Time Series Analysis & Forecasting Brent Oil Prices in the United States

A project work completed as a part of the course MTH517A

Submitted by

Anis Pakrashi (211264)

Krishnendu Paul (211322)

Rahul Ghosh Dastidar (211353)

Souraj Mazumdar (211393)

Under the supervision of

Dr. Amit Mitra

Professor

Department of Mathematics and Statistics

Indian Institute of Technology, Kanpur

Acknowledgement

We present our project report on "**Time Series Analysis & Forecasting Brent Oil Prices in the United States.**" However, as mere students of statistics, any achievement, in terms of a substantial project, would have been an uphill task, considering just our own efforts. It is apparent that there has always been a constant encouragement of noble minds for our efforts to bring this project to a successful completion. The satisfaction that accompanies the effectiveness of any task would be incomplete without mentioning those who made it possible for us to complete our desired project.

First and foremost, we want to express our sincere gratitude to our instructor **Dr. Amit Mitra** for his constant help, support and advice throughout the project preparation, which acted as a catalyst for effective completion of our work. Without his valuable guidance and motivation, it would have been nearly impossible to work as a team and imbibe the practical aspects of the course "**MTH517A: Time Series Analysis**". Besides, we are thankful to all faculty members of our department and our seniors, because without their support at various stages, this project would not have materialized.

We are also thankful to our friends for their critical appreciation and their questions and counter-questions solved a huge spectrum of doubts related to our work. Last but not the least, we are grateful to our parents for their constant motivation on the way of materializing the project.

Anis Pakrashi

Krishnendu Paul

Rahul Ghosh Dastidar

Souraj Mazumdar

Abstract

Time Series analysis and the financial market go hand-in-hand. We have considered a dataset on prices of Brent oil over several years (1987 - 2021). The deterministic components were eliminated from the training set. After trend elimination by differencing, the series turned out to be random. Post randomness check, we included test for stationarity with a success. We fitted an ARMA model to the training set. Hence, in reality, we fitted an ARIMA model and used it to predict the oil prices for the test set. The accuracy measures were noted and hence the success of the model was justified.

Contents

1 Introduction

Oil, or petroleum, often referred to as "black gold" is a natural treasure that the earth bestows on us. We require oil in almost every sphere of life. The food that we eat to the vehicles that we avail, the products we use to the machines that develop such products, all are dependent heavily on the petroleum industry. This is why a rise or fall in the price of oil is a matter of concern to economists and financial experts across the globe. An excess or a deficit of oil supply can decrease sale of automobiles, reduce the supply of raw food, shut down markets or can even start a war!!

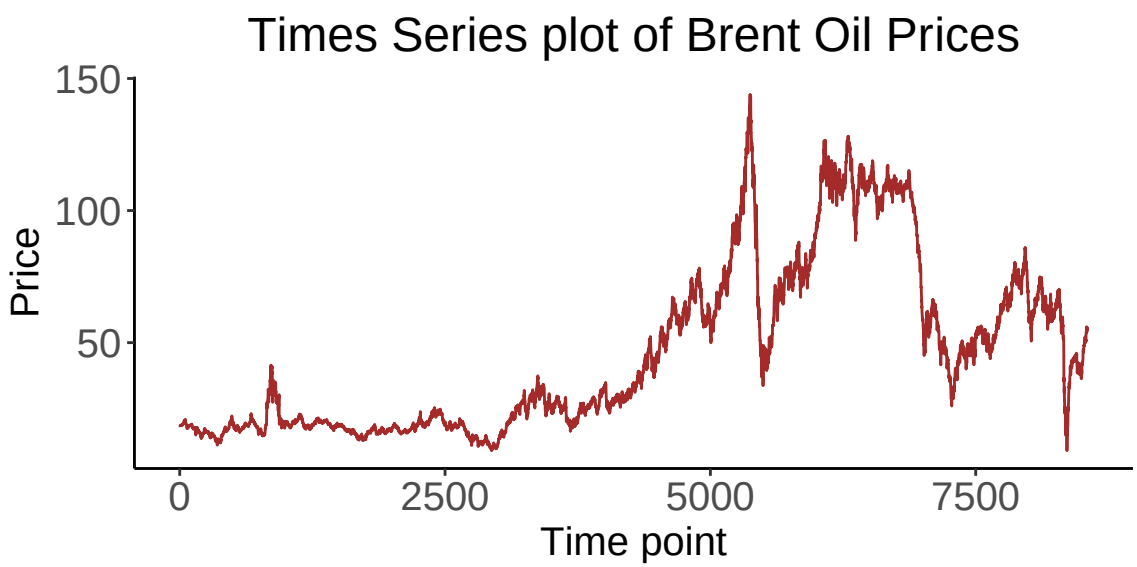
In our project, we have consided price of Brent oil over several years. *Brent* is the name used for a comparatively light crude oil obtained from a blend of different crude oils from 19 different oil fields in the North Sea. *Brent Crude* is one of the three main benchmarks for crude oil prices per barrel, along with *West Texas Intermediate* (WTI) from North America and *Dubai Crude* from the Persian Gulf. Brent is sometimes also referred to as the name of an oil field located in the North Sea off the coast of Scotland, discovered in 1971, which started production in 1976. Brent is an acronym for Broom, Rannoch, Etive, Ness and Tarbert – the five geological formations that form the Middle Jurassic field.

