Kwame Nkrumah University of Science and Technology, Kumasi

COLLEGE OF SCIENCE

FACULTY OF SCIENCE

DEPARTMENT OF MATHEMATICS

COMPARATIVE STUDY OF STOCK PRICE FORECASTING USING ARIMA AND ARIMAX MODELS

BY

KOFI AGYARKO ABABIO

JUNE 2012

Kwame Nkrumah University of Science and Technology, Kumasi

COLLEGE OF SCIENCE

FACULTY OF SCIENCE

DEPARTMENT OF MATHEMATICS

COMPARATIVE STUDY OF STOCK PRICE FORECASTING USING ARIMA AND ARIMAX MODELS

BY

KOFI AGYARKO ABABIO

A thesis submitted to the Department of Mathematics, Kwame Nkrumah University of Science and Technology in partial fulfillment of the requirements for the degree of Master Philosophy

in

Applied Mathematics

JUNE 2012

DECLARATION

I hereby declare that this submission is my own work towards the award of the M.Phil degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

KOFI AGYARKO ABABIO	………………	……………………
(20131689)	Signature	Date

Certified by:

DR. F. T. ODURO	………………	……………………
(SUPERVISOR)	Signature	Date

Certified by:

MR. F. K. DARKWAH	………………	……………………
(Head, Department of Mathematics)	Signature	Date

ACKNOWLEDGMENT

God Almighty has enabled me to conceptualize, undertake and complete this work which isconducted as part of the academic requirement of my master's program.

Secondly I am grateful to my family, especially to my wife Rita, my son Kofi and lastly Aunty B, the new addition to the family for their support throughout my course of studies.

I wish to acknowledge my gratitude to my supervisor Dr. F. T. Oduro for providing perfectguidelines and supporting me from the beginning till the end of project. Withouthis timely corrections, comments and valuable suggestions, I would not have been able to complete this thesis in time.

I extend my gratitude to Dr. Z. K. M. Batse, Dean, Faculty of Applied Science, Kumasi Polytechnic for providing valuable commentsduring the dissertation progress and to all my M.Phil mates.

Special and grateful appreciations are due to my very special friend Mr. Eric Aidoo for his constant moral support and encouragement in difficult times.

DEDICATION

To my father, Mr. K. A. Ababio for his care, support and love for the family.

ABSTRACT

This thesis evaluates the in-sample forecasting accuracy of two forecasting models namely ARIMA and ARIMAX. . Data used was monthly adjusted close price of four stocks in the Oil and Gas Industry in the London Stock Exchange from 2005 - 2010 with a total of 72 observations. The Mean Square Error (MSE), Mean absolute Error (MAE) and Root Mean Square Error (RMSE) serve as the error matrices in evaluating the forecastability of the models. The effect of Akaike Information Criterion (AIC) and the linear correlation on candidate models among the considered stocks were tested.The ARIMAXmodels performed well with lower error metrics as compared to the ARIMA models in all time regimes.The Linear Correlation on the other hand had little or no influence at all on the in – sample forecastability as compared to the AIC which had significant influence on the error metric. The results support that themarket is efficient and hence no investor has undue advantage of gaining from it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

“Prediction is a very difficult art, especially when it involves the future” -Neils Bohr (Nobel Laureate Physicist).

“Forecasting is the process of making statements about events whose actual outcomes (typically) have not yet been observed” Wikipedia. Words such as predicting are used to also refer to forecasting. The art of forecasting into the future is a very vital but important exercise to many stakeholders in diverse industries. As farmers would like to know the future rainfall pattern in order to properly sow their seeds and at the right time so do financial analyst expect to know the future performance of various market stocks to guide investment options available to them.

There are various important variables that are of much interest to the economy and social well-being of a country that can be forecasted. Such variables of interest can be the Gross Domestic Product (GDP), Inflation, Unemployment rate, Birth rate, Mortality rate, and Import/Export amongst others.

The primary advantage of forecasting is that it provides various stakeholders with valuable information that can be used to make decisions about the future. While forecasting/ projecting estimates for short periods, the art of predicting estimates into a distant future. It is not possible to accurately forecast the future. Forecasting into the future comes with a margin of error. The margin of error widens especially when forecasting deep into the future, in other words when predicting. Variables and their expected influence may change (with social, economic and political change) and new variables may emerge (e.g. the recent Ghana oil find). These errors arise as a result of the level of inaccuracy of the base information used and the method used to forecast into the future. This makes the choice of the forecasting method pivotal when predicting into the future. In many cases forecasting uses quantitative data rather than qualitative data which depends on the judgment of experts. But forecasting that uses quantitative data is more scientific than that of qualitative data.

Future price movement can be forecasted using two approaches: the fundamental or intrinsic value analysis and the "chartist" or technical analysis. Technical analysis is widely used in practice applying different methods as compared to fundamental analysis. "The basic assumption of all the chartist or technical theories is that history tends to repeat itself, that is, past patterns of price behavior in individual securities will tend to recur in the future. Thus the way to predict stock prices (and, of course, increase one's potential gains) is to develop a familiarity with past patterns of price behavior in order to recognize situations of likely recurrence" (Fama, 1986). This technique is seen by majority of market professionals as very obscure, in other words, it is surrounded by a degree of mysticism. The best known instance where the chartist approach was applied to forecast stock prices is the Dow Theory. "Dow Theory on stock price movement is a form of technical analysis that includes some aspects of sector rotation" Wikipedia. Whiles technical analysis uses past prices to foretell the future; the fundamental analyses integrate all events that affects the market (i.e. market forces) and can be used to estimate future price movement. "The assumption of the fundamental analysis approach is that at any point in time an individual security has an intrinsic value (or, in the terms of the economist, an equilibrium price) which depends on the earning potential of the security. The earning potential of the security depends in turn on such fundamental factors as quality of management, outlook for the industry and the economy, e.t.c" (Fama, 1986). There is no doubt that the price movement of stocks is highly influenced by economic factors such as demand and supply of the commodity. It is assumed here that the final adjusted price of a share is a summary of all the economic factors that influences the price movement of the stocks in general.

1.1 Background of Study

This thesis will concentrate on the use of technical analysis in forecasting future price movement of stocks in the Oil and Gas Producers. The industry was selected by convenience. The companies under the selected industry are as follows:BG – BG GROUP PLC, BP – BP PLC, CNE – CAIRN ENERGY PLC, ESSR – ESSAR ENERGY PLC, RDSA – ROYAL DUTCH SHELL PLC 'A', RDSB – ROYAL DUTCH SHELL PLC 'B', TLW – TULLOW OIL PLC.

