International Journal of Arts, Humanities and Social Studies
Abbreviated Key Title: *Int J Arts Huma Social Studies*
Home page: https://ijahss.in/
Volume: 5 Issue:6 (Nov-Dec 2023); Page No: 01-15
DOI: : http://doi.org/10.5281/zenodo.10115026

OPEN ACCESS
ISSN: 2582-3647
Original Article

IJAHSS

Application of R for ARIMA Modeling and Prediction of Stock Prices

Ashton Banda[1], Gu Jianya[1]

[1]School of Science, Zhejiang University of Science and Technology, Hangzhou, 310023, China

***Corresponding Author**
Ashton Banda

ABSTRACT

In recent years, stock trading has emerged as a pivotal aspect of the corporate landscape, garnering increasing significance. The corporate realm invests substantial time in both the execution of stock trades and the anticipation of stock prices. Extensive research endeavours have been undertaken to enhance the precision of stock price predictions. Traders leverage this information to capitalize on favourable price movements, while investors strategically allocate their resources based on projections of which stocks are poised to augment their net worth. Diverse methodologies are employed in the pursuit of stock price prediction. This study delves into the application of the ARIMA (Autoregressive Integrated Moving Average) model for forecasting stock prices. The paper includes pertinent R command lines and outputs for elucidation. The construction of the model relies on historical data sourced from Yahoo! Finance, specifically from the NASDAQ stock exchange. Estimations are executed using the forecast and predict packages in R. The outcomes are then juxtaposed with real-time stock prices, revealing promising results. The findings underscore the robust potential of ARIMA models in achieving accurate stock price predictions.

Key Words: *ARIMA, stock price prediction, time series, R, stationary, correlation*

INTRODUCTION

Stock price prediction refers to the act of trying to tell the future market price of an equity. The stock (also capital stock) of a corporation is all of the shares into which ownership of the corporation is divided. A single share of the stock represents fractional ownership of the corporation in proportion to the total number of shares. In American English a share is commonly referred to as a stock so we will be using the word stock to mean a share. Ownership of shares in a company entitles a shareholder or a stockholder part of the company's earnings and voting rights in annual general meetings so the more shares one holds the more powerful they are in influencing the decisions to be made in the company and the more earnings they get when the company performs well. More so the higher the number of shares held in a company the more they lose in unfortunate cases of liquidation. Stocks can be traded privately or on the stock exchange.

We will be talking about stocks traded openly on the stock exchange i.e., stocks of public limited companies. Everyday millions of dollars' worth of stocks is traded across the globe and every trader hopes to make profit from his/her sell. Investors who can make the right buy or sell decisions end up making lots of profits at the end of the day. This profit motive has resulted in people trying to predict the future price of a stock. Financial managers want to know the future price of their firm to know its capitalization and this helps them when they want to issue new shares. Stock price modelling goes back as early as when the stock market was opened and since then a lot of models have been made and refined to forecast the prices of stocks. Like any other price, stock price is determined by demand and supply but it's not that simple. Factors which determine the prices of stocks can be divided into two broad categories namely internal and external factors. Internal factors include things like earnings per share of the stock, its price earnings ratio, market capitalisation, profitability of the company, accounting scandals, introduction of new products.

External factors are factors beyond a company's control or influence. The factors include interest rates, if market interest rates are generally low investors prefer to invest in stocks other than bonds so this will likely push up the stock prices and if the interest rates are high, it means it's more profitable to invest in bonds and other interest-based securities, so the stock prices are likely to go down. Economic strength, company stocks tend to track with the market and with their sector or industry peers. Catastrophes e.g., the crash of the Ethiopian Airlines Boeing 737 max led to sharp fall of the Boeing stock price. Some prominent investment firms argue that the combination of overall market and sector movements—as opposed to a company's individual performance—determines most of a stock's movement. (Research has

suggested the economic/market factors account for 90% of it.) For example, a sudden negative outlook for one retail stock often hurts other retail stocks as "guilt by association" drags down demand for the whole sector.

Market sentiment generally affects the prices of a stock. This refers to the psychology of market participants, individually and collectively. Of course, it is subjective and has a bias, but it tends to generally affect the price of stocks. Market sentiment is being explored by the relatively new field of behavioural finance. It starts with the assumption that markets are apparently not efficient much of the time, and this inefficiency can be explained by psychology and other social science disciplines. The idea of applying social science to finance was fully legitimized when Daniel Kahneman, a psychologist, won the 2002 Nobel Memorial Prize in Economics (the first psychologist to do so). Many of the ideas in behavioural finance confirm observable suspicions: that investors tend to overemphasize data that come easily to mind; that many investors react with greater pain to losses than with pleasure to equivalent gains; and that investors tend to persist in a mistake. These are some of the few examples of factors which affect stock prices.

So, to make the right decisions investors must make decisions based on technical analysis such as company charts, stock indices etc and fundamental analysis such as interest rates announcements by reserve banks or the FOMC in the case of America. It is impossible for the human mind to effectively analyse all the data available without the use of machines and software thus a lot of programs like R, Python and Machine Learning have been developed to do that work.

LITERATURE REVIEW

The most efficient way to forecast the future is to understand the present scenarios. The author, Banerjee D [1] tried to develop an appropriate model that helps to forecast the unseen values of the Indian stock market, based on the information collected on the monthly closing stock indices. Based on the ARIMA model they predict the future stock indices which have the strong performance of the Indian economy. It is very important to understand the present status of the market because for many economists, investors and researchers the Indian stock market is the centre of interest. It has been predicted that the performance of the Indian stock market presents a suitable time series ARIMA (1,0,1) model which helps to create the appropriate values of the future indices.

The authors Li Bing and Chan [2] have extracted the ambiguous text through MLP techniques to get the real stock price movements and public sentiments. It has been said that public emotions may be co-related that has shown through Twitter. The authors have used data mining algorithms to mine Twitter data in order to forecast the stock trends using sentimental analysis which comes under fundamental analysis. The estimation of the latest social media analysis works on up-to date public views mining. Through this methodology for social media data mining, it has fulfilled the recognized research problems with adequate experimental results. The algorithm has defined the complete relationships surrounded in social media as a graph with several layers. The top layered attributes and intermediate layered attributes have direct relations; and bottom layered attributes and intermediate layered attribute have in-direct relations.

To forecast stock price trend, the authors, Tao Xing and Yuan Sun [3] have introduced a method based on Hidden Markov Model. Hidden Markov Model first proposed by Baum and Egon, which is a kind of Markov Chain and is used for the pattern recognition technique. This paper finds the hidden relationship existing between the Hidden Markov Model and stock prices. The experimental results show that, this method can get attractive accurate result, particularly efficient in short period prediction.