The data comprises values of prices of Brent oil, on weekdays, from May 20,1987 to January 25, 2021.

```
#Viewing the dataset
dim(data)

[1] 8554    3

head(data)

  Serial      Date Price
1      1 20-May-87 18.63
2      2 21-May-87 18.45
3      3 22-May-87 18.55
4      4 25-May-87 18.60
5      5 26-May-87 18.63
6      6 27-May-87 18.60

str(data)

'data.frame': 8554 obs. of  3 variables:
 $ Serial: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Date  : chr  "20-May-87" "21-May-87" "22-May-87" "25-May-87" ...
 $ Price : num  18.6 18.4 18.6 18.6 18.6 ...
```

```
#Checking for presence of missing value
colSums(is.na(data))

Serial   Date  Price
     0      0      0
```

Consider the plot of the entire data in Figure below:



1.1 Objectives

The main aim of our project is to analyse the above data. We aim to check for the deterministic components, eliminate them and check for the randomness of the data. Then, we check for stationarity, and if stationary, we fit an appropriate model based on accuracy measures. We then use this model for forecasting purposes.

2 Train-Test Split

Our data comprises 8554 values. We split the data into training and testing sets. Since the data under study is a time series, we do not use a random sample from the whole dataset as our train data. Instead, we consider the first 98% values as our training set and the remaining 2% as our test set. Thus, our training data has the first 8404 values while the test set has the final 150 values. We proceed with model building using the training set and use the test set for comparison of the predicted values.

```
#Splitting of main data
t1=data$Price[1:8404]
head(t1)
```

```
[1] 18.63 18.45 18.55 18.60 18.63 18.60

t2=data$Price[8405:8554]
head(t2)

[1] 41.18 40.97 41.58 41.64 42.18 43.19
```

3 Deterministic Aspects of Time Series

3.1 Initial Theory -

A time series may be roughly composed of deterministic and random components. The deterministic components are linked with some sort of pattern in the data. Hence they must be eliminated in order that we are able to fit some model to the data to make the data usable for future analyses. Such deterministic components include:-

- Trend - it characterizes a long term movement/ pattern in the time series
- Seasonal component - it is a distinguishable pattern of regular annual variation in the time series
- Cyclical component - it comprises regular long range swings above or below some equilibrium level or trend line, moving through the stages of upswing, peak, downswing and trough

Time series may one or a combination of the above deterministic factors. We begin with the test for existence of trend.

3.2 Testing for the existence of trend -

We test for the existence of trend using a non-parametric testing procedure called the **Relative Ordering Test.** The test is given as follows :

H_0 : no trend against H_1 :trend is present

Let the time series be denoted by $Y_1, Y_2, ..., Y_N$.

Define

$$q_{ij} = \begin{cases} 1 & if \quad Y_i > Y_j \quad when \quad i < j \\ 0 & o.w. \end{cases}$$

Define

$$Q = \sum_i \sum_j q_{ij}$$

where the sum is taken over all i, j such i < j.