The historical adjusted close quote data for the considered time range/horizon for ESSAR ENERGY PLC (ESSR), ROYAL DUTCH SHELL PLC 'A' (RDSA) and ROYAL DUTCH SHELL PLC 'B' (RDSB) in the Oil & Gas industry are unavailability.

1.2 Problem Statement

The question of whether forecasts are necessary crops up from time to time. In all, the inescapable conclusion is that no matter what type of enterprise you are in, or what function you perform, there is a need for some kind of future estimate upon which to build a plan.

Marketing practitioners need forecasts to determine which new products or services to introduce or discontinue; which markets to enter or exit; and which products to promote, Salespeople on the other hand, use forecasts to make sales plans, since sales quotas are generally based on estimates of future sales.

Finance professionals use forecasts to make financial plans. They also use them to report about their earnings expectations. Investors invest their hard earned capital in stocks with the expectation of gaining from their investment through a positive payoff. Hence having a good knowledge about share price movement in the future serves the interest of financial professionals and investors. This knowledge about the future boosts their confidence by way of consulting and investing. It goes without saying that forecasting methods which will predict the future movement of share prices with the least error margin will be of much interest to financial professional and investors.

There are many forecasting methods in projecting price movement of stocks such as the Box Jenkins method, Black-Scholes model, and Binomial model amongst others.

In this study, less sophisticated methods will be employed in forecasting the price movement of stocks under the above selected industries in the London Stock Exchange (LSE). These less sophiscated methods will go a long way for non-statisticians and nonprofessionals to understand and use.

1.3 Objective to the Study

This study seeks to investigate which forecasting method under consideration, gives the minimum forecasting error by considering three error metrics.

The specific objectives of this study are as follows:

1. to compare the in-sample forecasting efficiency of two different methods (i.e. ARIMA and ARIMAX).
2. to investigate the effect of Akaike Information Criterion (AIC) and linear correlation in evaluating the in-sample forecasting accuracy of the Box-Jenkins Method with/without an exogenous variable.
3. to test if the considered stocks follows a random walk.

1.5 Methodology

Secondary data wasobtained from Yahoo finance. Geometric Mean Regression, Exponential Smoothing and Box Jenkins Method (model selection based on residual analysis and Akaike information criterion) were the main statistical methods to be used for analysis. Data analysis wasperformed using R statistical package. Forecasting efficiency was derived based on the following error measures MAE – Mean Absolute Error, RMSE – Root Mean Square Error and MSE – Mean Square Error

1.6 Significance of Study

The results of the research will go a long way to help the managers of financial portfolios and to understand and appreciate the underlying factors behind the in-sample forecasting accuracy of stocks in the London Stock Exchange and other exchanges. This will further boost the confidence of stakeholders in the financial industry to do more business with less risk. Other beneficiaries of the research are investors, shareholders, directors, regulators and other financial institutions as well as researchers in the academia.

1.8 Organisation of the Thesis

This thesis looks at two standard methods of forecasting market share price in the London Stock Exchange (LSE). This thesis contains five main chapters. Each of the five chapters has headings containing what the headings reflects, namely

Chapter 1: Introduction - This section provides a background that establishes the relevance for the study within the context of previously published research

Chapter 2: Literature review - This section includes a comprehensive theoretical framework description of the related literature in the field and develops the theoretical framework for the study.

Chapter 3:Methodology - This section is a detailed discussion of the research process

Chapter 4:Analysis and Results - This section includes a comprehensive description of all research results and data and

Chapter 5: Conclusions and Recommendations - This section provides an interpretive critique and discussion of the results of the study.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

As explained in chapter 1, the main objective of this study is to compare various methods used for forecasting share prices. This Chapter describes, compares, contrasts and evaluates the major theories, arguments, themes, methodologies, approaches and controversies in the scholarly literature on the subject of this study. A general overview of share price forecasting methods and forecasting efficiency models are presented and evaluated.

2.3Forecasting Models

There are several forecasting models and methods in practice. The popularity of a forecasting model as against another is solely based on their risk metrics. Forecasting of stocks is generally believed to be a very difficult task. The use of untranslated raw data for stock market forecasting has been widely established using different methods such as fuzzy logic, artificial neural network, text mining, rough set, genetic algorithms and time series analysis amongst others. Common time series forecasting models are as follows: Box – Jerkin's Methods, Holt Winters Exponential Smoothing and simple linear regression. The result always favours one of the methods of forecasting and this can be ascertained by the use of error metrics.

Effective stock forecasting is usually an investor's delight as investors wish to be sure whether the stock they are putting their money on, is able to pay back its liabilities, has enough working capital and is generally in good state of financial health. It is suggested that a potential investors need would be met when the information that could enable him make a good and wise investment decision as regards stocks to purchase or invest in is easily accessible, available, understandable and reasonably rational.

2.3.1 Artificial Neural Network and Time Series Model

Since the early 90's, when the first practically usable Artificial Neural Network (ANN) types emerged, it had been rapidly used in different fields of sciences. The areas in which ANNs have been successfully implemented are speech, handwritten character, finger print and electrical recognition (Sejnowski, *et al*, 1990). Other areas are prediction of bank failure, stock market performance and pattern recognition (Tam, *et al* (1992). Further areas of application include system identification and control (vehicle control, process control), game-playing and decision making (chess, racing), medical diagnosis, financial application and data mining. A large number of studies have been reported in the literature with reference to the use of ANN in modeling stock prices in advanced countries. E. Birgul*et al* used ANN and Moving Average to predict Turkey's Stock Market (Birgul,*et al*). S.I Wu and H. Zheng developed recurrent neural networks to forecast the daily closing prices of these stock indexes (Wu,*et al*, 2003). K. Kim and W. Lee also used ANN with optimal feature transformation to predict stock market trends. In their study, a genetic algorithm (GA) was incorporated to improve the learning and general ability of ANNs for stock market prediction. The results obtained by a feature transformation method using the GA when compared to other two feature transformation ANN methods without GA showed that the performance of the improved ANN-GA model was better. Cirna, *et al*, (2010) introduced a genetic programming technique called Multi-Expression programming for the prediction of two stock indices and compared the performance with an artificial neural network trained using Levenberg-Marquardt algorithm, support vector machine andTakagi-Sugenoneuro-fuzzy model (Cirna,*et al*, 2010). The use of ANN in stock prediction comes in because ANN can learn to detect complex pattern in data. As ANNs are essentially non-linear statistical models, their accuracy and predictive capabilities can be both analytically and empirically tested. The performance of ANNs has been extensively compared to that of various statistical methods within the areas of prediction (Ripley, (1994). In particular, a fair amount of literature has been generated on the use of ANNs in time series forecasting. Lapedes and Farber concluded that basic neural network substantially outperform conventional statistical methods (Lapedes, *et al* 1984). J. M. Hutchinson used time series analysis to measure stock market performance (Hutchinson, 1994).It was noted that researchers on stock price prediction used untranslated data.