It is a tedious task for the stock market financial specialists to guesstimate the pattern of the stock exchange costs as effectively as could be allowed to settle on the best exchanging choices.

The authors, Vishwanath R Ha and Leena Sa [4] have proposed a system called APST, which performs the pre-processing of verifiable stock time arrangement information to produce the grouping of approximated values by utilizing multi-scale segment mean methodology. To locate the closest neighbour objects, they utilize the Euclidian separation way to recognize the comparative arrangement of articles. The experimental results of this system show that the executed framework has shown 74% of the exactness.

The authors [Ayodele A] [Adebiyi] [5] have used the ARIMA model to develop an extensive process of building stock price predictive model by obtaining data from NYSE and NSE. Artificial Neural Networks (ANNs) model is very popular due to its ability to learn patterns from data and infer solution from unknown data. Hybrid approaches also engaged to improve stock price predictive models by exploiting the unique strength of each of them. The results obtained from real-life data demonstrated the potential strength of ARIMA models to provide investors short-term prediction that could aid investment decision making process.

The paper by Qasem A. Al-radaideh, Adel Abu Asaf, Eman Alnagi [6] likely explores the application of data mining techniques to forecast stock prices. Data mining involves extracting patterns and insights from large datasets, and in the context of stock prices, it could involve analysing historical market data to identify trends or factors influencing stock movements. The authors may have employed various data mining methods and algorithms to develop predictive models

for stock prices. The article is likely to discuss the methodology, findings, and implications of using data mining in the domain of stock market forecasting.

Stock exchange markets facilitate savings and investment that are beneficial to increase the effectiveness of national economy. The author Li Zhe [7] have used the method of technical analysis in which trading rules were established based on the ancient data of stock trading price and volume. Technical analysis uses various methods that aim to predict future stock price movements based on the assumption that history repeats itself and future market directions can be determined by examining historical stock prices.

ARIMA MODELS

Introduction

ARIMA stands for autoregressive integrated moving average. It is one of the models in the Box-Jenkins models. Box-Jenkins model is a mathematical model designed to forecast data from a specified time series. The Box-Jenkins Model can analyse many different types of time series data for forecasting. Its methodology uses differences between data points to determine outcomes. The Box-Jenkins Model is also generally best suited for short-term forecasting of 18 months or less. Overall, the methodology allows the model to pick out trends, using autoregression, moving averages and seasonal differencing in order to generate forecasts. In ARIMA modelling, data is differenced in order to make it stationary. A stationary time series is a time series process one whose probability distributions are stable over time in the following sense: if we take any collection of random variables in the sequence and then shift that sequence ahead *h* time periods, the joint probability distribution must remain unchanged. In short stationarity means mean, variance, correlation and autocorrelation remain constant. This assumption makes intuitive sense: Since ARIMA uses previous lags of series to model its behaviour, modelling stable series with consistent properties involves less uncertainty

The Box-Jenkins model has four iterative steps namely identification, parameter estimation, diagnostic checking and forecasting. In identification step the main goal is to check the occurrence of a trend in data series movement by plotting time series, data is transformed to make the series stationary. The main reason of wanting to have a stationary series is because stationary data is easy to analyse, and it allows us to apply some laws and theories like the Central Limit Theorem and Law of Large numbers. All this is done to get meaningful sample statistics such as mean, variance and correlations. Such statistics are useful as descriptions of future behaviour if and only if the series is stationary. The stationary process is a foundation in building an ARIMA (*p, d, q*) model where (p, d, q) represent the following;

p: the number of lag observations in the model; also known as the lag order.

d: the number of times that raw observations are differenced; also known as the degree of differencing.

q: the size of the moving average window; also known as the order of the moving average. Here if d=0, then the model becomes ARMA which is a linear stationary model. The same stationary and in-variability conditions that are used for autoregressive and moving average models apply to this ARIMA (*p, d, q*) model. Selecting the appropriate values for p, d and q can be challenging. The AUTO.ARIMA () function in R will do it automatically.

Auto Regression AR ()

Auto Regression techniques estimate the future prices basing on the current price. The first order of auto regression written as AR (1) represents that the next value depends on the current value. AR (2) means the current value depends on the preceding two values. In general AR (n) where n is a whole number means that the current value depends on the previous n numbers. It is possible for n=0, this suggests that the current value is uncorrelated to any preceding terms.

The word auto regression says that it is a regression of the variable against itself.

Moving Average

A moving average technique is a bit different from AR (). Moving average is a technique to find the overall idea within a data set while AR specifically states the number of variables a value is dependent on. Moving average find future trends based on the overall past behaviour. Two commonly used moving average techniques are simple moving average (SMA) and exponential moving average (EMA).

ARIMA function in R

The AUTO.ARIMA () function in R is a very useful function, it makes the computation easy. However, this function examines over conceivable models within the edict limitations provided and returns the best ARIMA model. The value of d also has an effect on the prediction intervals i.e., the more complex the value of d, the more rapidly forecasting intervals surge in size. For d=0, the long-term prediction average deviance will go to the typical deviance of the historic data. It is usually not possible to tell merely from a time plot; what values of p and q are suitable for the specific kind of data. Sometimes it is conceivable to use the ACF plot and closely related PACF plot to govern the appropriate values.

When the observed time series presents trends and non-seasonal behaviour, data transformation and differencing are applied to the data series in order to stabilize variance and to remove the trend before an ARIMA model is applied. One way of making stationary data is by detrending it e.g., by fitting a trendline and subtracting it out prior to fitting a model.

In R we can use the Augmented Dickey-Fuller Test (adf test) to test for stationarity, it's part of the tseries library in R. In addition to that we can use the autocorrelation function (acf) and partial autocorrelation function (pacf) to identify an ARIMA model. ACF and PACF plots are the core of ARIMA modelling. They provide a way to identifying an ARIMA model. They are used in the following way

Step1; If ACF cut off after lag n and PACF dies down, then identify the order of MA (q) in ARIMA (0, d, q) model

Step2; if ACF dies down and PACF cut off after lag n then identify AR (p) in ARIMA (p, d, 0) model

Step3; if ACF and PACF die down means, then we get mixed ARIMA (p, d, q) model, time series needs to differencing (d).

ACF and PACF are primary tools for clarifying the relations that may occur within and between time series at various lags. In the beginning of fitting ARIMA model, the idea of model parameterization as possible yet still be capable of explaining the series i.e., p and q should be 3 or less, or the total number of parameters should be less than 3 in view of Box-Jenkins method. The more parameters the greater noise that can be introduced into the model and hence standard deviation and PACF of squared residuals will help confirm if the residuals are not independent and can be predicted. A strict white noise cannot be predicted either linearly or non-linearly while the general noise might not be predicted linearly yet done so linearly.