Note that Q counts the number of decreasing points in the time series.
Under H_0,

$$P\left(q_{ij}=0\right)=P\left(q_{ij}=1\right)=\frac{1}{2}$$

$$\Rightarrow E\left(Q\right)=\frac{n(n-1)}{4}$$

If observed Q doesnot differ significantly from E(Q), it signifies no trend.

3.2.1 Results

```
#Test for existence for trend : Relative Ordering Test
trendtest <- function (y){
n=length(y)
Q=0
for(i in 1:(n-1)){
    for(j in (i+1):n){
                if(y[i]>y[j]){
                        Q=Q+1
                }
        }
}
E_Q=n*(n-1)/4
T=1-(4*Q/(n*(n-1)))
var_T=(2*(2*n+5))/(9*n*(n-1))
z=T/sqrt(var_T)
cat("The value of the test statistic is = ",z,"\n")
if(abs(z)>qnorm(0.975)){
        cat("Reject Null Hypothesis, i.e. Trend exists")
} else
{     cat("Null Hypothesis is not rejected i.e. No Trend in data")
} }
trendtest(t1)

The value of the test statistic is =  79.39012
Reject Null Hypothesis, i.e. Trend exists
```

We observe that there is presence of trend in our data. We need to eliminate the trend to proceed with further analyses.

3.3 Elimination of Trend -

We employ the **Method of Differencing** to eliminate the trend component from our data.

Define lag operator B such that

$$BY_t = Y_{t-1}$$

Hence,

$$B^j Y_t = Y_{t-j}; j = 1, 2, 3, ...$$

The first difference operator is given by ∇ such that

$$\nabla \equiv 1 - B$$

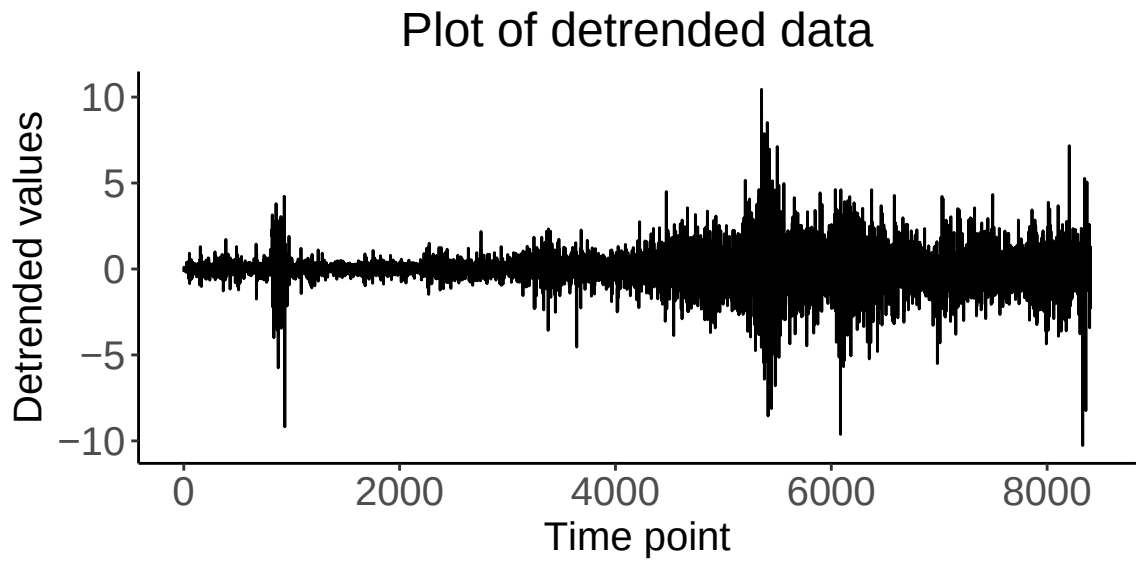
Hence,

$$\nabla^j Y_t = (1 - B)^j Y_t; j = 1, 2, 3, ...$$

Suppose we have a linear time trend $m_t = a + bt$. Then, $\nabla m_t = b$.

