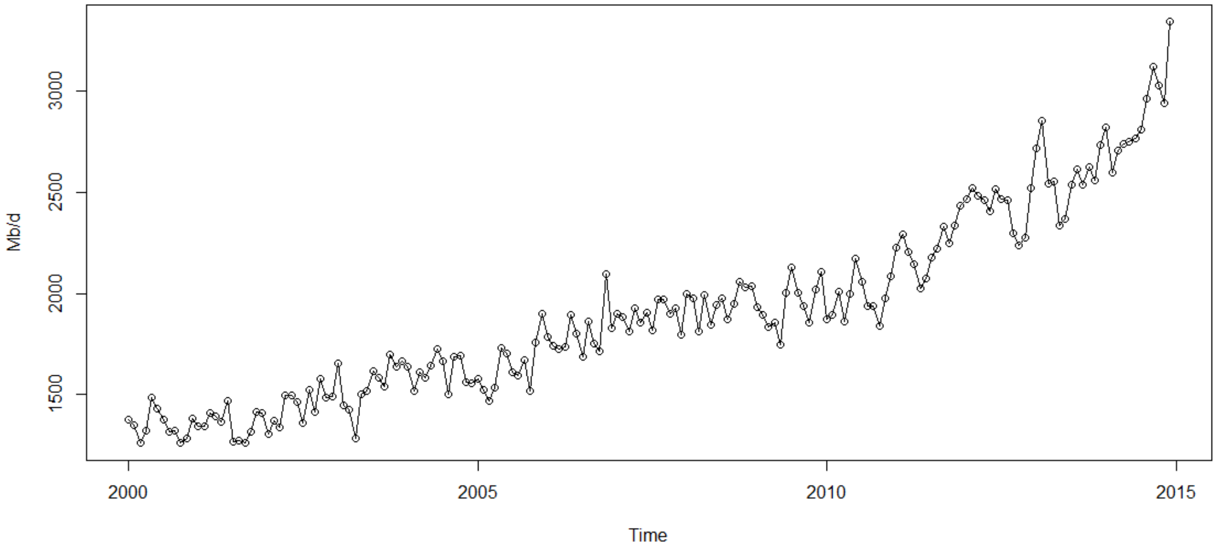
Cleaning and wrangling of the dataset: The downloaded data was in the excel format. One file was saved with data from 2000 till 2014 and the other file was saved with data from 2000 till 2015. The previous one will be used as training set and the latter as a testing set. They were saved as a .CSV file. Apart from that the data looked quite clean. The data was checked for missing values but none were found.

Preliminary exploration and initial findings: At first the data was read into R as a .csv file and then plotted using the ts() function. This ts() function converted the data into a time series format.

exportts <- ts(export, frequency = 12, start = c(2000,1), end = c(2014, 12))

exports

plot(exportts, type = "o", ylab = "Mb/d")



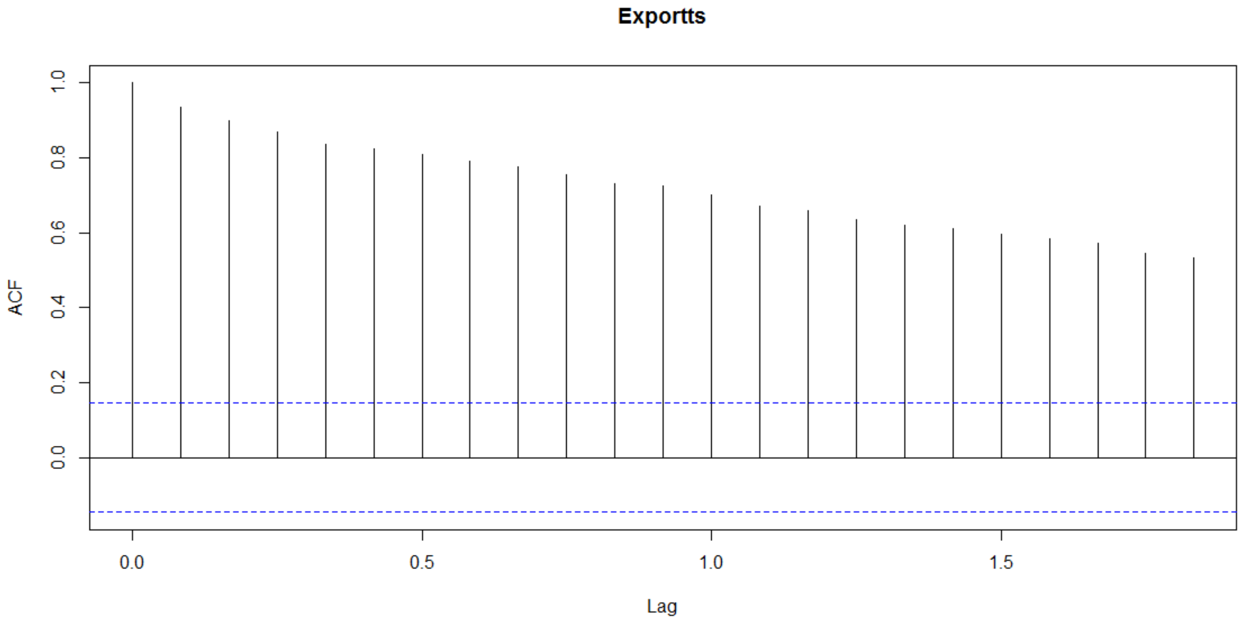
The dataset from 2000 till 2014 shows an increasing upward trend as the mean and variance also seems to be changing i.e. a function of time. Such a series is known as a non-stationary time series. Our time series also shows some spikes and troughs but it is not clear from data visualization whether they exhibit any seasonality or not. The data was again checked for stationarity using the ACF (auto correlation function) and PACF (partial auto correlation function) plots and the Dickey-Fuller test of stationarity.