An artificial neural network attempts to model the human brain which is composed of neurons and connections between the neurons called synapses. A neural network mimics the brain in that observed knowledge is acquired through a learning process and interneuron connection strengths (weights) store the knowledge. (Haykin, 1994). Prior knowledge may be embedded in the network through utilizing a pre-trainednetwork within the larger network.

A neural network is comprised of four parts: nodes, connections between nodes, activation functions, and a learning rule. Each neuron collects data from the weighted sum of other neurons. The activation function of the neuron is applied to this sum. An activation function can be different for each neuron; however, this may become confusing, therefore the code used in this thesis has only one activation function in the network. There are three common activation functions: the hard-limiter, the threshold function, and the sigmoid function. The hard - limiter creates an output of positive one, negative one, or zero as in the McCulloch-Pitts model. Haykin, (1994).

In 1943, McCulloch and Pitts presented an early artificial neuron model known as the linear threshold gate. This neuron has multiple inputs and one output. The neuron produces a binary output used to group the input set. The threshold and weight are fixed and the model is simple. A disadvantage of this model is the inability of the network to work with nonlinearly separable classes.

Yoon and Swales (1991) considered multiple Discriminant analysis (MDA) models and ANN to forecast stock price. MDA is a set of simultaneous equations. Data were gathered from two information sources: The Fortune 500 and Business Week's "Top 100". A stock's total return and market valuation were the error metrics used to evaluate the models. MDA model was 74% correct on the training set and 65% correct on the test set. Artificial Neural network was 91% correct on the training set and 77.5% correct on the test set. This showed that neural network is better than MDA. In addition to this result the authors have stated that MDA has the capability to explain the feature and significance of each input parameters.

Kudyba, (1998) compared neural net-based computer algorithm with Semtsa (structural econometric model time series analysis). Data about electricity demand in US from 1945 to 1990 were considered for the comparison of models. Neural network was developed using all the information used to develop the Semtsa model. Measures used were adjusted R-square in conjunction with t- and F-statistics. Also root mean square error (RMSE) and the mean absolute percent error (MAPE) were used to compare the accuracy of each model. The result showed that the neural network for electricity demand is better than complex Semtsa. It was also seen that Semtsa was more expressive than neural network in explaining outputs.

Gruca, *et al* (1999) compared artificial neural network withMultiplicative Competitive Interaction (MCI) models using sales data of coffee and A.C. Nielsen catsup. (MCI model is a set of equations and parameters.) Data were obtained from Sioux Falls,SD. The coffee dataset contained 52 weeks data from March 1981 to April 1982. This data set was split into two samples: 43 weeks data as estimation sample and 9 weeks data as hold-out sample. The catsup data was of 156 weeks from August 1985 to August 1988. This data set was also separated into two samples: a 146 week data as estimation sample and a ten week data as holdout sample. Coffee dataset was small dataset with only few observations and catsup dataset was large dataset with enough data. The mean absolute percentage error (MAPE) was used to compare models. The results showed that ANN is more accurate than MCI.

Moshiri, *et al*(2000) compared different types of Back Propagation NeuralNetwork (BPN) with econometric models using data about inflation rate. Different BPN considered were: BPN, BPN with ARIMA, and BPN with VAR model. Different time series models were considered as econometric models. These were: an ARIMA model, a vector autoregression (VAR) model, and Bayesian vector autoregression (BVAR) model. Monthly data from 1973:1 – 1994:12 about inflation rate, GDP gap, money supply, and import price inflation were obtained from the CANISM databank. The dataset was divided into two sets: data from 1970:1-1990:12 as training set and data from 1991:1 – 1994:12 as test set. The Root mean squareErrors (RMSE) and Mean Absolute Errors (MAE) were the metrics used for calculating forecast quality. Information test method was also used to compare usefulness of models according to information content. The test result showed hybrid BPN were similar or better than their equivalent econometric model in dynamic forecasting.

Camargo, *et al* (2009) undertook a comparative study between artificial Neural Networks and ARIMA model using 8 years sales data collected from a medium sized enterprise in Brazil.

The data were from January 2000 to December 2008. MAPE and residual variation were the performance metrics used. The results of the comparison were as follows shown in the Table 1,

Table 2.1:Values of Observed Metrics

FORECASTING MODEL	VALUE OF THE OBSERVED	
	RMSE	**MAPE**
ARIMA	0.1235	0.6812
NEURAL NETWORK	0.1027	0.4765

Source: (Dinesh Bajracharya, 2010)

Experiments showed that artificial neural network adjusted well with the sales data and provided satisfactory forecast.

Tjung,*et al* (2010) compared neural network with the regression model (OLS method was used to estimate parameters of regression model) using financial stock data. Eight explanatory variables were used for forecasting financial stock prices. SPSS program was used to create unique regression model and AlyudaNeuroIntelligence program was executed to create neural network model. The mean and standard deviation of the % error were the evaluation metrics used to compare models. Authors also calculated Adjusted R-square value. The result of comparison was that neural network is more accurate than OLS. The accuracy of neural network was 96% while accuracy of OLS was only 68%. The study also showed some difficulties with neural networks. Neural network is complex; it requires more training time to find the best model. It has problem of over-fitting and it cannot be assured perfectly that the model created is the best because it is a blind search.

2.3.2 <u>Time Series Forecasting Models</u>

The Box-Jenkins approach to modeling and forecasting time series data is but one of a large family of quantitative forecasting methods which have been developed in the fields of operations research, statistics, and management science. Box-Jenkins models are also known as "ARIMA" models, the acronym standing for Autoregressive Integrated Moving Average. This terminology is made clear in the following sections. Exponential smoothing, linear regression, Bayesianforecasting and generalized adaptive filtering are some of the other techniques which are termed "extrapolative" forecasting (Makridakis *et al*., 1982).