3.3.1 Results

```
#First order backward difference function
order1_diff<- function(x){
        k=1
        y=array(0)
        for(i in 2:length(x)){
        y[k]=x[i] - x[i-1]
        k=k+1}
        return(y)
}
#Applying first order differencing in our data
detrended<-order1_diff(t1)
head(detrended)

[1] -0.18  0.10  0.05  0.03 -0.03  0.00
```



We now confirm our detrendation with the Relative ordering test on the detrended data.

```
The value of the test statistic is =  1.591205
Null Hypothesis is not rejected i.e. No Trend in data
```

Now, we are sure that the trend component has been removed. We must also check for the presence of other deterministic components in the data. However, prior to this, we check for randomness. If the series turns out to be random, we need not check for the other components as the series is already free of any such components.

4 Test for Randomness of a time series

To test whether a series is random, we make use of the **Turning Point Test**. The test is given as follows :

H_0 : series is purely random against H_1 : series is not purely random

A turning point is either a *peak* when a value is greater than its two neighboring values or a *trough* when a value is less than its two neighboring values.

Let the time series be denoted by $Y_1, Y_2, ..., Y_N$.

Define

$$U_i = \begin{cases} 1 & if \quad Y_i \quad is \quad a \quad turning \quad point \\ 0 & o.w. \end{cases}$$

Define the total number of turning points as

$$P = \sum_{i=2}^{n-1} U_i$$

Under H_0,

$$E\left(P\right) = \frac{2}{3}(n-2)$$

$$Var\left(P\right) = \frac{16n-29}{90}$$

The asymptotic test is based on

$$Z = \frac{P - E(P)}{\sqrt{Var\left(P\right)}}$$

,which is asymptotically normally distributed with mean as 0 and variance as 1. We reject null hypothesis at level 0.05 if observed |Z| $> \tau_{0.025}$.

4.1 Results

```
 The observed value of test statistic is =  -1.046206
Null Hypothesis is not rejected i.e. Series is purely random
```

This implies that the detrended series is itself **random**, free from any other deterministic components. Hence, we may procees with further study without testing for presence of other components.

5 Test for Stationarity of a time series

To test for stationarity in time series, we implement the **Augmented Dickey-Fuller Test**. Consider the AR(1) model for sake of illustration.

$$X_t = \phi X_{t-1} + \epsilon_t$$

The test is as follows:
$H_0 : \phi = 1(\text{series is not stationary})$ against $H_1 : \phi < 1(\text{series is stationary})$
It is easy to see that

$$X_t - X_{t-1} = \phi_0 + \left(\phi_1 - 1\right) X_{t-1} + \epsilon_t$$

$$\nabla X_t = \phi_0 + \phi_1^* X_{t-1} + \epsilon_t$$

Here we compare the value of the test statistic with the value of Dickey-Fuller distribution. If the value of the test statistic is less than the value of the Dickey-Fuller distribution, we reject the null hypothesis.

5.1 Results

```
Warning in adf.test(detrended):  p-value smaller than printed p-value
 The observed value of test statistic is =  -18.54358
 The p value is(approx) =  0.01
Reject Null Hypothesis, i.e. Series is stationary
```

The time series is **stationary**, as the p-value is less than the level of significance, 0.05.