Interpretation of ACF

x- shows strong persistence, meaning the current value is close relative to those that preceded it while y- shows a periodic pattern with a cycle length of approximately 4 observations meaning the current value is relatively close to the observation four before it, z- does not exhibit any clear pattern.

The lag for each autocorrelation estimate is indicated by the horizontal axis. Estimating the autocorrelation function at many lags allows us to assess how a time series relates to its past. Correlation shows association

METHODOLGY AND DATA PROCESSING

We use the AUTO.ARIMA function which is inbuilt within R. We convert the prices to the logarithmic format as log returns are supposed to be stationary, at least in periods that are not too long say 18 months. Log pricing is used in option pricing, asset pricing, irreversible investments and other important prices in finance. A logarithm of a series is used to help stabilize a strong growth trend

We are going to look for two different types of stocks which are uncorrelated, and which have different behaviours in terms of price movements. We are going to use Johnson & Johnson and Adobe Systems Inc for our case studies. These two stocks are uncorrelated; their beta correlation is very low. Johnson & Johnson is an American multinational medical device, pharmaceutical and consumer packaged goods manufacturing company founded in 1886. Its common stock is a component of the Dow Jones Industrial Average while Adobe Systems is a software company which historically focused upon creation of multimedia and creativity software products with a more recent foray towards digital music software. Adobe Systems Inc (ADBE) is sensitive due to the nature of business (technology) while Johnson & Johnson is more stable as its prices don't totally depend on the economic activity.

Here is the general outline method of using the ARIMA modelling technique

- Download historical closing prices of the stock in subject in a csv file.
- Convert the dollar prices to logarithmic format as explained earlier.
- Conduct the Augmented Dickey-Fuller test to test whether data is stationary (ADF test). As mentioned earlier these models only work with stationary data. The null hypothesis for an ADF test is that the data are non-stationary. So large p-values are indicative of non-stationarity, and small p-values suggest stationarity. Using the usual 5% threshold, differencing is required if the p-value is greater than 0.05
- Based on the unit test results we identify whether the data is stationary or not. If the data is stationary, then we choose optimal ARIMA models and forecast the future intervals. If the time series is non- stationary, then we difference the time series (computing the differences between consecutive observations). Use ndiffs (), diff () functions to find the number of times differencing needed for the data & to difference the data respectively.

Differencing the series can help in removing its trend or cycles. The idea behind differencing is that, if the original data series does not have constant properties over time, then the change from one period to another might. The difference is calculated by subtracting one period's values from the previous period's values:

$$Y_{dt}=Y_t - Y_{t-1}$$

Higher order differences are calculated in a similar fashion. For example, second order differencing ($d = 2$) is simply expanded to include second lag of the series:

$$Y_{d2t}=Y_{dt} - Y_{dt-1}= (Y_t - Y_{t-1}) - (Y_{t-1}-Y_{t-2}) \qquad 1.0$$

Similarly, differencing can be used if there is a seasonal pattern at specific lags. In such a case, subtracting a value for the "season" from a previous period represents the change from one period to another, as well as from one season to another:

$$Y_{dt} = (Y_t - Y_{t-s}) - (Y_{t-1} - Y_{t-s-1}) \quad 1.1$$

The number of differences performed is represented by the *d* component of ARIMA.

- Retest for stationarity by applying the ADF test again or use autocorrelation function plots, acf (). If the plots show stationarity, then go ahead by applying ARIMA models. These plots can also help to choose the order of parameters for ARIMA model. If the series is correlated with its lags, then, generally, there are some trend or seasonal components and therefore its statistical properties are not constant over time. ACF plots display correlation between a series and its lags. In addition to suggesting the order of differencing, ACF plots can help in determining the order of the MA *(q)* model. Partial autocorrelation plots (PACF), as the name suggests, display correlation between a variable and its lags that is not explained by previous lags. PACF plots are useful when determining the order of the *AR(p)* model. R plots 95% significance boundaries as blue dotted lines.
- Identify Seasonality/Trend. The seasonality in the data can be obtained by the stl () function.
- The next step is model fitting. The forecast package allows the user to explicitly specify the order of the model using the ARIMA **()** function, or automatically generate a set of optimal *(p, d, q)* using AUTO.ARIMA (). This function searches through combinations of order parameters and picks the set that optimizes model fit criteria. There exist a number of such criteria for comparing quality of fit across multiple models. Two of the most widely used are Akaike information criteria (AIC) and Baysian information criteria (BIC). These criteria are closely related and can be interpreted as an estimate of how much information would be lost if a given model is chosen. When comparing models, one wants to minimize AIC and BIC.

While AUTO.ARIMA () can be very useful, it is still important to complete steps described above in order to understand the series and interpret model results. Note that AUTO.ARIMA () also allows the user to specify maximum order for *(p, d, q)*, which is set to 5 by default.

- So now we have fitted a model that can produce a forecast, but does it make sense? Can we trust this model? We can start by examining ACF and PACF plots for model residuals. If model order parameters and structure are correctly specified, we would expect no significant autocorrelations present. Calculate the mean error i.e. the difference between the actual prices and the prices forecasted by the model.

The Figure 1 below shows the general outline of the procedures describe above.

Figure 1: Summary of modelling procedure

Data collection is done during the downloading stage. Data pre-processing and data wrangling is the process of change the dollar closing prices of stocks to logarithmic format and constructing a time series plot to confirm that a time series exist. This can be done without necessarily converting the prices to the logarithmic format. Data training and data processing involves all the processes of checking for stationarity by conducting the Unit root test (Dickey-fuller test), and

construction and of autocorrelation and partial autocorrelation function plots. After data is prepared, we can forecast the results, this is wholly done by ARIMA, it automatically picks up the number of auto regresses, the size of the moving average window, the lag order and the degree of differencing. After making the forecasts, we plot graphs to aid the analysis processes as it easier to understand with the aid of graphs and tables. Finally, we compare the results from the forecasts and the actual live prices, we calculate the mean error to see how accurate the predictions were made.

One last step which more of a test for confirmation, is to conduct the Ljung-Box test. The Ljung - Box test also provides a different way to double check the model. Basically, Ljung- Box is a test of autocorrelation in which it verifies whether the autocorrelations of a time series are different from 0. In other words, if the result rejects the hypothesis, this means the data is independent and uncorrelated; otherwise, there still remains serial correlation in the series and the model needs more modification. The procedure includes observing residual plot and its ACF and PACF diagram, and check Ljung- Box result. If ACF and PACF of the model residuals show no significant lags, the selected model is appropriate.

CASE STUDIES AND RESULTS

In this section we are going to look at two different cases as way of illustrating how ARIMA modelling works. We chose two different companies which are in different industries and are uncorrelated.