Many of these methods have a common element; they utilize only the previous values of a series of numbers to forecast the future values of interest. Hence, they are referred to as univariate models, since the values from a single variable are used to predict the future values of the same variable. This is in contrast to multivariate models, where the variable of interest is also considered to depend on other variables.Examples of real application arepresented in Table 2.

Table 2.2: A List of Examples of Real Applications

Dataset	Forecast horizon	Benchmark	Reference
Univariate ARIMA			
Electricity load (min)	1–30 min	Wiener filter	Di Caprio, Genesio, Pozzi, and Vicino - 1983
Quarterly automobile insurance paid claim cos	8 quarters	Log-linear regression	Cummins and Griepentrog (1985)
Daily federal funds rate	1 day	Random walk	Hein and Spudeck (1988)
Quarterly macroeconomic data	1–8 quarters	Wharton model	Dhrymes and Peristiani (1988)
Monthly department store sales	1 month	Simple exponential smoothing	Geurts and Kelly (1986, 1990),Pack (1990)
Monthly demand for telephone services	3 years	Univariate state space	Grambsch and Stahel (1990)
Yearly population totals	20–30 years	Demographic models	Pflaumer (1992)
Monthly tourism demand	1–24 months	Univariate & multivariate state space	du Preez and Witt (2003)
Dynamic regression/transfer function			
Monthly telecommunications traffic	1 month	Univariate ARIMA	Layton, Defris, and Zehnwirth (1986)
Weekly sales data	2 years	n.a.	Leone (1987)
Daily call volumes	1 week	Holt–Winters	Bianchi, Jarrett, and Hanumara (1998)
Monthly employment levels	1–12 months	Univariate ARIMA	Weller (1989)
Monthly and quarterly consumption of natural ga	1 month/1 quarter	Univariate ARIMA	Liu and Lin (1991)
Monthly electricity consumption	1–3 years	Univariate ARIMA	Harris and Liu (1993)
VARIMA			
Yearly municipal budget data (in-sample)	Yearly	Univariate ARIMA	Downs and Rocke (1983)
Monthly accounting data	1 month	Regression, univariate, ARIMA,transfer fu	Hillmer, Larcker, and Schroeder (1983)
Quarterly macroeconomic data	1–10 quarters	Judgmental methods, univariate ARIMA	Oller (1985)
Monthly truck sales	1–13 months	Univariate ARIMA, Holt–Winters	Heuts and Bronckers (1988)
Monthly hospital patient movements	2 years	Univariate ARIMA, Holt–Winters	Lin (1989)
Quarterly unemployment rate	1–8 quarters	Transfer function	Edlund and Karlsson (1993)

Source: (De Gooijer and Hyndman, 2006)

2.3.3 <u>Text Mining</u>

The area of Knowledge Discovery in Text (KDT) and Text mining (TM) is fast growing mainly because of the strong need for analyzing the vast amount of textual data that reside on internal file systems and the web, (Karanikas and Theodoulidis, 2002).

In today's information age, we have witnessed and experienced an ever increasing flood of information. The Internet makes available a tremendous amount of information, on an amazing variety of topics that has been generated for human consumption. Unfortunately, the hundreds of millions of pages of information make it difficult to easily identify or find information of interest to specific users or useful for particular purposes. The amount of text is simply too large to read and analyze early. Furthermore, it changes constantly, and requires ongoing review and analysis if one wants to keep abreast of up-to-date information. Working in this ever-expanding sea of text becomes extremely difficult, (Wen, 2001).

As stated by Grobelnik*et al* (2000) with the emergence of the World Wide Web, there is a need for extending the focus to mining information from unstructured and semi – structured information sources such as on – line feed, corporate archives, research papers, financial reports, medical records and e-mail messages amongst others.

While the amount of textual data available is constantly increasing, ability to understand and process this information remains constant. According to Tan (1999), approximately, 80% of information of an organization is stored in unstructured textual forms such as reports, e-mails. The need for automated extraction of useful knowledge of huge amount of textual data in order to assist human analysis is fully apparent. Merrill Lynch, 2000 cited by Karanikas and Theodoulidis, (2002).

Table 2.3: Examples of Textual Financial Data

Textual Source	**Types**	**Example**	**Description**
Company Generated Sources	SEC Reports	8K	Reports on significant changes
		10K	Annual reports
Independently Generated Sources	Analyst Created	Recommendations	Buy/Hold/Sell assessments
		Stock alerts	Alerts for share prices
	News Outlets	Financial Times	Financial News stories
		Wall Street Journal	Financial News stories
	News Wire	PRNews Wire	Breaking financial news articles
		Yahoo Finance	45 financial news wire sources
	Discussion Boards	The Motley Fool	Forum to share stock-related Information

Source: (Schumaker,*et al*, 2006)

The knowledge map with respect to text mining as a way of predicting stock prices is illustrated below in table 4:

Table 2.4: Article Related to the Prediction of Stock Market Using Text Mining

Articles	Authors
Daily Stock Market Forecast from Textual Web Data	Wuthrich, 1998
Activity Monitoring: Noticing Interesting Changes in Behavior	Fawcett, 1999
Electronic Analyst of Stock Behaviour (Ǽnalyst)	Lavrenko, 1999
Language models for Financial News Recommendation	Lavrenko, 2000
Mining of Concurrent Text and Time Series	Lavrenko, 2000
Integrating Genetic Algorithms and Text Learning for Prediction	Sycara*et al*. 2000
Using News Articles to Predict Stock Price Movements	Gidofalvi, 2001
News Sensitive Stock Trend Prediction	Fung *et al*. 2002
Stock prediction: Integrating Text Mining Approach Using News	Fung *et al*. 2003
Forecasting Intraday Stock Price Trends with Text - mining	Mittermayer, 2004
Stock Broker P – Sentiment Extraction for the Stock Market	Khara*et al*, 2004
The Prediction Power of Textual Information on Financial Markets	Fung *et al*. 2005
Text Mining for Stock Movement Prediction – A Malaysian Approach	Phung, 2005
Textual Analysis of Stock Market Prediction Using Financial News	Schumaker, 2006

Source: (Falinouss, 2007)

