ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 16, Issue, 01, pp.26822-26830, January, 2024
DOI: https://doi.org/10.24941/ijcr.46491.01.2024

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

A COMPREHENSIVE ANALYSIS OF PRICE CHANGES IN CRUDE OIL DATA INVOLVING TIME SERIES MODELLING

Joseph Justin Rebello[1] and Mary Mrudhula, P.L.

AC, Mahatma Gandhi University, Kottayam, Kerala, India

ARTICLE INFO

Article History:
Received 25th October, 2023
Received in revised form
27th November, 2023
Accepted 15th December, 2023
Published online 19th January, 2024

Key words:
Time series, ARIMA model, SARIMA model, ARCH model, GARCH model, Time series forecasting, Data driven approach.

***Corresponding author:**
Joseph Justin Rebello

ABSTRACT

This research model explored different time series modelling approach over Crude oil prices. In time series analysis, we assume that the current price of crude oil reflects the effect of all the influencing factors. So that the price forecasting of the crude oil can be done using the past crude oil prices. The main assumption in this time series modeling is that the past crude oil prices can be used to predict the future crude oil price. Although the time series analysis can find the trend, there will be limitations to the forecasting capability of the model that we use in the analysis when the reversal in trend is observed in the data taken or the pattern repeated may not be followed by the future prices. Different types of trend patterns such as increasing trend, decreasing trend or periodic patterns can be obtained. Time series analysis is more useful and will give better forecasting only when the data follows any of these trends. In this work, data analysis on crude oil data set is performed. A novel time-series forecasting approach based on Auto-Regressive Integrated Moving Average (ARIMA) model, Seasonal Auto-Regressive Moving Average (SARIMA), ARCH (Auto Regressive Conditional Heteroscedasticity) model, GARCH (Generalized Auto Regressive Conditional Heteroscedasticity) are also proposed using the R programming language for statistical computing and graphics. The results will help the researchers from various community to gauge the trend and improvise containment strategies accordingly.

Citation: ***Joseph Justin Rebello and Mary Mrudhula, P.L.*** **2024.** "A Comprehensive Analysis of Price Changes in Crude Oil Data Involving Time Series Modelling*". International Journal of Current Research*, 16, (01), 26822-26830.

INTRODUCTION

Various studies were conducted on this topic previously by many scholars. From the study of Sakinat Olurabukanla Folorunso, Joseph Bamidele Awotunde, Oluwatobi Oluwaseyi Banjo, Ezekiel Adebayo Ogundepo & Nureni Olawale Adeboye (2022) it was observed that a 14- multi step ahead forecast system for active coronavirus cases are built, analyzed and compared using ARIMA and Exponential Smoothing. In another study of Sai Manoj Cheruvu (2021) stock price has been predicted using Exponential Smoothing and it outperformed other models like ARIMA. Apart from this, the study conducted by Bilgi Yilmaz & A.Sevtap Selcuk Kestel (2020) estimate and forecast Turkey's house price using GLM, multivariate (VAR) and univariate time series models and exponential smoothing approaches. Supplementary to these studies Christian Brownlees & Matteo Barigozzi (2019) proposes novel network analysis techniques for multivariate time series. This approach is based on a VAR approximation of the process. Another study conducted by Anil Namdeo, Chuleekorn Tanathitikorn, Nikki Rousseau & Richard McNally (2018) use time series to forecast the climatic change as well as created a system for health warning. Also, Balaji Prabhu B V& M Dakshayini (2018) in Predictive analysis of the regression and time series predictive models using parallel implementation for agricultural data compares the forecasted values of various models with the actual market values and analyzed the performance to select the best model. According to Atul Wadagale V, Jagganath Dixit V, Varsharani Vithalrao Kendre & Vaishali Bahattare N (2017) Forecasting for Measles vaccine was done up to 2018-19 and it was obtained that simple seasonal time series analysis can be used to forecast Measles vaccine requirement. By the study conducted by Anna Klimovskaia, Manfred Claassen & Stefan Ganscha (2016) used time series to study the single cell structure function using Stochastic reaction networks and Sparse Regression. Apart from this Adam M. Sykulski, Sofia C. Olhede, Jonathan M. Lilly & Eric Danioux (2015) proposed time series models are used to summarize large multivariate datasets and infer important physical parameters of inertial oscillations and other ocean processes. Bhalla N and Rakesh S (2018) analyzed the crude oil data using ARIMA techniques. Along with this a Non stationary time series methods are employed to account for the spatiotemporal variability of each trajectory. Here we consider a comprehensive approach of Time series modelling with a special reference to Oil data.