Case study 1: Adobe Systems Inc

We downloaded the monthly closing price data of ADBE from NASDAQ for the periods 2000-01-01 to 2018-12-01. We have 229 data points. Adobe Systems is a software company which historically focused upon creation of multimedia and creativity software products with a more recent foray towards digital music software. Technology stocks have high betas (usually above 1) because the demand for their products usually vary with the business cycle. ADBE is a component of the S&P 500 which is is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE, NASDAQ, or the Cboe BZX Exchange

We are going to use data from 2014 to 2017 to build our model then use the closing prices of 2018 to compare the live prices with the prices generated from the ARIMA model.

A good starting point is to plot the time series and visually examine it for any outliers, volatility, or irregularities. The diagram (figure 4.1) below shows the time series monthly closing prices of Adobe Systems Inc (ADBE). Of course, there are a lot fluctuation, but it is clear that from the year 2000 to 2018 the stock was an uptrend with the highest price recorded in 2018 and the lowest recorded in 2003, surprisingly the stock didn't tumble much during the 2008 global crisis. The company has seen a significant growth in terms of market capitalisation. The price per share has risen by more than 150% since the early months of the year 2000. Take for example a shareholder who bought the share in 2000 when the price was $43 and now the share is currently trading at $200, this represents a profit of around $150 plus the dividends received in the past 10 years. Growth of the stock is partly due to increase in demand of its products.

It should be noted that if there are any suspected outliers that could bias the model by skewing statistical summaries. R provides a convenient method for removing time series outliers: tsclean () as part of its forecast package. tsclean () identifies and replaces outliers using series smoothing and decomposition. The time series plot is made to confirm the existence of a time series. At this stage it's not necessary to convert the prices to the logarithmic format as we are merely confirming the existence of a time series.

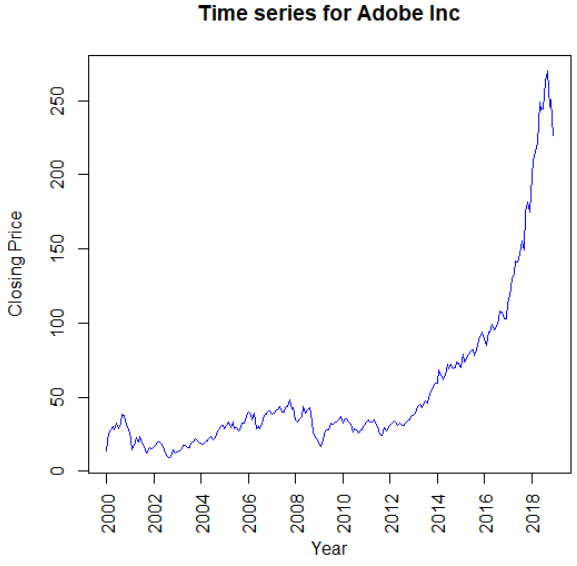

Figure 2: Time series for Adobe Inc

The next step is to convert the closing prices to the logarithmic format. We conduct the Dickey-Fuller Test (ADF test) to determine the stationarity of the time series. The results are shown below.

```
        Augmented Dickey-Fuller Test

data:  Close
Dickey-Fuller = -0.74478, Lag order = 6, p-value = 0.9654
alternative hypothesis: stationary
```

The p-value is 0.9654 which is greater than 0.05, which means the data is non-stationary. Differencing is needed to make the data stationary. Further evidence of non-stationarity can be seen from the ACF& PACF plots below. The lag.max= 30.

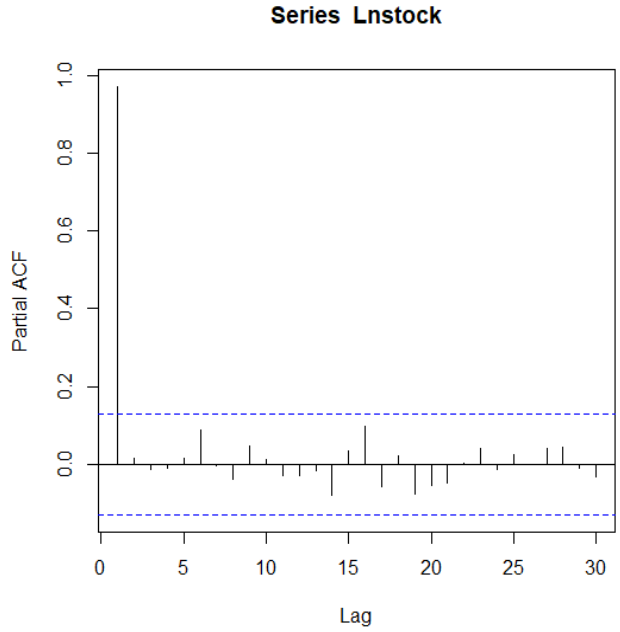

Figure 3: PACF plot for original log prices

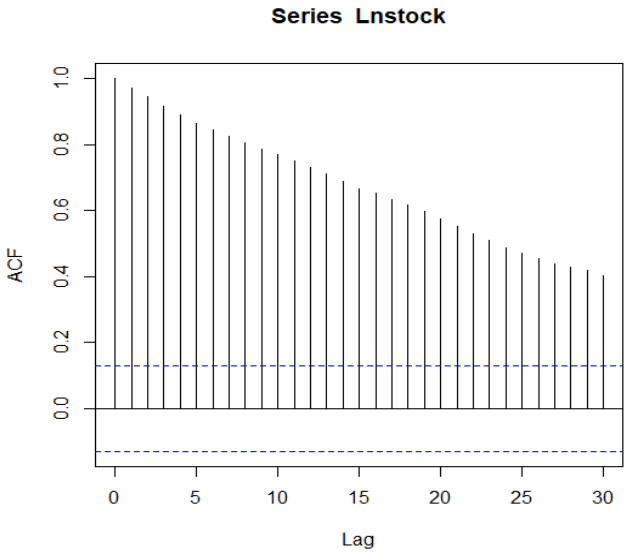

Figure 4: ACF plot for original log prices

Estimating the autocorrelation function (ACF) at many lags allows us to assess how a time series for the closing prices relates to its past. In the plots produced by acf (), the lag for each autocorrelation estimate is denoted on the horizontal axis and each autocorrelation estimate is indicated by the height of the vertical bars.

The figure 4 above shows a gradual descent of the ACF plot, it shows weaker correlation between values as they get

far apart each other. The PACF immediately dies down which shows no correlation between values. At lag=0, the autocorrelation is equal to 1 and as the number of lags increase, the autocorrelation gradually falls showing weak correlation between values which are far apart in terms of time. The correlation for lag=30 is 0.4 which shows a weak positive correlation but it's still positive, which is normal for stock prices, option prices, and currency exchange rates as the daily ranges or closing prices are correlated and usually depend on the previous values.