2.3.4 Properties of Error Metrics

Armstrong (1985) has mentioned several observations about the error metrics. Some of them are:

i. **MAPE** is biased if the data series contains only positive numbers and it favors low forecast

ii. **RMSE** is strongly influenced by the scale of series and is unreliable if data contains outliers

iii. **Adjusted MAPE** or similar error metrics are more reliable than MAPE

According to Decision 411 forecasting (2010), there are no absolute criteria for a "good" value of RMSE or MAE as they depend on the units of variable and degree of forecasting accuracy.RMSE is always greater than MAE. If the difference is great, then there will be great variance in the individual errors in the data. If RMSE equals MAE then all the errors will be same.The fluctuations in data are taken in account by WMAPE and if the fluctuations in data are smallWMAPE simply turns into the ordinary MAPE (Schutz, *et al*, 2011)

In 1981, a survey found RMSE was more popular than MAPE, 48% of 62 academics and 33% of61 practitioners used RMSE while only 24% of academics and 11% of practitioners used MAPE.But a decade later MAPE was found to be the most commonly used metric (52%) compare toRMSE (10%), Armstrong, (1985).

Shu Chang and Burn D.H. (2003) forecasted flood frequency using ensemble of ANN and singleANN. The accuracy of ensemble ANN was found better than single ANN and it was found that ensemble is less sensitive to the choice of initial parameters. The evaluation metrics used for the comparison between ensemble of ANN and single ANN were relative squared error (RSError), percent relative error (PRError) and relative bias (RBias). Further they compared ensemble ofANN with multiple regression and found that ensemble is better than it. The major finding of their experiment was that properly designed ensemble of ANN is better than single ANN and multiple regression model.

Error metrics can affect the ranking of forecasting methods as they are affected by several factors (Armstrong, 1985). So it is obvious that the rank of one forecasting method differs from the rank of another forecast method on the same data set. Some of the factors that influence the error metrics are as follows:

a. Scale of data
b. Nature of data
c. Outliers in the data

If infrequent large values are not a problem in decision situation then the MAE or MAPE be more relevant error metric than RMSE (Decision 411, 2010).

CHAPTER 3

METHODOLOGY

3.0 Introduction:

This chapter describes various concepts/terminology/techniques which are useful for understanding the problem, choosing appropriate techniques/models and carrying out analysis. It includes types of forecasting, modeling techniques and brief description of forecasting models used for optimization of error metrics.

3.1 Theoretical Framework

This section describes compares and evaluate major theories, themes and approaches subject of this study. According to Mun, (2010) there are two types of forecasting techniques namely: Quantitative and Qualitative forecasting. There are categorised on the basis of knowledge, experiences, judgment, historical data and statistics.

Qualitative Forecasting

The qualitative forecasting techniques are use when historical and comparable data is not available. These techniques are based on the expert's knowledge, opinions and market research etc. The techniques use the secondary information to create a model which predicts future values. These techniques are also used where the historical data is available but due to unexpected circumstances the use of the data cannot be trusted.

Quantitative Forecasting

Quantitative forecasting techniques only use historical data to create forecast models. These techniques try to find relationships between the dependent variable and one or more independent variables and then use these relationships to forecast values of the dependent variable.

3.2 Theories of Stock Market Forecasting

Stock markets have been studied over and over again to extract useful patterns and predict their movement (Hirshleifer and Shumway, 2003). Stock market prediction is not only a concern to financial market analysts and researchers, but also to investors. Whiles researchers want to minimize stock market short term prediction margin of error, investors are concerned with their future portfolio value. This can be done using different forecasting techniques as applied by market analysts. Two most used theories in stock market forecasting will be explained briefly. A clear and more precise distinction between the two conventional theories will be provided in this section.

There are two main theories available for forecasting future prices of stock market security. These are

a. Efficient Market Hypothesis (EMH), first introduced by Fama (1964) and
b. The Random Walk Theory (RWT). (Malkiel, 1996).

The following sections describe the distinction between these two main common theories:

Efficient Market Hypothesis (EMH)

The efficient market hypothesis (EMH) is a backbreaker for forecasters. In its crudest form, it effectively says that the series we would very much like to forecast, the returns from speculative assets, are unforecastable. This is a venerable thesis, its earliest form appearing a century ago as the random walk theory (Bachelier, 1964). This theory was confirmed empirically in the 1960s (Cootner, 1964) and many times since. Soon after the empirical evidence appeared, the EMH was proposed, based on the overpowering logic that if returns were forecastable, many investors would use them to generate unlimited profits. The behavior of market participants induce returns that obey the EMH, otherwise there would exist a 'money-machine' producing unlimited wealth, which cannot occur in a stable economy.

Intellectually, that might appear to be the end of the story. However, despite the force of the argument, it seems not to be completely convincing for many forecasters. Everyone with a new prediction method wants to try it out on returns from a speculative asset, such as stock market prices, rather than series that are known to be forecastable. Papers continue to appear attempting to forecast stock returns, usually with very little success.

Definitions of Market Efficiency

Jensen (1978) defines market efficiency as follows

A market is efficient with respect to information set Xt if it is impossible to make economic profits by trading on the basis of information set Ω_t.

A closely related definition of market efficiency is provided by Malkiel (1992).

A capital market is said to be efficient if it fully and correctly reflects all relevant information in determining security prices. Formally, the market is said to be efficient with respect to some information set, Ω_t, if security prices would be unaffected by revealing that information to all participants. Moreover, efficiency with respect to an information set, Ω_t, implies that it is impossible to make economic profits by trading on the basis of Ω_t.

Three points are emphasized in these definitions, namely

i. the importance of the information set adopted in the test, Ω_t;

ii. the ability to exploit this information in a trading strategy; and finally

iii. that the yardstick for testing if the EMH holds is measured in economic (i.e. risk-adjusted and net of transaction costs) profits.

3.3The Information Set

Three forms of market efficiency are commonly entertained in the EMH literature based on the set of variables contained in the information set, Ω_t .(Roberts (1967) and Fama (1970)). If Ω_t only comprises past and current asset prices (as well as possibly dividends and variables such as trading volume), the EMH in its weak form is being tested. Expanding Xt to include all publicly available information gives rise to the EMH in its semi-strong form. Finally, if all public and private information is included in Ω_t, market efficiency in the strong form is being tested.