Concepts and Methods

Time series modeling techniques: Several time series techniques are proposed to build a time series. They include AR (Auto Regressive) model, MA (Moving Average) model, ARMA (Auto Regressive Moving Average) model, ARCH (Auto Regressive Conditional Heteroscedasticity) model, GARCH (Generalized Auto Regressive Conditional Heteroscedasticity), ARIMA (Auto Regressive Integrated Moving Average) model, SARIMA (Seasonal Auto Regressive Integrated Moving Average), Holt- Winter's Exponential Smoothing. Various hybrid models are also suggested as a combination of two models with support vector regression, genetic algorithms and wavelets. We have used ARIMA and other methods for modeling the crude oil prices since these models covers both linear and non- linear time series modeling.

ARIMA Modelling: Most of the time series are non- stationary model so we make differencing of feasible order to get into a stationary model. Thus we have a model capable of describing certain types of non- stationary series, called an Integrated model. Thus, an ARIMA(p,d,q) process can be defined and it consist of 3 parts: AR component which is a linear combination of the previous values, MA component is the linear combination of past error terms and I (Integrated) replaces the original series with the differenced series. To build an ARIMA model, we use the ADF (Augmented Dickey- Fuller) test in order to test the stationarity of the given time series. The proper orders of p and q for the model can be obtained by plotting the SACF (Sample Auto- Correlation Function) and SPACF (Sample Partial Auto-Correlation Function). A time series plot of Price of the crude oil appears in Figure 1. Here we can see that the price rises and falls through the time and it is clear that data exhibit high level of volatility during the period of analysis.

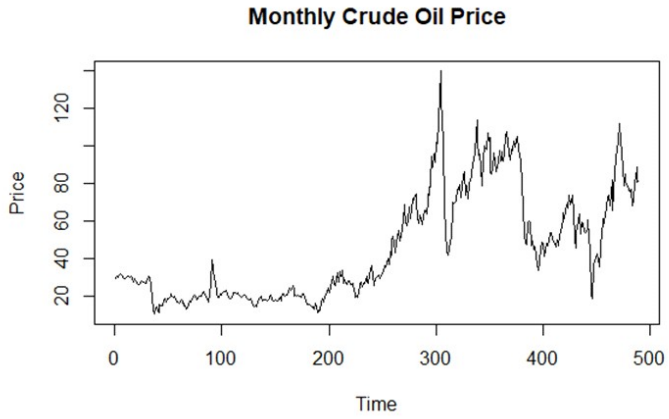


Figure 1. Time series plot of crude oil price from March 1983 to November 2023

The time series plot shows that there is a small increase in variance and an upward trend in the data. Also, it is clear that the prices arise and falls through time, so its mean may not be stationary.

Normality test- To check the normality of our data we use both graphically by constructing the Histogram & Q-Q plot and statistically by Shapiro- Wilk test.

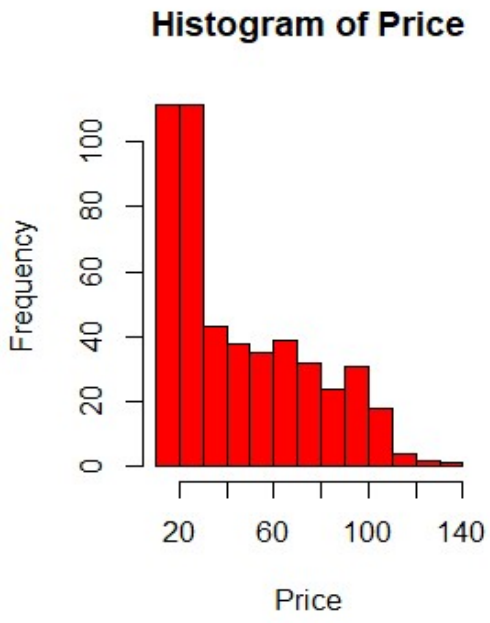


Figure 2. Histogram of crude oil price.

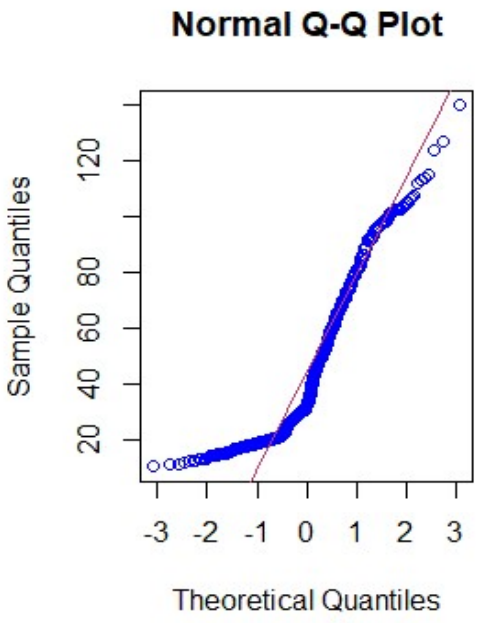


Figure 3. Q-Q plot of crude oil price.