summary(exportts)

```
Min.   :1261
1st Qu.:1534
Median :1858
Mean   :1908
3rd Qu.:2175
Max.   :3339
```

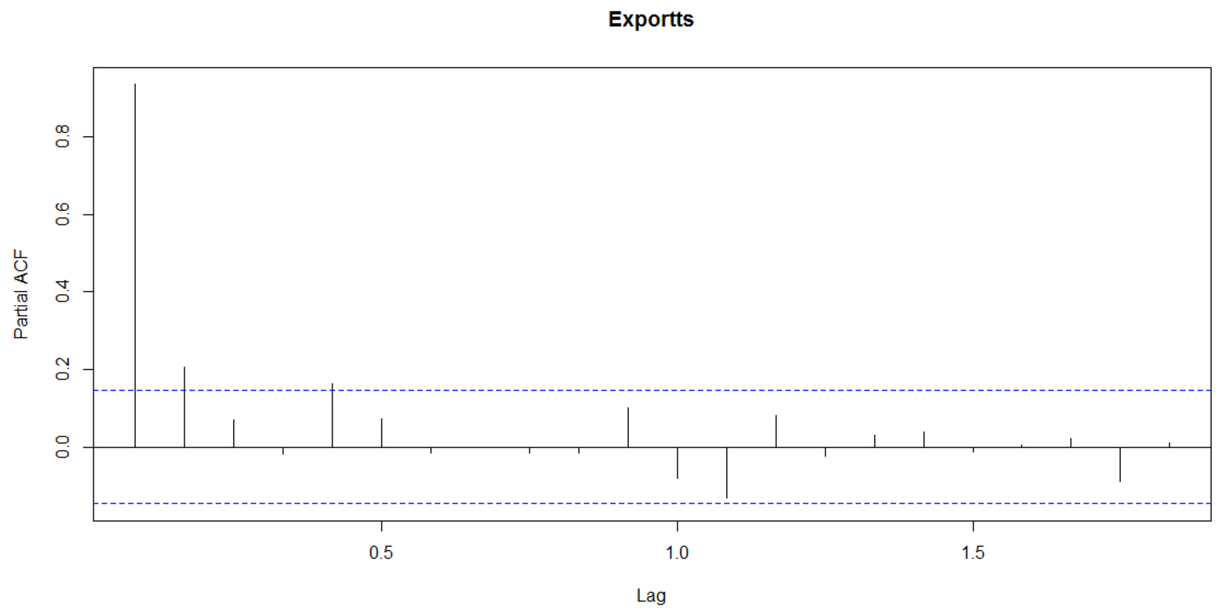
The summary of the timeseries shows that it is non-stationary and has a non-zero mean of a large value (1908). In order to build time series models or execute forecasting, the data has to be stationary.

acf(exportts)



The ACF plot indicates non-stationarity of our data as at each lag, as the spikes are well above the confidence bands (blue dashed lines) and decays very slowly with increasing lag which imply that there are significant correlations within the data at each lag. The lag spikes of a stationary data decays to zero at quite faster with increasing lag.

pacf(exportts)



In the PACF plot, we observe one very strong lag and the PACF cuts off after the 2nd lag.