We need to difference the data to make it stationary. To confirm that whether our time series is stationary now, we conduct a Dickey-Fuller test on the differenced time series. The results are shown below,

```
        Augmented Dickey-Fuller Test

data:  diffLnstock
Dickey-Fuller = -6.6658, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diffLnstock) : p-value smaller than printed p-value
>
```

From the information above the p-value is 0.01 which is less than the threshold of 0.05 meaning the data is stationary. Furthermore, there is an additional warning message saying the printed value is less than 0.01 suggesting that 0.01 was an estimate round figure so the p-value was likely to be something like 0.0095…. or more which was rounded to 0.01.

We also construct the ACF and PACF plots to show that the differenced time series of closing prices is stationary. We used the same lag to effectively compare the difference between a differenced time series from the original one. The results are show in the following diagrams,

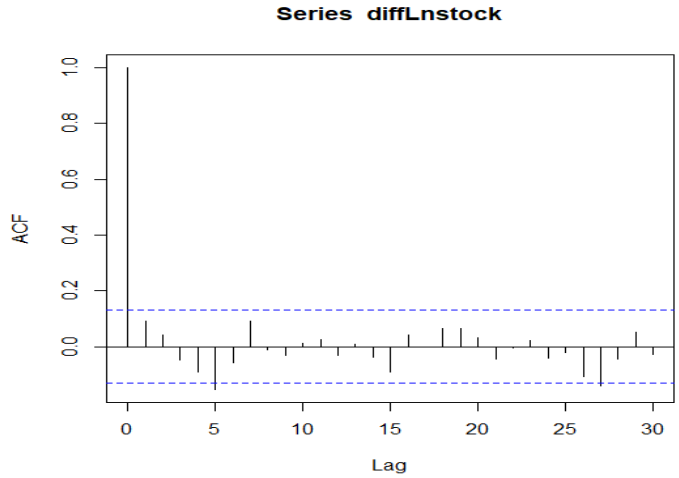

Figure 5: ACF plot for differenced log prices

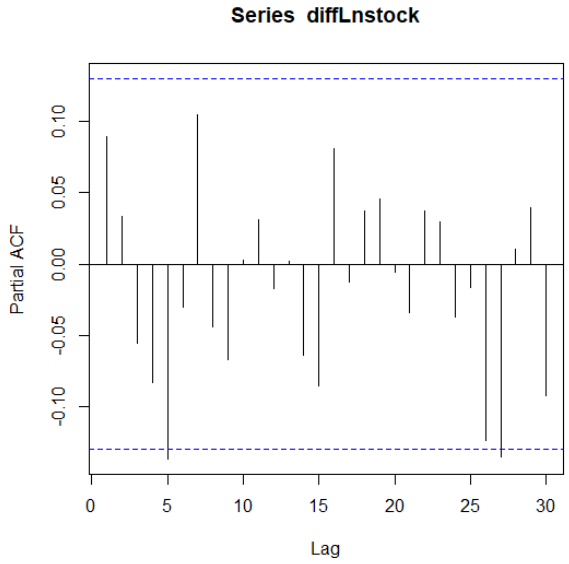

Figure 6: PACF plot for differenced log prices

We can clearly see the difference between the figures 5 & 6, The latter shows differenced data while the former shows undifferenced data. Now that our data is stationary, we can go ahead with the ARIMA model construction. The ACF plot of the differenced data does not show any correlation whatsoever, the correlation at lags 2 to 30 are ranging from 0.2 to -0.2 which shows weak correlation between preceding closing prices.

As mentioned earlier, we are using the data from 2014 to 2017 for model building the use the values of 2018 for testing the data. We selected a short period because ARIMA works best in the short run usually periods less than 18 months. We used the AUTO.ARIMA and forecast functions to predict the monthly closing prices for the year 2018. The results are shown by the figure 4.1.6 below. The forecasts are represented by the blue colour. The forecasts are from ARIMA (0,1,1) which means p=0, d=1 and q=1.In this case "drift" simply means d=1, the AUTO.ARIMA() function from the forecast package in R automates the inclusion of a constant. By default, for d=0 or d=1, a constant will be included if it improves the AIC value; for d>1 the constant is always omitted. This implies that the current price is not dependent on the previous price, the trend that existed was removed by differencing the series once i.e. the d=1 and the order of the moving average is 1 as well.

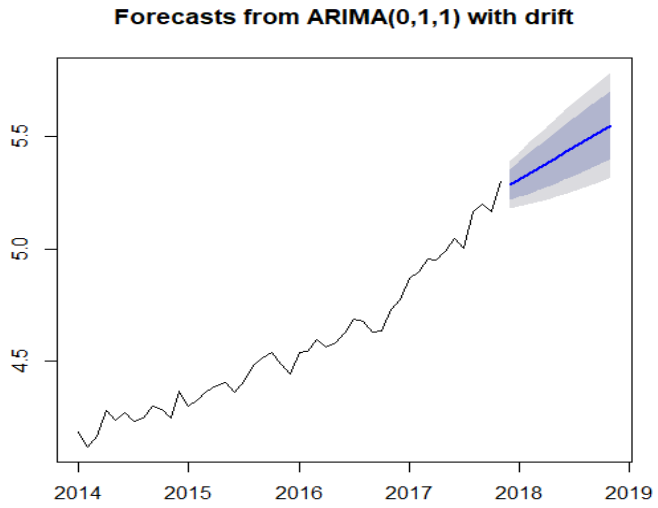

Figure 7: Forecasts from ARIMA (0 1 1)

The AUTO.ARIMA() function automatically picked those values automatically. The blue line is somehow straight which is a bit unrealistic given the past behaviour of the series. Recall that the model is assuming a series with no seasonality and is different from the original non-stationary data. In other words, plotted predictions are based on the assumption that there will be no other seasonal fluctuations in the data and the change in the closing price from one day to another is more or less constant in terms of its mean and variance. This forecast may be a naive model, but it illustrates the process of choosing an ARIMA model and could also serve as a benchmark to grade against as more complex models are built.

Other forecasting techniques, such as exponential smoothing, would help make the model more accurate using a weighted combination of seasonality, trend, and historical values to make predictions. In addition, the price of the stock is probably highly dependent on other factors, such new product announcements, holidays e.g., Christmas, time of the year etc. One could try fitting time series models that allow for inclusion of other predictors using methods such ARMAX or dynamic regression. These more complex models allow for control of other factors in predicting the time series. These are beyond the scope of this thesis.

Case Study 2: Johnson & Johnson (JNJ)

Johnson & Johnson is an American multinational medical device, pharmaceutical and consumer packaged goods manufacturing company founded in 1886. Its common stock is a component of the Dow Jones Industrial Average. It's a bit different from Adobe, JNJ is less volatile as demand for personal and household products have little relation to the state of the company. It has a positive beta but it's less than 1. It is also part of the S&P 500.