From the figure 2 & 3 it is clear that our data is not normally distributed.

Shapiro- Wilk normality test

Data is normally distributed. Vs : Data is not normally distributed.

Shapiro-Wilk normality test

data: Price
W = 0.88391, p-value < 2.2e-16
The p value in the Shapiro- Wilk normality test is less than 0.05. Hence, we reject the null hypothesis. Hence, we can conclude that our data is not normal. Now we decompose the time series into its components to evaluate.

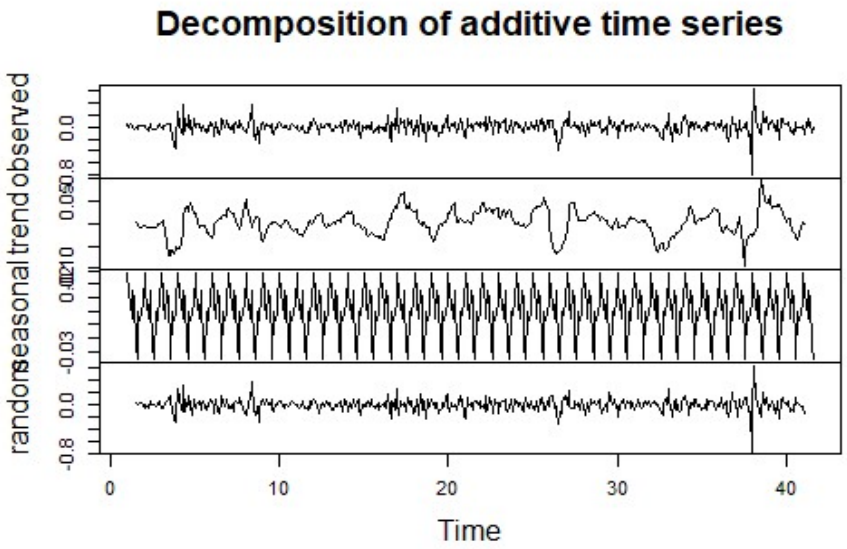


Figure 4. Decompose plot of additive time series.

This plot shows the decomposed components of the time series. It consists of three components; Trend, Seasonality and Residuals.

Stationarity: From our time series plot, we analyze that our data is non- stationary and it contains the trend and seasonal component. In order to make this assumption more specific we construct the ACF, PACF plot and Dickey- Fuller test.

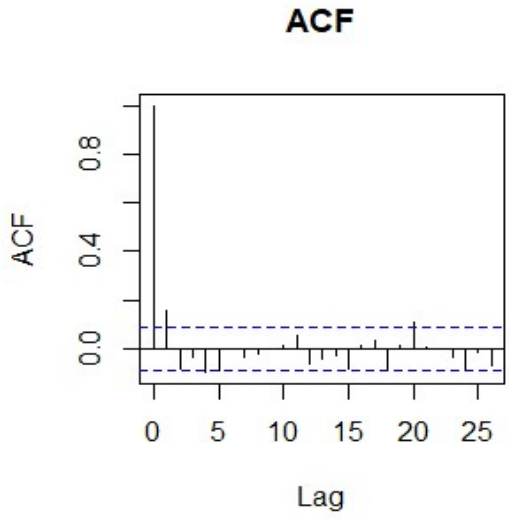


Figure 5. ACF of monthly crude oil price.

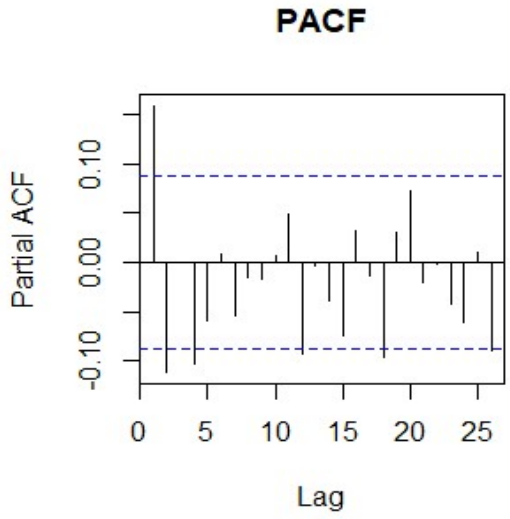


Figure 6. PACF of monthly crude oil price.

From the above two figures, it is clear that the data is not stationary. Similarly, the stationarity of the time series can be tested using Augmented Dickey- Fuller test. The hypothesis is as follows. The time series is non- stationary Vs. The time series is stationary.
Augmented Dickey-Fuller Test

data: Price

Dickey-Fuller = -2.8592, Lag order = 7, p-value = 0.2146 alternative hypothesis: stationary
It is clear that p value is greater than 0.05. Hence, we accept the null hypothesis. So that our time series data is non- stationary.