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

```
        Augmented Dickey-Fuller Test

data:  exportts
Dickey-Fuller = -4.5715, Lag order = 0, p-value
= 0.01
alternative hypothesis: stationary
```

adf.test(exportts, alternative = "explosive", k = 0)

```
        Augmented Dickey-Fuller Test

data:  exportts
Dickey-Fuller = -4.5715, Lag order = 0, p-value
= 0.99
alternative hypothesis: explosive
```

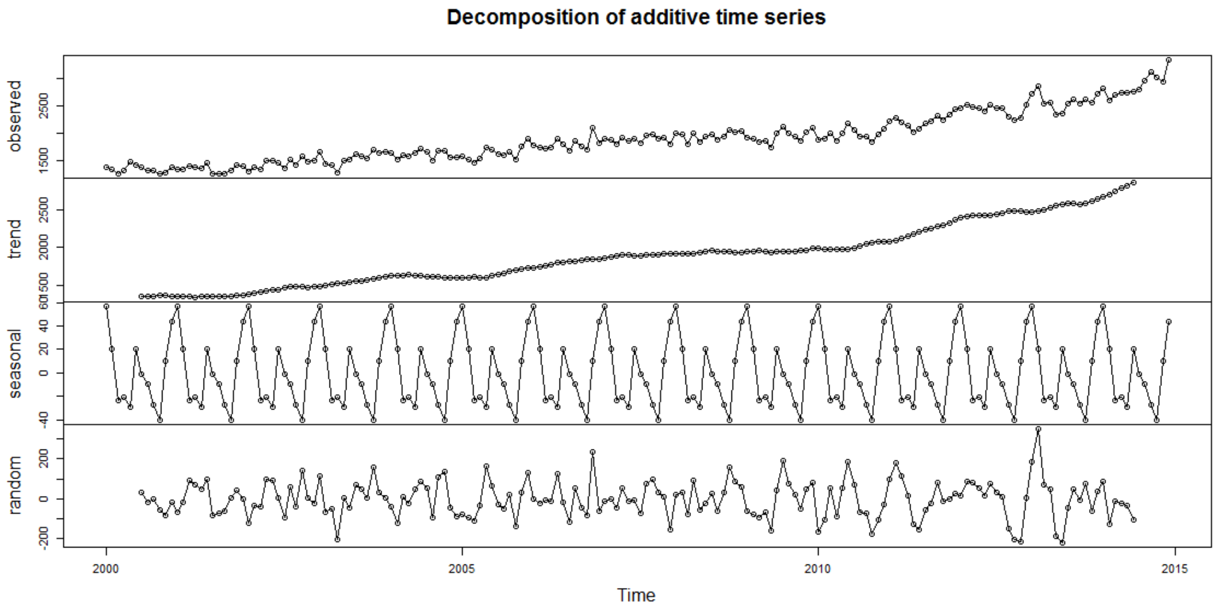
adf.test(exportts)

The augmented Dickey-Fuller test shows that based on p-value (< 0.05), we cannot reject our null hypothesis of a unit root and accept that our data is non-stationary and based on another p value (>> 0.05) we accept the alternative hypothesis that our data is explosive.

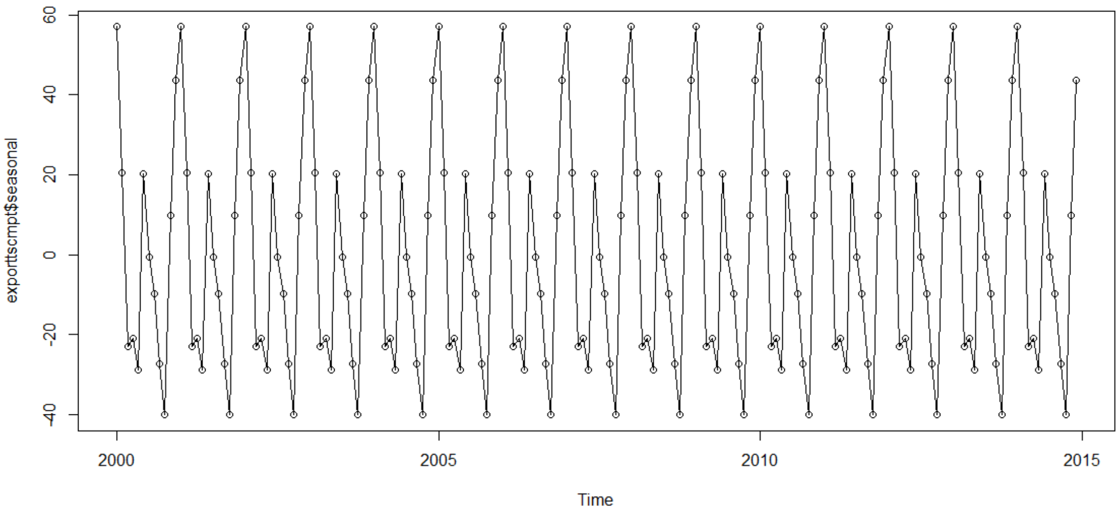
Our data was next subdivided into its various components using the decompose() function to check for the seasonality and random factor present in our time series.

```
exporttscmpt <- decompose(exportts)
```

```
plot(exporttscmpt, type = “o”)
```



```
plot(exporttscmpt$seasonal)
```



Our time series data also shows a seasonal component of increased export during January while a decline in export volume in each year October. This time series data will be converted to a stationary dataset and forecasting will be done using ARIMA method.