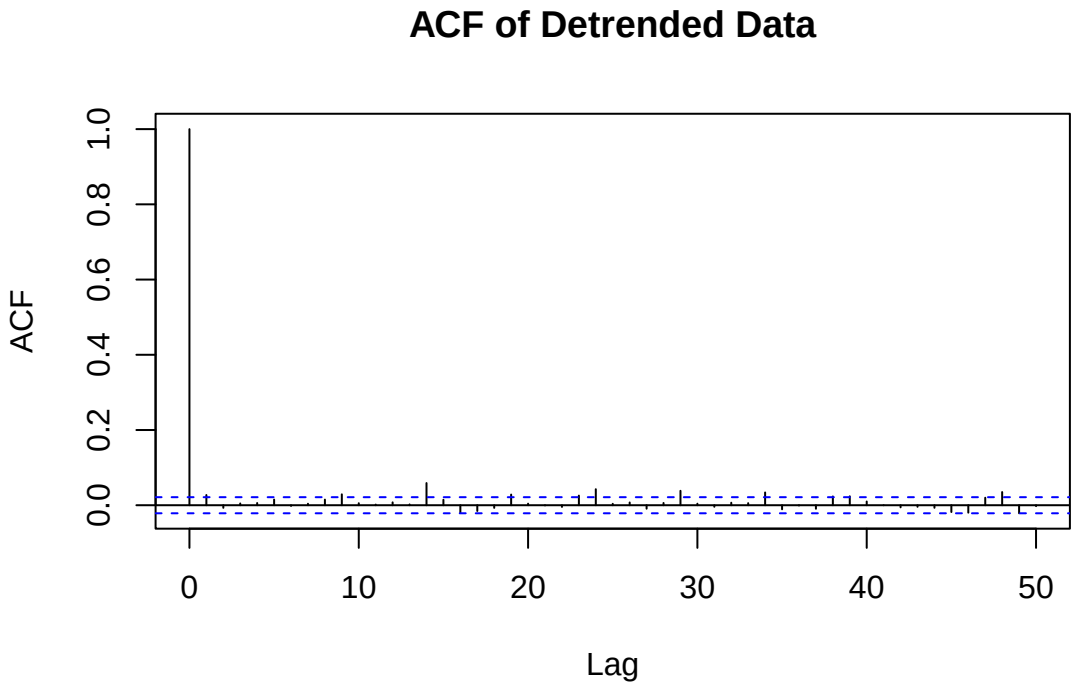
6 Model Fitting

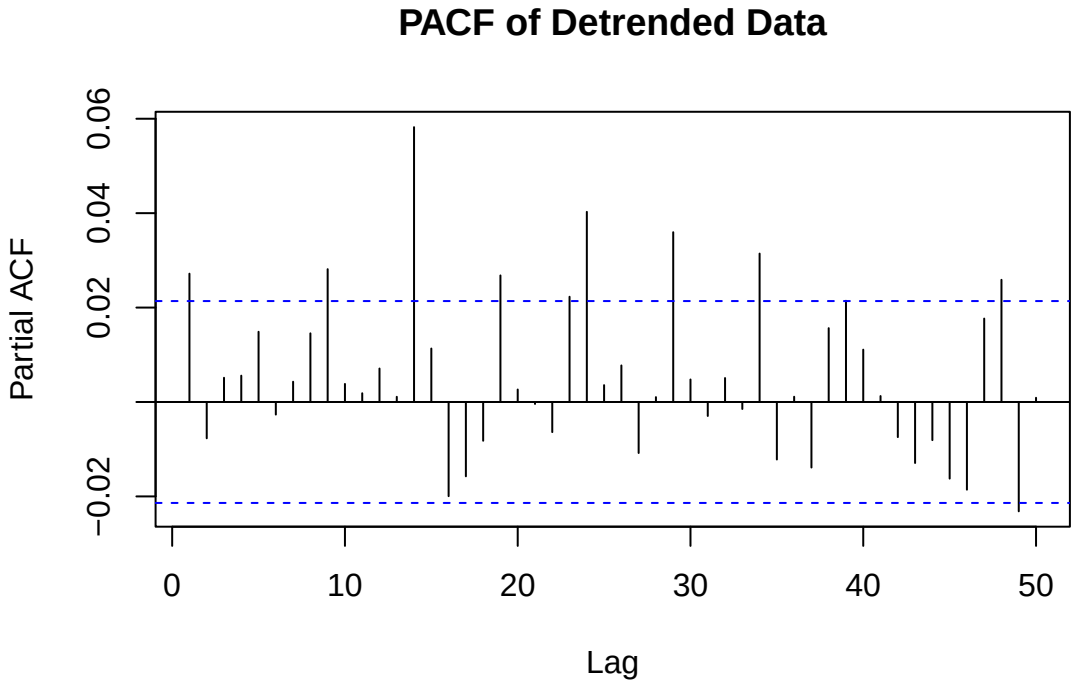
6.1 ACF and PACF plots

Autocorrelation and partial autocorrelation plots are used to graphically summarize the strength of a relationship between an observation of a time series and observations at prior time steps. Plots of autocorrelation function (ACF) and partial autocorrelation function (PACF) give us different viewpoints of time series.