The plot of PACF shows a non- constant mean and variance. Using differencing, and because of non- constant variance, we work with the log returns of our data. The log returns approach is considered as

Where, represents the price of crude oil and is its differenced series.

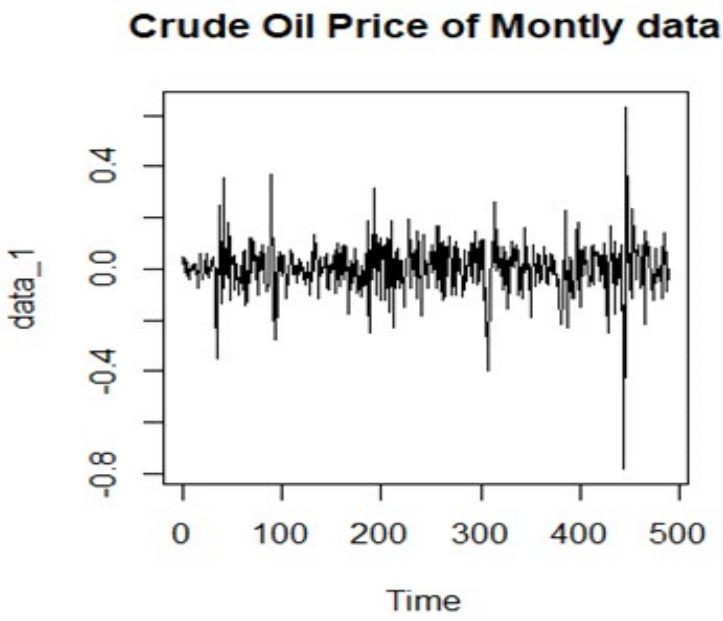


Figure 7. Time series plot of log differenced data.

From the time series plot of log differenced data, it is clear that after applying the transformation, the trend and seasonality is removed. Also, our data becomes stationary in mean. But our is not constant in variance. Hence, the further analysis can be done with the help of ARIMA, ARCH and GARCH. In order to substantiate that the transformed data is stationary, we use the unit root test which include Augmented Dickey- Fuller test (ADF), Phillips-Perron test and Kwiatkowski- Phillips- Schmidt- Shin (KPSS) test.

The time series is non- stationary and has a trend component which cannot be removed bydifferencing the data Vs. The time series is stationary.
Augmented Dickey-Fuller Test

Augmented Dickey-Fuller Test

data: data_1

Dickey-Fuller = -8.8557, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

From the ADF test, we can see that the p value is less than 0.05. Thus, we reject the null hypothesis. Hence, our time series with log difference is a stationary process.

Phillips-Perron Unit Root Test

data: data_1

Dickey-Fuller Z(alpha) = -361.51, Truncation lag parameter = 5, p-value = 0.01

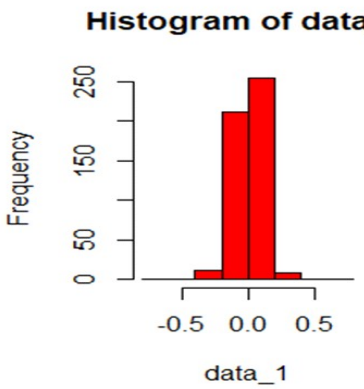
alternative hypothesis: stationary

Here, the Dickey- Fuller statistic has a negative value which shows strong evidence against our null hypothesis and suggest that our log differenced time series is stationary. The truncated lag parameter of 5 indicates that PP test consider up to lag 5 of the Auto Regressive parameter in estimating our model. Also, it shows there is no serial correlation and heteroskedasticity in our model and since the p- value is 0.01 which indicates that our time series is stationary.

KPSS Test for Level Stationarity data: data_1

KPSS Level = 0.050875, Truncation lag parameter = 5, p-value = 0.1

In KPSS (Kwiatkowski- Phillips- Schmidt- Shin), the p- value is greater than 0.05. Hence, we accept the null hypothesis that our data is stationary.



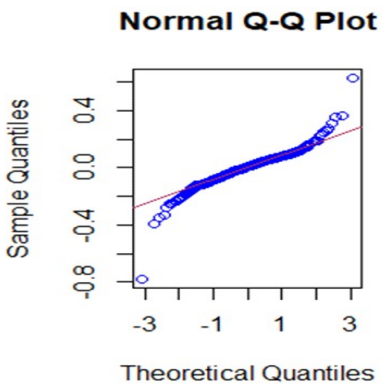


Figure 8. Histogram of data_1

Figure 9. QQ plot of data_1

Since, after transforming our data by applying the log differencing, the data becomes stationary. Now, we will check the normality of transformed data using Histogram and Q-Q Plot.