Similar to Adobe, we downloaded the monthly closing prices of the stock from the period 2000-01-01 to 2018-12-31. We have in total 228 data points. We are going to use the first 216 data points for model building then use the remaining 12 points to calculate the accuracy of the model

The assumptions and steps are the same as for case study 1 so we will not rewrite the general outline, instead we will go direct to the results.

The first step is to check if a time series exists on the given data. The diagram (figure 8) below confirms that a time series does exist. The figure depicts the original pattern of the series and allows us to have general overview of whether the time series is stationary or not. From the graph below the time series have random walk pattern. Despite of the random walk the stock is on an uptrend with the lowest closing price of $35 recorded in the early months of the year 2000 in March to be precise. Ever since that time the closing price never dropped below $35 with the highest price of $147 recorded at the end of 2018. This increase in share price represents an increase in market capitalisation by more than 100% and this is good for shareholders and management. Shareholders will get more returns for their investment through dividends and profits from the rise in the market price of the share. Managers of big companies are usually paid in stocks as part of their remuneration packages Management can as well easily access lines of credit, most lending institutions are likely lending to companies with performing stocks than to firms with struggling stocks, easy access to credit lines can improve the cashflow of a company. This enables companies to meet short term obligations like repayment of short-term loans and payment of dividends. It should be noted that an increase in a share price can mean the share was initially under-priced but it's unlikely in this case, Johnson & Johnson's increase is a result of the good performance of the company. JNJ has been declaring dividends of $0.70/share on average for the past 5 years. The payments are made quarterly. That's a reasonable measure that the company has been doing well.

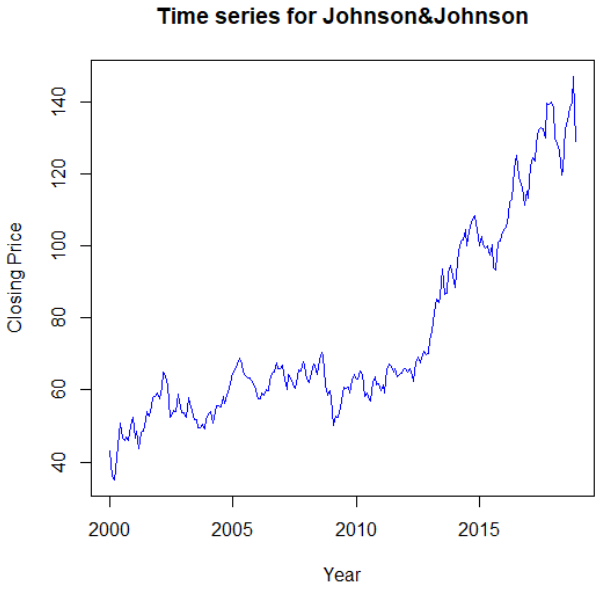

Figure 8: Time series for JNJ

The next step is to conduct the unit root test. This is done to test for stationarity. We conduct the ADF test of the original time series at 5% significance level. The results are shown below.

```
        Augmented Dickey-Fuller Test

data:  Lnstock
Dickey-Fuller = -1.469, Lag order = 5, p-value = 0.7988
alternative hypothesis: stationary

>
```

According to the unit root test conducted the series is not stationary. The p-value=0.7988 which is greater than 0.05. The series needs to be stationary before using the ARIMA or AUTO.ARIMA functions. Further evidence of non-stationarity is shown by the ACF and PACF plots shown below.

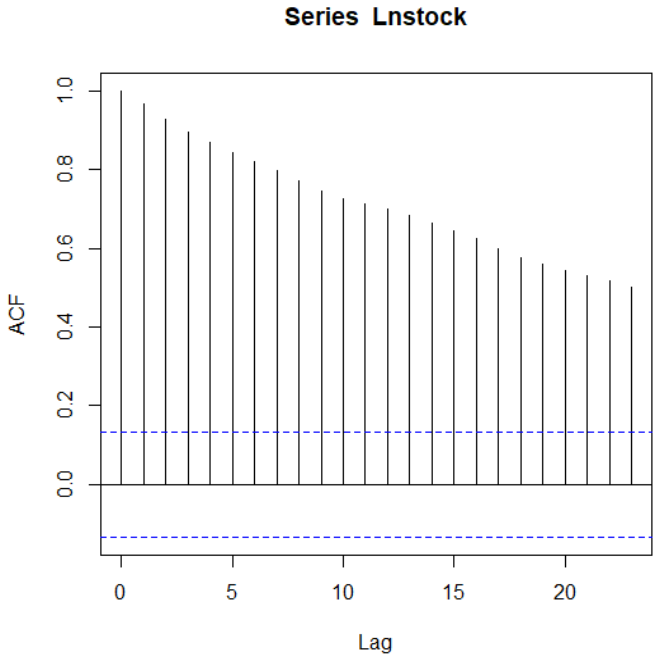

Figure 9: PACF plot for original log prices

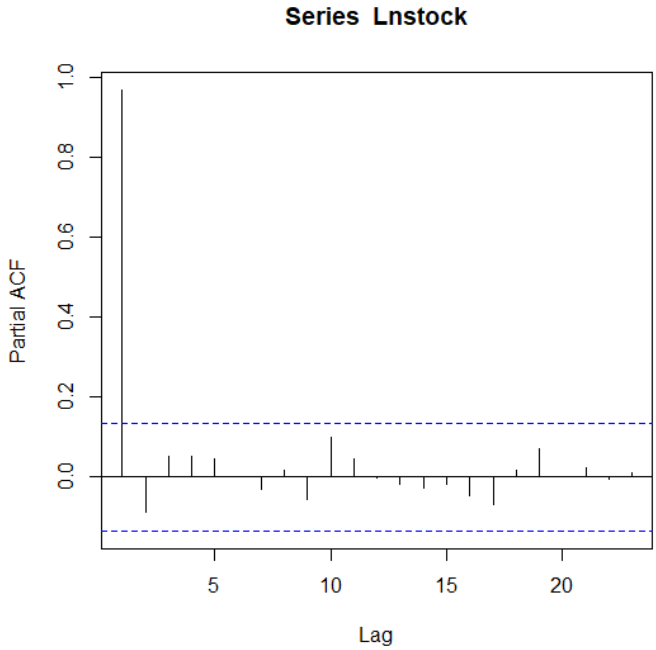

Figure 10: PACF plot for original log prices

From the graphs shown above, the PACF dies down extremely slowly which simply means that the time series is non stationary. The figure 9 above shows a gradual descent of the ACF plot, it shows weak correlation between values as they get far apart each other. The values are strongly positively correlated, the rho ranges from 1.0 to 0.5. The next step should be an attempt to make the series stationary. One way to do this is by differencing. Below are the ADF test results of the differenced time series.

```
        Augmented Dickey-Fuller Test

data:  diffLnstock
Dickey-Fuller = -6.2921, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diffLnstock) : p-value smaller than printed p-value
> |
```