A partial autocorrelation plot is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed i.e. PACF only describes the direct relationship between an observation and its lag. This would suggest that there would be no correlation for lag values beyond k. While ACF describes the autocorrelation between an observation and another observation at a prior time step that includes direct and indirect dependence information.

We use the ACF and PACF plots to determine the order of AR and MA. For both the plots, we take that value of the lag, after which the values of ACF and PACF are not significantly different from 0. From the ACF plot, we get we get the order of MA(q) i.e., the value of 'q' and from PACF plot, we get the order of AR(p), i.e., the value of 'p'.





We can see that, in both the cases, the values of ACF and PACF can be considered indifferent from 0, even for small values of lag.

6.2 Model Order Estimation and model fitting

We consider maximum 15 lags for 'p' and maximum 5 lags for 'q'. From these 96 models, we will choose that model, for which we get the minimum value of **AIC** (**Akaike Information Criteria**). The results, that we have obtained, shows that, AIC attains the minimum value

for **ARMA(6,5) model**. So, we choose the value of 'p' and 'q' as 6 and 5 respectively. Now, recall that we have obtained the detrended series by considering first order difference. Hence, the true model to be fitted to our data is **ARIMA(6,1,5).**

Consider the results below:

```
model_order

[1] 6 1 5

aic

[1] 25402.17

final_fitted_model=arima(t1, order = model_order ,
        optim.control = list(maxit = 1000))
final_fitted_model$coef

        ar1         ar2         ar3         ar4         ar5         ar6
-0.21748539  0.10895412 -0.15999109  0.19872379  0.95964480 -0.02428496
        ma1         ma2         ma3         ma4         ma5
 0.24607286 -0.09601397  0.16067733 -0.18125276 -0.94395345
```

7 Forecasting of future values

We use the above model to forecast future values. We predict the values of the time series for as many number of time points ahead as is equal to the number of points in the test dataset. Here, we predict the values for 150 observations forward.

```
Time Series:
Start = 8405
End = 8554
Frequency = 1
  [1] 40.31625 40.46587 40.44448 40.42573 40.55965 40.53761 40.70136 40.61403
  [9] 40.66353 40.74165 40.75216 40.89079 40.77134 40.87589 40.89381 40.95615
 [17] 41.03686 40.92601 41.05573 41.02958 41.14257 41.14830 41.08099 41.20015
 [25] 41.16019 41.30284 41.23778 41.23281 41.31203 41.29177 41.43053 41.31859
 [33] 41.37385 41.39955 41.42410 41.52413 41.40112 41.49570 41.47393 41.55147
 [41] 41.58805 41.49035 41.59260 41.54608 41.66517 41.63165 41.58503 41.66372
 [49] 41.62343 41.75698 41.66663 41.67899 41.71373 41.70794 41.82220 41.70367
 [57] 41.76378 41.75140 41.79591 41.86149 41.74949 41.83197 41.78684 41.87962
 [65] 41.88092 41.80528 41.87994 41.82837 41.95021 41.89011 41.86693 41.90916
```