Histogram and Normal Q-Q plot shows that the data is normally distributed. Data testing and training are of paramount in time series analysis for several reasons. Testing and training data allows us to assess the performance of our time series model. By comparing model predictions with the actual values in testing dataset, we can measure how well our model generalizes to unseen data. This evaluation is crucial for understanding the model effectiveness in making the future forecast. The time series model needs to capture underlying patterns and trends in the data. Training data helps the model to learn these patterns, while testing data assess its ability to generalize beyond the training set. Ensuring that a model can generalize well is essential for making accurate forecast. During model training, we may need to adjust the hyper parameters such as the lag order in ARIMA. Testing data are crucial for assessing the impact of different hyper parameter choices and selecting the best configuration for optimal forecasting performance. Dividing a dataset into training and testing sets is a fundamental step in machine learning and time series analysis. The goal is to separate the data into two distinct subsets: one for training the model and another for evaluating the performance. To maintain the temporal order and ensure that the model learns from past data to predict future data, we typically use the sequential splitting approach. The dataset is divided into two parts: the training set and testing set. The training set contains the earliest observations in the time series data, covering a substantial portion of the historical data. For example, we might use the first 80% of the data as the training set. The testing set contains the most recent observations representing a portion of the data that the model has to be seen during training. This portion is around 20% of the data but can vary depending on the specific requirements. It's crucial to consider the temporal aspects when splitting the data. The training data should precede the testing data in terms of time. The training data should provide the model with a sufficient historical context to learn from past patterns and trends. The testing data should represent a future time period that the model needs to forecast accurately. Dividing a time series dataset into training and testing sets involves careful considerations of the temporal order of the data. The goal is to provide a model with historical context while ensuring that it can make accurate predictions for future time points.

Since, our data is normally distributed and stationary, we can fit a model.

Box cox lambda value1.275122

While applying the Box- Cox transformation to find the optimal value of lambda, we got the value of lambda as greater than zero so that our transformed time series data can be used to build a more accurate model for our underlying process like ARIMA.

arima(x = data_1.train, order = c(1, 1, 2))

Coefficients:

	ar1	ma1	ma2
	-0.0814	-0.7328	-0.2672
s.e.	0.2420	0.2334	0.2333

sigma^2 estimated as 0.008663: log likelihood = 406.98, aic = -805.97

It is clear that the best fitted model is ARIMA(1,1,2) which means that our model has the lag 1 of AR , difference d=1 and MA has residual lag

In fitting ARIMA(p,d,q) model, Thus, our model becomes:

Where,represents the value of time series at time represents the value of time series at previous time step , is the error term which represents the difference between the observed value at time t and the predicted value at time t based on our ARIMA model, is the error term at the previous time step is the error term at the previous time step The model equation is indicating that the value of the time series at time depends on its value at the previous time step and the error terms at the current time step and the previous two time steps and The coefficients on the error terms (-0.7328 and -0.2672) indicate that the current value of the error term is negatively correlated with the error terms at the previous time steps, with the strength of the correlation decreasing as the time lag increases. The coefficient on the lagged value of the time series (-0.0814) indicates a positive correlation between the value of the time series at time t and its value at the previous time step, with a strength of -0.0814.

The values of AIC, AICc, BIC are -1.902452, -1.902234, -1.855199 respectively. These less values shows that our model is a good fit for this data.

Diagnostic Checking: In order to check whether our ARIMA model is a best fit to the data, we should do the diagnostic checking. ie, we have to check the residuals which is very helpful in checking whether the model has adequately captured the complete information from the data.

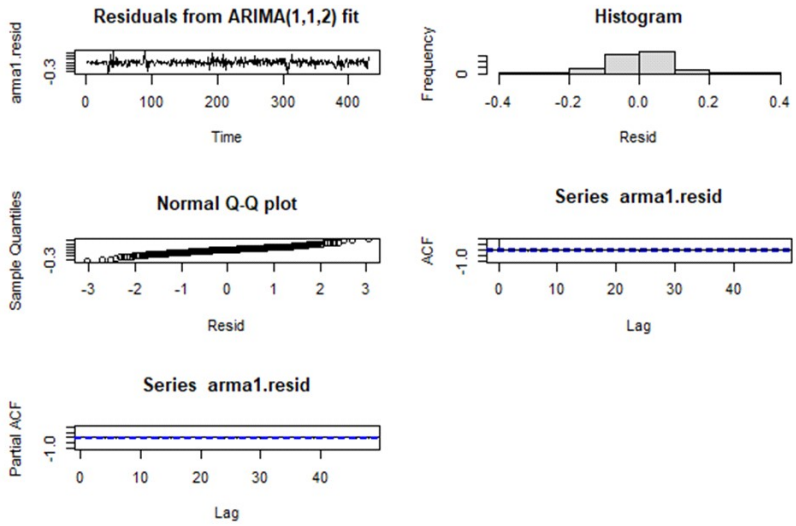


Figure 10. The Residual Plot, Histogram, Q-Q Plot, ACF and PACF of log differenced data.