The p-value=0.01 which is less than 0.05, this supports the hypothesis that the series is stationary. The additional warning which states that p-value is smaller than printed p-value means the actually p-value is approximately at least 0.095…. To add more, we construct the ACF and PACF plots of the differenced series. The results are shown below;

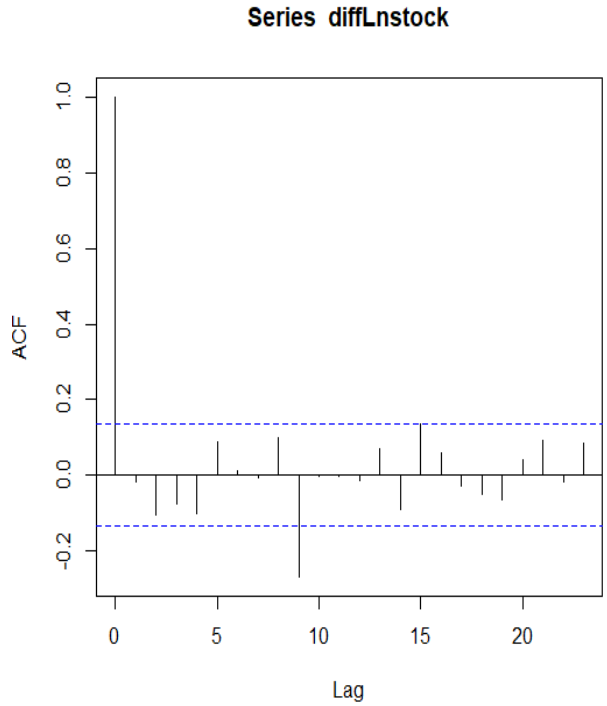

Figure 11: ACF plot for differenced log prices

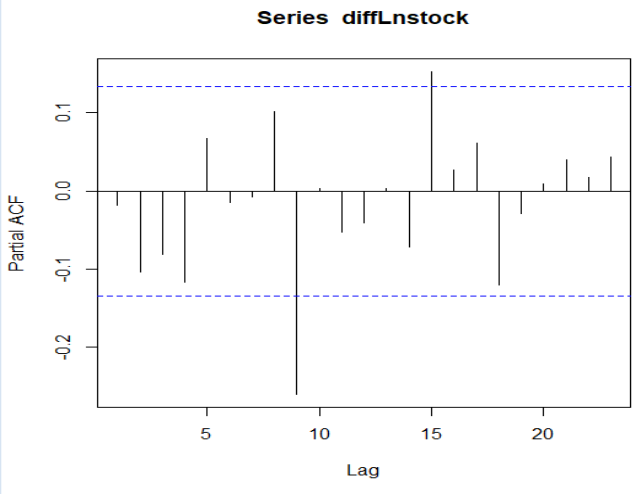

Figure 12: PACF plot for differenced log prices

The plots above clearly show that there is weak or no correlation between preceding closing prices. To be more specific the correlation ranges between -0.2<rho<0.2, this can be interpreted as a weak correlation, but this doesn't mean that a time series does not exist but what this shows is that there is no clear linear trend between closing prices in a differenced time series of JNJ.

As mentioned earlier, we are using the data from 2014 to 2017 for model building the use the values of 2018 for testing the data. We selected a short period because ARIMA works best in the short run usually periods less than 18 months. We used the AUTO.ARIMA and forecast functions to predict the monthly closing prices for the year 2018. The results are shown by the figure 13 below. A closer analysis shows that the stock price is more likely to go up than to go down. The trend is quite clear that the stock is on an uptrend and it would be quite strange for the stock to suddenly fall unless there has been a news release which negatively affects the company for example JNJ manufactures baby powder and if one its products is discovered to cause skin rash, this will definitely slow down its baby powder sales and a fall in demand can reduce the profit earned and most likely will reduce the earnings per share thereby reducing the demand for its share. A fall in demand will force the stock price down. It should be noted ARIMA in R cannot predict such falls or rises especially when they are sudden.

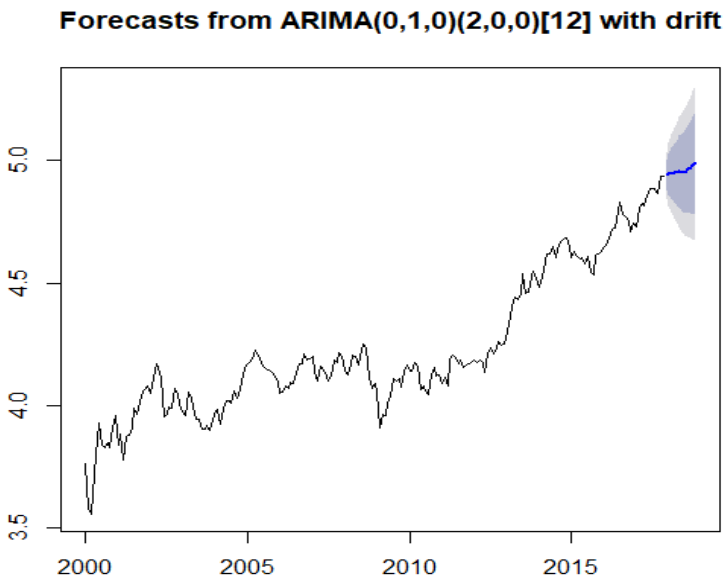

Figure 13: Forecasts from ARIMA (0, 1, 0) (2, 0,0) with drift

A closer analysis shows that the stock price is more likely to go up than to go down. To help simplify that a sample of the results from the forecast and the actual prices are compared in the following table 1. The mean error is -0.08204953 which approximately 8%. The error margin is good for an estimate. It gives investors, managers, traders, economic analysts etc a rough idea of what the stock price is likely to be in the next 12 months. In this example it would be odd for JNJ's price to suddenly jump to $200 in the next few months.

The forecasted prices are rising faster than the actual prices but both prices are moving in the same direction. It can be noticed that when the actual price falls from its previous value, the forecasted price follows the same trend for most values in the table.

Table 1: Comparison of forecast results and the actual results.

Actual Price	Forecasted Price
138.19	139.80
129.88	140.42
128.15	140.87
126.49	141.16
119.62	141.41
121.34	142.04
132.52	141.46
134.69	141.74
138.17	143.50
139.99	144.63
146.90	145.60
129.05	147.26

The forecasted prices are rising faster than the actual prices but both prices are moving in the same direction. It can be noticed that when the actual price falls from its previous value, the forecasted price follows the same trend for most values in the table above.