Most studies in the literature on predictability of stock market returns test the EMH in its weak or semi-strong form. For example, papers on the predictive performance of technical trading rules test weak form market efficiency since only past prices and maybe volume information are used as predictor variables. Studies that include an extended set of predictor variables such as default premia, term spreads and other business cycle indicators test semi strong efficiency. Restricting the information set in this way is designed to rule out private information that is harder to measure and perhaps also more expensive to acquire. For example, it is not usually asserted that a market is efficient with respect to inside information since this information is not widely accessible and hence cannot be expected to be fully incorporated in the current price. Strong form efficiency can be tested indirectly, for example by considering the performance of fund managers and testing if they manage to earn profits net of risk premia after accounting for the cost of acquiring private information. Surveys of market efficiency such as Fama (1970, 1991) have focused on testing informational efficiency. Fama (1970) concludes that the empirical evidence is largely supportive of weak form and semi-strong form efficiency, while Fama (1991) reports stronger evidence of predictability in returns based both on lagged values of returns and publicly available information.

The contribution of Fama, with respect to efficient market hypothesis, is very significant and cannot be overlooked. The Efficient Market Hypothesis states that at any given time, security prices fully reflect all available information. Making the assumption that capital markets are efficient, stock prices should always be reflecting the present value of future cash flows (Fama, 1965; Kothari, 2001; Ou& Penman, 1989). That is if the price would be unaffected by revealing the information set to all market participants (Malkiel, 1992). However there are different kinds of information that influence the security values. The analysis of the information available can be used to give indications about mispriced stocks (Abarbanell&Bushee, 1998). This implies that without market information, forecasting of market prices is virtually impossible. Fama's theory breaks the Efficient Market Hypothesis into three: Weak, Semi – Strong and Strong. (Schumakerand Chen, 2006). This implies that the kind of information available is the sine qua non for distinguishing between efficient market hypothesis versions.

Weak Efficient Market Hypothesis

Weak EMH considers the past price of market stock and historical information as the only component of the current price. This implies that weak EMH rules out any form of forecasting based on the price data alone, since prices follow random work where successive changes have zero correlation. **Semi –** ***Strong Efficient Market Hypothesis***

In the case of Semi – Strong EMH, the information needed is just a buildup of the Weak EMH which incorporates all historical and current public information into the price.

Strong Efficient Market Hypothesis

Strong EMH requires more information with respect to forecasting market security than the Weak and Semi – Strong EMH respective. Whiles Weak and Semi – Strong EMH only require the past price of market security and historical and public information, Strong EMH requires historical, public, and private information including insider information on the stock price. This form means that prices on the stock market fully reflect all information about company value, both publicly available and unannounced.

The weak and semi – strong form of EMH has been fairly supported in a number of research studies (Low and Webb 1991; White, 1998). However, most recent published reports suggest that the EMH is defeated on the scale of balance. Fama, (1991). In his article “Efficient Capital Market” Fama (1991) states that the efficient market hypothesis surely must be false. Due to the shortage of data, the strong form of EMH has been difficult to be tested.

Lee (2001) believes that market efficiency is an inadequate and naive notion. He furthermore stated that stock prices do not adjust to their fundamental value instantaneously, but that the adjustment is rather a process requiring time and effort.

3.4 Random Walk Theory

A different perspective on forecasting comes from the random walk Theory. (Malkiel, 1996).Random walk theory gained popularity in 1973 when Burton Malkiel wrote "A Random Walk Down Wall Street", a book that is now regarded as an investment classic. Random walk is a stock market theory that states that the past movement or direction of the price of a stock or overall market cannot be used to forecast its future movement. Originally examined by Maurice Kendall in 1953, the theory states that stock price fluctuations are independent of each other and have the same probability distribution, but that, over a period of time, prices maintain an upward trend. Random Walk Theory is very similar in theoretical thrush to semi – strong EMH where all public information is said to be available to all. However, Random Walk Theory declares that even with such information, future forecast is ineffective.

Approaches to Stock Market Forecasting

Out of the two theories (i.e. EMH and Random Walk) discussed so far, two distinct conventional approaches to financial market forecasting philosophies are unveiled, namelyTechnical approach and Fundamental approach.

A clear distinction between the two forecasting philosophies approach or approaches is provided below.

<u>Technicians Trading Approach</u>

The term technical analysis denotes a basic method of stock investing where the past prices are studied, using charts as the basic tool. Technical analysts also referred to as the "technicians" argue that their approach allows them to profit from changes in the psychology of the market. The following statement expresses this view:

"*The technical approach to investment is essentially a reflection of the idea that prices move in trends which are determined by the changing attitudes of investors toward a variety of economic, monetary, political and psychological forces…Since the technical approach is based on the theory that the price is a reflection of mass psychology ("the crowd") in action, it attempts to forecast future price movements on the assumption that crowd psychology moves between panic, fear, and pessimism on one hand and confidence, excessive optimism, and greed on the other*" Pring (1991, pp. 2-3)

Many hundreds of methods for forecasting of stock prices have been developed and are still being developed on the foundation of these principles (Hellmstrom and Holmstrom, 1998). Technical analysis was first developed in the context of the stock market; its exponents argue that it applies in one form or another to all asset markets. Technical analysis (Pring, 1991) is based on numerical time series data and tries to forecast the stock markets using indicators of technical analysis. Technicians believe that all reactions of the market to all sort of news are incorporated in the real – time prices of stocks. Based on this, it is very clear that technicians ignore news but rather are concerned with the identification of existing trend and anticipate the future trends of the stock market from charts. However charts or numeric time series data only contain the event and not the reason why it happened (Kroha and Baeza – Yates, 2004). Technicians utilize charts and modeling techniques to identify trends in price and volume. They rely on historical data to forecast future outcomes. (Schumaker and Chen, 2004). Since the early 1970s, when floating of exchange rate became prominent, foreign currency dealers have widely incorporated this approach to trading. At least some technicians clearly believe that the foreign exchange market is particularly prone to trending.

"*Currencies have the tendency to develop strong trends, stronger than stocks in my opinion because currencies reflect the performance of countries*" (Gand, *et al*, (2010 p. 84))

"*It has been our longstanding experience that nothing trends as well or as clearly as a major currency market — not equity market indices, not commodity markets and not even long-term Treasuries*" (Zimmermann*et al*, (2010 p. 197))

Even though, the technical approach has been used for quite some time now, Park and Irwin (2007) concluded that technical analysis is profitable in foreign exchange and commodity futures markets but not in stock markets. This same conclusion was arrived at by Silber (1994).