Figure 10: The Residual Plot, Histogram, Q-Q Plot, ACF and PACF of log differenced data. From the graph of Residuals it is clear that all the points do not cross the control limit lines. Hence, there is no autocorrelation is present in between the residuals. Also, mean is constant for residuals at zero. But there exists volatility in the ARIMA model. The ACF and PACF plot shows the correlation of residuals with the previous time points and it is not significant. To make it significant by plotting ACF and PACF of Residual square.

Shapiro-Wilk normality test
data: arma1.resid
W = 0.98193, p-value = 3.365e-05

The Q-Q plot of residuals shows that residuals are normally distributed. But in the Shapiro- Wilk test the p value is 3.365e-05 which is less than 0.05. It shows that the residuals are not normally distributed. Hence, we check the model adequacy by Ljung Box Portmanteau test as follows.
: The model does not show lack of fit.
: The model shows lack of fit.

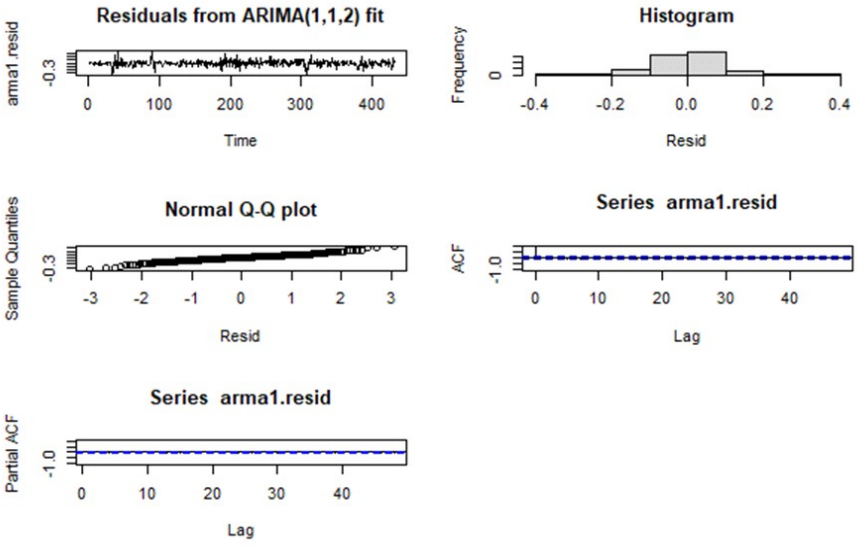


Figure 11. Squared Residual Plot, Histogram, Q-Q Plot, ACF and PACF of log differenced data.

From the figure above, it is clear that there is no AR and MA effect (since the lag is 0 in ACF and PACF plot). But the is volatility still present. Since, the residuals vary randomly around zero and the spread of the residuals are not same throughout the plot, we can conclude that the random shocks is normally distributed with zero mean and non- constant variance. Hence, it is necessary to fit ARCH and GARCH models.

ARCH and GARCH Modeling: A change in the variance over time can cause problems when modeling a time series data with classical methods like ARIMA. An ARCH refers to a type of time series where the conditional variance of the series changes over time. The ARCH modeling technique expect that the time series is stationary. To check whether the variance is non- constant and if we could apply ARCH model, we analyze the ARCH effect. The hypothesis is given by,

: There is no autocorrelation in the squared residuals of the model(the variance is constant and does not depend only on the past observation) Vs.: The variance is conditional on past observation and it is not constant over time.

ARCH LM-test; Null hypothesis: no ARCH effects
data: log_price Chi-squared = 95.091df = 12p-value = 5.066e-15

Here, the p- value is less than 0.05. Thus, the null hypothesis is rejected in favour of alternative hypothesis indicating the presence of ARCH effect in time series. So that GARCH modeling will be appropriate according to the data. .In other words, the test suggests that the variance of the time series data changes over time and is dependent on past values of the series, indicating the presence of conditional heteroscedasticity or volatility clustering in the data. This can have implications for modeling and forecasting the time series data. The GARCH model is a statistical model which is used to describe the conditional variance of a financial time series. The GARCH model extends the ARCH model by allowing for time varying volatility in the conditional variance of time series. This model includes an Auto Regressive term for conditional mean and a Moving Average term for conditional variance which is modeled as a function of past squared residuals and past variances.