Besides using the AUTO.ARIMA() function, we could have used ARIMA () where we could pick our own (p d q) values but the main advantage of using AUTO.ARIMA() is it picks the best combination basing on the AIC and BIC values derived from the time series.

LIMITATIONS

In as much as we may try to predict the stock prices, no model can accurately predict the future price stock. This is due to the inherent nature of stocks which is a random walk. The Efficient Market Hypothesis states that it is difficult or

impossible to beat the stock market. Our research was mainly based on the weak form which states that current stock prices reflect all the information contained in past prices. Many investors trade on the stock market, and the universe are aware of all past prices. Buyers and sellers of shares have come to a consensus viewpoint of the value of investment, and that value is reflected in the current price of the stock. It is not possible to profit based on information on past prices. Evidence of randomness was first discovered by Kendall Maurice in 1953, his results were rather shocking and disturbing to most economists of his time. He suggested that prices were likely to go up as they were likely to go down on any particular day regardless of past performance. The data provided no way to predict price movements. His results seemed to confirm the irrationality of the market but in fact it was the opposite. For example, take a company ADBE whose stock is currently trading for $165 per share, then we use ARIMA model to predict the price of the stock next month to be $180 per share.

Obviously, investors would rush to buy the stock now and all those holding will not be willing to sell their stock, the net effect would be a jump of the stock to $180 however, if markets functioned this way investors would reap unending profits by simply using computer models to predict prices. If we look closely at this, we would realise that this is not sustainable for a long time. This simple example illustrates why Kendall's attempt to find recurrent patterns in stock price movements was doomed to failure. A forecast about favourable future performance leads instead to favourable current performance, as market participants all try to get in on the action before the price jump. More generally, one might say that any information that could be used to predict stock performance should already be reflected in stock prices. As soon as there is any information indicating that a stock is under-priced and therefore offers a profit opportunity, investors flock to buy the stock and immediately bid up its price to a fair level, where only ordinary rates of return can be expected.

Other forms of the efficiency market hypothesis are the semi-strong form and the strong form. The semi strong-form hypothesis states that all publicly available information regarding the prospects of a firm must be reflected already in the stock price. Such information includes, in addition to past prices, fundamental data on the firm's product line, quality of management, balance sheet composition, patents held, earning forecasts, and accounting practices. Again, if investors have access to such information from publicly available sources, one would expect it to be reflected in stock prices.

The strong-form version of the efficient market hypothesis states that stock prices reflect all information relevant to the firm, even including information available only to company insiders. This version of the hypothesis is quite extreme. Few would argue with the proposition that corporate officers have access to pertinent information long enough before public release to enable them to profit from trading on that information. In different countries there are different laws put in place to prevent insiders from profiting by exploiting their privileged situation. Defining insider trading is not always easy, however. After all, stock analysts are in the business of uncovering information not already widely known to market participants. The distinction between private and inside information is sometimes impossible to see.

Further evidence is of randomness is explained by Brealey, Meyers, and Allen. They created scatter plot for prices changes of four stocks. They plotted the percentage change on day $t+1$ against the percentage change on day t. There was no pattern; points were distributed evenly over the four quadrants. Moreover, the autocorrelation coefficients of the time series of percentage changes were close to zero.

Of course, the problems described above relate to all stock price predicting models, ARIMA is also included, more so ARIMA models do not take into account external random factors which affect stock prices for example, new releases like resignation of directors, abuse and sex scandals of employees especially by top management. All these factors massively affect the prices of stocks. There are other techniques which use text mining like machine learning but ARIMA in R is not one them. It should be noted that even if these problems exist, traders still make profitable trades, in short managers, investors, economic analysts should not solely rely on computer models, these models should act as guidelines along other tools. Besides we noticed that the results generated are not 100% correct there is a margin of error though its small and insignificant.

More so ARIMA models directly rely on past values, and therefore work best on long and stable series. Also note that ARIMA simply approximates historical patterns and therefore does not aim to explain the structure of the underlying data mechanism.

CONCLUSION

In this thesis we made an attempt to develop a prediction model for forecasting the stock trends based on technical analysis using historical time series stock market data and data mining techniques. Substantial volume of research exists on the topic, very little is aimed at long term forecasting while making use of machine learning methods. The accuracy steadily decreases with the number of predictions made; it cannot be used for long term predictions in stock market. The experimental results obtained demonstrated the potential of ARIMA model to predict the stock price indices on short-term basis. This could guide the investors in the stock market to make profitable investment decisions whether to buy/sell/hold a share. With the results obtained ARIMA model can compete reasonably well with emerging forecasting techniques in short-term prediction.

There are many steps in our model creation process which are ready for further exploration and improvement. This paper can be extended by integrating the technical analysis and fundamental analysis techniques. Through the evaluation of social media analysis particularly on public opinions using fundamental analysis techniques can be incorporated in order to obtain better results. ARIMA (p, d, q) model focused on analysing time series linearly and it does not reflect recent changes as new information is available in the data. Therefore, in order to more accurately develop the model, researches need to incorporate new data and estimate parameters again for forecasting

REFERENCES

[1] Banerjee, D., "Forecasting of Indian stock market using time-series ARIMA model", 2nd IEEE International Conference on Business and Information Management (ICBIM), [J] January 2014, pp. 131-135

[2] Li Bing, Chan, K. C. C., C. Ou, "Public sentiment analysis in Twitter data for prediction of a company's stock price movements", 11th IEEE International Conference on e-Business Engineering (ICEBE), [J] November 2014, pp. 232-239

[3] Tao Xing, Yuan Sun, Qian Wang, Guo Yu, "The analysis and prediction of stock prices", [J] IEEE International Conference on Granular Computing (GrC), December 2013, pp. 368–373

[4] Vishwanath R. Ha, Leena Sa, Srikantaiah K. Ca, K. Shreekrishna Kumar b., P. Deepa Shenoya, Venugopal K. Ra, S. S. Iyengarc, L. M. Patnaik, "Forecasting stock time-series using data approximation and pattern sequence similarity", International Journal of Information Processing (IJIP), [J] September 2013, pp. 90-100.

[5] Ayodele A. Adebiyi, Aderemi O. Adewumi, Charles K. Ayo, "Stock price prediction using the ARIMA model"[J], 16th IEEE International Conference on Computer Modelling and Simulation (UKSim), March 2014, pp. 106 -112

[6] Qasem A. Al-radaideh, Adel Abu Asaf, Eman Alnagi, "Predicting stock prices using data mining techniques" [J], The International Arab Conference on Information Technology 2013

[7] Li Zhe; "Research on China's stock exchange markets: problems and improvements" [J], International Conference on Education and Management Technology, 2010. pp 465-469