Many promising forecasting methods to forecast stock market movements from numeric time series such as autoregressive, moving average, simple linear regression and exponential smoothing among others are some of the famous stock trend forecasting techniques which have dominated the time series for some years now. A book named "Technical Analysis from A to Z" by Achelis, 1995, outlines most common technical indicators.

In summary, Technical analysis does not try to gain deep insight into a company's business. It assumes the available public information does not offer a competitive trading advantage. Instead, it focuses on studying a company's historical share price and on identifying patterns in the chart. The intention is to recognize trends in advance and to capitalize on them.

Fundamentalist Trading Approach

Technical analysis looks at the price movement of a security and uses this data to predict its future price movements whereas fundamental analysis looks at economic factors, known as fundamentals.

This approach tries to identify promising companies by analyzing their fundamental attributes. These include characteristics such as financial results, growth forecasts and anticipated product development. It is important to note that this type of analysis is not static; newly released financial information, corporate announcements and other news can influence the fundamental outlook of a company. Fundamental analysis requires expertise in a particular sector and is often conducted by professional analysts. Their investment recommendations are regularly updated and published.

3.5 Error Metrics

Forecasting models need to be evaluated from different perspectives. These include determining how much forecasted values deviate from actual values; whether the model used to forecast is useful or not; and the strength of linear relationship between dependent and independent variables.

Armstrong (1985) has found that the use of only one metrics for the comparison of a forecasting model is not suitable as the result of different metrics differ. He has suggested the use of multiple error metrics for testing forecasting models. He added that, the result of one metric may not be reliable and above all RMSE is one of the worst metrics. One metric may perform well in one situation but the same metric may not perform that well in another situation.

Forecasted values obtained from different forecasting models and methods may differ. The error metrics show how risky the forecast model or method is. The model is tested by taking difference between the actual value and the forecast value. The smaller the difference, the better the model is. Several criteria can be used to compare different forecasting models. According to (Zeng, *et al*, 1998), some methods for evaluating forecasting models are as follows:

a. Root Mean Square Error (RMSE)
b. Mean Absolute Error (MAE)
c. Mean Absolute Percentage Error (MAPE)
d. Weighted Mean Absolute Percentage Error (WMAPE)

Rummel R.J. (1976) has considered linear correlation as the workhorse of quantitative research and analysis. This motivates to consider linear correlation in comparing models (odd). In this thesis, three different error metrics are considered for the evaluation of forecasting models. They are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE) and linear correlation.

3.6 Data Collection

In a research project there are two types of data; primary and secondary data. Primary data is data collected solely for the research. Secondary data is data that is collected earlier for another purpose. (Eriksson ochWiedersheim-Paul, 2001)

Data from the Oil and Gas Producers in London Stock Exchange was conveniently sampled out of forty one (41) industries. The time horizon of the data is seven years spanning from January, 2005 to December, 2011. The data is a secondary data and can be accessed from Yahoo Finance. The historical adjusted close prices of the respective stock in the Oil and Gas on the London Stock Exchange will be used in this thesis.

3.7 Sampling

The sampling method in this thesis is convenient sampling which is a non-probabilistic sampling. In this thesis, the Oil and Gas industry wasselectedby the simple random sampling technique.

3.8 Forecasting Models Description

According to Wang, *et al*, (2003), thefollowing are the most commonly used forecasting models:Logistic Regression

a. Linear Regression
b. Autoregression
c. Autoregression Moving Average (ARMA)

However, in this thesis the Autoregression Integrated Moving Average with/without an exogenous variable (ARIMA) model will be used.

3.8.1Time-series Models

In time series model, the past behavior of a time series is examined to infer something about its future behavior instead of searching for effect of one or more variables on the forecast variable. Howrey (1980) stated that the time series models are based on a relatively weak nonparametric formulation of the model. They put more emphasis on the data analysis for simplification of the model.

Different patterns or trends can be seen in the time series data. The time series is influenced by several factors like random components, seasonal components, cyclic components etc. The random component in the time series may shield the influence of other components and make it difficult to describe the observed trends or patterns in the data. Removal or reduction of random components from the time series will result in better forecasting or interpretation of series. To remove or reduce the effect of the random component from the time series, smoothing techniques are used. Smoothing of the time series is done before forecasting. Smoothing techniques like moving average and exponential smoothing can remove random components and seasonal components from the time series data (McClave,*et al*, 1998).

The Random Walk Model

Random walk is a simplest model containing stochastic trends given by following equation:

$$Y_t = Y_{t-1} + z_t$$

Here, u_t is called error term or white noise. In the simplest random walk process, future value of time series is given by its immediate previous (one step back) value.

ARIMA model is subset of univariate model in which time series is expressed in terms of past values of it, current and lagged value of a 'white noise' or error term. ARIMA models do not assume any knowledge about underlying economic model or structural relationships between variables (Meyler*et al*, 1998). ARIMA model is formed by combining two models: Autoregressive model and Moving Average model.

a. ***Autoregressive Model***

Autoregressive model represents current value of time series as combination of one or more previous values of the series. It shows the dependency of one value with its nearest previous values. Autoregressive process is a difference equation determined by random variables (difference equation shows current value of series as function of its previous values).

Autoregressive model has order term, *p* that determines how many previous values are to be included in the difference equation to estimate current value. A difference equation relates a variable X_t at time *t* with its previous values (Horvath, *et al*, 2006).

The autoregressive AR (1), *p* = 1, includes only one previous value. It is a standard linear difference equation and written as:

$$Y_t = \phi Y_{t-1} + z_t \text{, t = 0, ± 1, ± 2…} \qquad (1)$$

Where, u_t is error term and ϕ is parameter to be estimated.

The p^{th} order AR time series is AR (p) and is given by the following expression:

$$Y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + z_t \quad t = 0, \pm 1, \pm 2, \qquad (2)$$

Where, $\phi_0, \phi_p \neq 0$, and z_t are uncorrelated random variables.

Using difference equation, value of Y_t can be obtained from Y_{t-1}, value of Y_{t-1} is obtained from Y_{t-2} and so on.