Coefficient(s):

a0	a1	b1
0.0007987	0.4841569	0.5435296

The coefficient terms of GARCH model represents the AR and MA terms of the variance equation. Here, the coefficients, a0 represents the intercept of GARCH process, a1 represents the coefficient of lagged variance term in GARCH process and b1 represents the coefficient of lagged squared of residual term in GARCH equation. The order of GARCH model refers to the number of lagged term and number of lagged squared of residual term. Hence we can use the GARCH(1,1) model.

GARCH models: Call:garch(x = rprice, grad = "numerical", trace = FALSE) Coefficient(s):

a0	a1	b1
0.0007987	0.4841569	0.5435296

Information Criteria

Akaike	-1.9528
Bayes	-1.8659
Shibata	-1.9537
Hannan-Quinn	-1.9186

Nyblom stability test:
Joint Statistic: 2.1025

Individual Statistics

mu	0.14464
ar1	0.06106
ar2	0.12995
ar3	0.13410
ma1	0.03907
ma2	0.24519
ma3	0.10703
omega	0.46020
alpha1	0.08424
beta1	0.22601

Here, we can see that the coefficients of ARCH and GARCH terms are equal. Hence, there exist symmetry in GARCH model. We can also see that all the coefficients' terms ar1, ma1, omega, alpha, beta are statistically significant and also the coefficients of conditional variance (omega and alpha) are positive and less than 1. Thus, the assumptions of GARCH model are satisfied. It seems that the GARCH (1,1) model with an ARFIMA (3,0,3) mean model and Normal distribution is a reasonable fit for the data. The estimated parameters are statistically significant, and the model passed several diagnostic tests, including the weighted Ljung-Box test and the Nyblom stability test. Additionally, the Information Criteria values suggest that the model has good explanatory power. Overall, the GARCH (1,1) model provides a useful framework for modeling the volatility dynamics of the time series data. also, the ARCH coefficient (alpha) measures the impact of past squared residuals on the current conditional variance and the GARCH coefficient (beta) measures the impact of past conditional variance on the current conditional variance. Similarly, the lambda coefficient measures the impact of skewness on the conditional variance and the gamma coefficient measures the impact of kurtosis on the conditional variance.

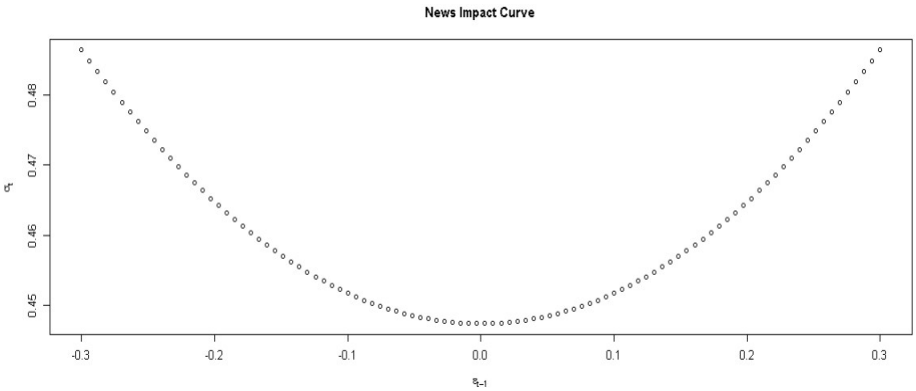


Figure 12: The New Impact Curve (plots the remaining patterns in the residuals after fitting the ARIMA model).

It is clear from the new impact curve that no asymmetries are present in response to positive and negative shocks. We can turn the model to include asymmetry as well by using EGARCH or TGARCH. The estimation results for this model shows that the estimated model GARCH (1,1) fits the volatility of ARIMA (1,1,2). The parameters estimated for GARCH (1,1) model is,

mu	ar1	ar2	ar3	ma1	ma2	ma3	omega	alpha1
0.14464	0.06106	0.12995	0.13410	0.03907	0.24519	0.10703	0.46020	0.08424

beta1 0.22601

Thus, the ARIMA- GARCH model for the return series is given by,

Where,

Forecasting: One of the most important objective of time series modeling is to forecast future values. In forecasting our objective is to produce an optimum forecast that has no error or possibly little error, which leads to minimum mean square error forecasting. Since the best fitted model for our data is ARIMA (1,1,2)- GARCH (1,1) is statistically significant. Also, our model satisfies the assumptions of stationarity and invertibility. Thus, we use this model to forecast the monthly crude oil price for the next 50 months.

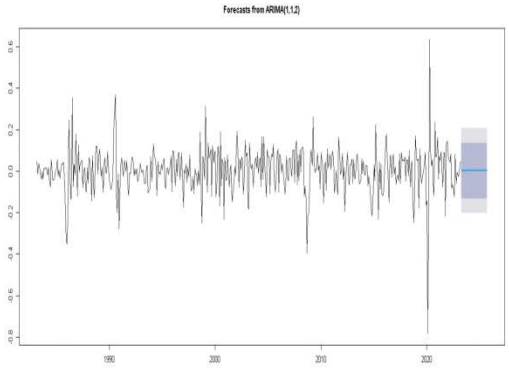


Figure 13: Forecast using ARIMA (1,1,2).