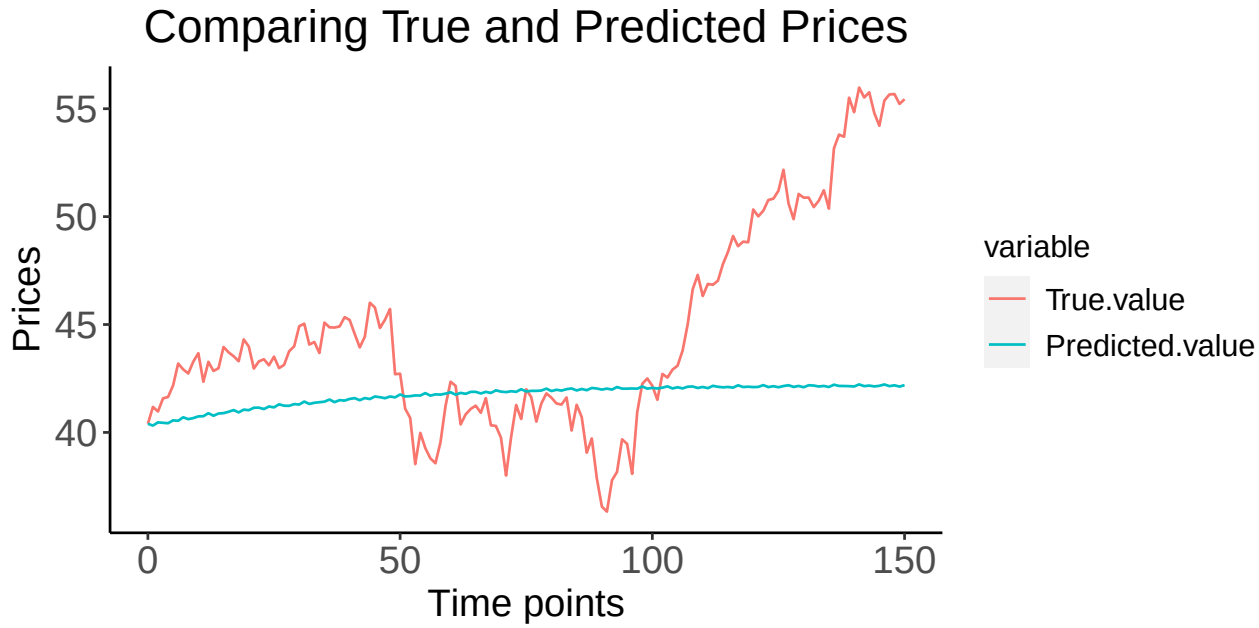
```
 [73] 41.87998 42.00087 41.89940 41.92691 41.92583 41.94029 42.02918 41.91676
 [81] 41.97722 41.93877 42.00327 42.03799 41.94559 42.01227 41.95659 42.06048
 [89] 42.03455 41.98415 42.03083 41.98493 42.10396 42.02817 42.02651 42.03647
 [97] 42.02472 42.12891 42.02743 42.06495 42.03633 42.07207 42.13527 42.03754
[105] 42.09275 42.03871 42.11966 42.12770 42.05907 42.10653 42.05033 42.15930
[113] 42.11399 42.08809 42.10740 42.07420 42.18460 42.10261 42.11786 42.10051
[121] 42.10867 42.19302 42.10001 42.14130 42.09329 42.14813 42.18663 42.10881
[129] 42.15350 42.09296 42.18490 42.17131 42.12723 42.15338 42.10415 42.21175
[137] 42.15477 42.14999 42.14398 42.12739 42.22424 42.14409 42.17033 42.13145
[145] 42.15906 42.22203 42.14351 42.18243 42.12287 42.19258
```

We compare these predicted values with the actual values for the time points, which are nothing but the true data values in the test dataset. We use the **Mean Absolute Percentage Error (MAPE)** as a measure of comparison.

```
#MAPE
mape = mean((abs(c$pred - t2)/t2))*100
mape

[1] 8.590808
```

We obtain an MAPE value of **8.590808**, which is quite low. Hence we may consider our prediction to be quite satisfactory.

Consider the plot below:



We have plotted the predicted portion along with corresponding the true values for the final 150 values (which were considered in test data). In the above plot, '0' represents the last time

point in our training set. We observe that the prediction is not that satisfactory even if the MAPE for the model was low.

8 Conclusion

The time series anaysis on Brent Oil Prices in the US reveals that the data can be well modelled using an **ARIMA(6,1,5).** The forecasted values, when compared with the test data, yield an MAPE of 8.590808, which is low. But the forecast doesnot seem to be a good one. However, we may improvise upon this model by considering other models, which capture the variations in the series more accurately. However, we have not explored those cases here, in our project.

References

[2013] P.J.Brockwell, R.A.Davies, *Introduction to Time Series and Forecasting,* Springer

[1995] Wayne A. Fuller, *Introduction to Statistical Time Series,* Wiley

[2019] P.J.Brockwell,R.A.Davies, *Time Series: Theory & Methods,* Springer

[2020] J.D.Hamilton, *Time Series Analysis,* Princeton University Press