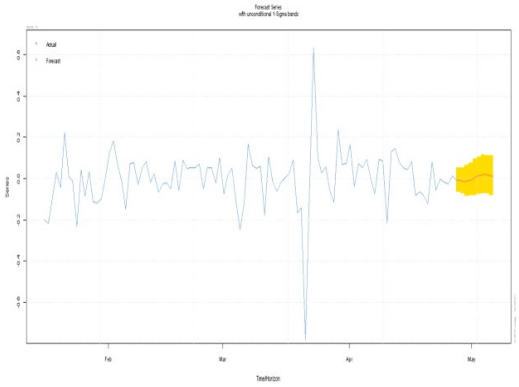


Figure 14: Forecast series with unconditional 1-sigma bands.

Accuracy: We can check the accuracy of our forecast values using MAPE, ME, MAE and MASE.
Box-Ljung test

data: arma1.resid

X-squared = 44.782, df = 48, p-value = 0.6055

Here, the p value is greater than 0.05. Hence, we accept the null hypothesis that the proposed model is sufficient for the data.

Figure 16: Forecast Values of Crude oil price using the given data.

	ME	RMSE	MAE	MPE	MAPE
Training set	-542799.10015	6.053252e+07	3.820345e+06	-7.453269e+06	7.453272e+06
Test set	11.15559	2.546073e+01	2.095369e+01	4.761473e+00	2.928906e+01

	MASE	ACF1
Training set	9.988417e-01	-0.0003490441
Test set	5.478411e-06	NA

CONCLUSION

Traditional statistical models such as ARIMA and Exponential Smoothing methods may not perform well with complex or non-linear data patterns. With the advent of machine learning, several models have been applied to time series forecasting and these models can capture non-linear relationships, handle large feature sets, and provide good performance in many cases. In this study we suggest various models with special reference to OIL Data and these models excel at capturing long-term dependencies, temporal patterns, and non-linear relationships in the data. The choice of model for time series analysis depends on various factors, including the nature of the data, the desired interpretability, the available computational resources, and the specific forecasting requirements. There is no universally superior model, and it is often necessary to experiment with different approaches and select the one that best fits the specific problem at hand. In this study, it is evident that our model performance is relatively good in terms of ME, RMSE, MAE, MASE and ACF1 values. We conclude that these model analyses pave a way to researchers in this area to do more practices on similar approaches in future.

REFERENCE

Adam M. Sykulski, Eric Danioux, Jonathan M. Lilly, Sofia C. Olhede, 2015. *Lagrangian time series models for ocean surface drifter trajectories*, Journal of the Royal Statistical Society Series C (Applied Statistics),Vol 65(1) , pp. 29-50.

Anil Namdeo ,Chuleekorn Tanathitikorn , Nikki Rousseau, Richard McNally, 2018. *A time series analysis of associations between climate change and heat related illnesses and development of a heat health warning system in Thailand*, Newcastle University.

Anna Klimovskaia, Manfred Claassen, Stefan Ganscha, *2016. Sparse Regression Based Structure Learning of Stochastic Reaction Networks from Single Cell Snapshot Time Series,*PLoS Computational Biology*, Vol 12(12), pp. e1005234.*

*Atul Wadagale V, Jagganath Dixit V, Varsharani Vithalrao Kendre, Vaishali Bahattare N, 2017. Forecast measles vaccine requirement by using time series analysis,*Vol 6(28), pp. 2329-2333.

Bella N, Rakesh S, 2018, *A Time series forecast of crude oil prices in India: An empirical analysis using ARIMA techniques*. 7th National Conference on Business Transformation through Green Growth, Globalisation and Governance, new Delhi.

Christian Brownlees, Matteo Barigozzi, 2019. *NETS: Network estimation for time series*, Journal of Applied Econometrics, Vol 34(3), pp. 347-364.

Ezekiel Adebayo Ogundepo, Joseph Bamidele Awotunde, Nureni Olawale Adeboye, Oluwatobi Oluwaseyi Banjo, Sakinat Oluwabukonla Folorunso, 202. *Comparison of Active COVID-19 Cases per Population Using Time-Series Models,* International Journal of E-Health and Medical Communications.

Sai Manoj Cheruvu, 2021. *Stock Price Prediction Using Time Series*, International Journal for Research in Applied Science and Engineering Technology, Vol 9(12), pp. 375-381.

Sevtap Selcuk Kestel A, Bilgi Yilmaz, 2020. *Forecasting house prices in Turkey: GLM, VAR and time series approaches*, Pressacademia,

****