The AR (1) model is fitted with collected data by first estimating value of ρ. To estimate value of ρ least squares estimation method is used. It minimizes the sum of square of errors for the observed values with respect to ρ.

$$\frac{\partial}{\partial x}\sum_{t=2}^{n}(Y_t - \phi Y_{t-1})^2 = 2\sum_{t=2}^{n}(Y_t - \phi Y_{t-1})(-Y_{t-1}) \qquad (3)$$

Equating above equation (3) to zero and solving it further gives the value of least square estimator for ϕ:

$$\hat{\phi} = \frac{\sum_{t=2}^{n} Y_t Y_{t-1}}{\sum_{t=2}^{n} Y^2{}_{t-1}}$$

From estimated value of ϕ, distribution of error terms can be found.

$$\hat{z}_t = Y_t - \phi Y_{t-1}$$

Now using estimated value of ϕ and distribution of error data, the model can be fitted using equation (2).

b. *<u>Moving Average Process of order q, MA (q)</u>*

A time series is influenced by random shocks in noisy environment. As a result current value of series is affected by the random shocks appeared in previous values. Moving average terms are used to capture the influence of previous random shocks in the future value.

First order moving average or MA (1) is a simple time series, given by

$$Y_t = z + z_t + \alpha z_t \qquad (4)$$

This equation says, apart from mean, z , Y_1 is a weighted average of z_2 and z_1, Y_2 is a weighted average of z_2 and z_1 etc. The value of Y_t is defined in terms of random shocks z_t .

A Moving average of order *q*, MA (*q*) process X_t , is given by

$$Y_t = z_t + \theta_1 z_{t-1} + ... + \theta_q z_{t-q} \qquad (5)$$

Above equation (5) representing MA (*q*) process is always stationary**.** In fact MA process is inverse of AR model. The MA model is invertible if an MA model can be expressed as autoregressive (infinite order) model

c. *Autoregressive moving average process, ARMA (p, q)*

Autoregressive Moving Average model is formed by combining terms of AR and MA models. Autoregressive model or Moving Average can be used to approximate any stationary process with any degree of accuracy as desired.

Combining equation (2) and (5), ARMA model of order *p* and *q* is formed,

$$Y_t = \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + z_t + \theta_1 z_{t-1} + ... + \theta_q z_{t-q} (6)$$

d. *Autoregressive Integrated Moving Average Process - ARIMA (p, d, q)*

The ARMA model assumes that the time series data is stationary (that is statistical properties of data do not change over time). But the real data are not stationary in nature. Time series data is made stationary by differencing process. The first order differencing process of time series X_t is defined as $X'_t = X_t - X_{t-1}$. ARMA time series which is made stationary by differencing process is known as Integrated Autoregressive Moving Average (ARIMA) model. ARIMA model is represented by three parameters: *p* order of autoregressive model, *d* order of differencing, and *q* order of moving average model.

ARIMAmodel takes historical data and decomposes that data into an autoregressive (AR) process which maintains memory of past events, an Integrated (I) process which makes data stationary for easy forecast and a Moving Average (MA) process of forecast errors. It does not suffer from existence of serial correlation between the error residuals and their own lagged values.

An ARIMA (*p, d, q*) model can be checked if it is a good statistical fit for data or not, using Akaike Information Criterion (AIC) and Schwarz Criterion (SC) method. Autocorrelation (AC) and partial autocorrelation (PAC) statistics help to determine the right parameters for ARIMA model (Real Options Valuation, 2007).

Box-Jenkins has specified four stages for ARIMA model selection (Kahforoushan,*et al*, 2010). These are:

- Determining values of p, d, q
- Estimate parameters of the model
- Checking whether considered model fits data properly or not, if not consider another model

- Estimation using the best selected model.

ARIMA model has two main advantages, namely

- It requires only data on the time series in question and this is advantage in case of forecasting large number of time series.
- No problem of timelines of data

The Disadvantages of ARIMA model are:

- Model identification may require skill and experience of the forecaster
- No underlying theoretical model or structural relationships is assumed.

According to (Kmenta, *et al*, (1980), time series models are developed with little or no economic theory, so these models are not good in showing cause-and-effect relationship between different variables of the system under study.

e. *ARIMA (p, d, q) with Exogenous Variable*

An ARMAX model simply adds in the covariate on the right hand side:

$$y_t = \beta x_t + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} - \theta_1 z_{t-1} - \ldots - \theta_q z_{t-q} + z_t \qquad (7)$$

where x_t is a covariate at time t and β is its coefficient. While this looks straight-forward, one disadvantage is that the covariate coefficient is hard to interpret. The value of β is *not* the effect on y_t when the x_t is increased by one (as it is in regression). The presence of lagged values of the response variable on the right hand side of the equation mean that β can only be interpreted conditional on the value of previous values of the response variable, which is hardly intuitive.

If we write the model in (7) using backshift operators, the ARMAX model is given by

$$\phi(B)y_t = \beta x_t + \theta(B)z_t \text{ or } y_t = \frac{\beta}{\phi(B)}x_t + \frac{\theta(B)}{\phi(B)}z_t, \qquad (8)$$

where $\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - ... - \theta_p B^p$. Notice how the AR coefficients get mixed up with both the covariates and the error term.

3.8.2 <u>The steps in the ARIMA model-building</u>

STEP 1: ***Model Identification (Selecting an initial model)***

1.1 Determine whether the series is stationary or not by considering thegraph of ACF. If a graph of ACF of the time series values either *cuts off fairly quickly or dies down fairly quickly*, then the time series values shouldbe considered stationary. If a graph of ACF *dies down extremely slowly*,then the time series values should be considered *non-stationary*.

If the series is not stationary, it can often be converted to a stationaryseries by differencing. That is, the original series is replaced by a series ofdifferences. An ARMA model is then specified for the differenced series.

Differencing is done until a plot of the data indicates the series varies abouta fixed level, and the graph of ACF either cuts off fairly quickly or dies downfairly quickly.

The theory of time-series analysis has developed a specific language and aset of linear operators. According to equation (8), a highly useful operatorin time-series theory is the lag or backward linear operator (B) defined by

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

Model for non-seasonal series are called Autoregressive integratedmoving average model, denoted by ARIMA (p, d, q). Here p indicatesthe order of the autoregressive part, d indicates the amount of differencing,and q indicates the order of the moving average part. If the original seriesis stationary, d = 0 and the ARIMA models reduce to the ARMA models.

The difference linear operator (Δ), defined by

$$\Delta Y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1-B)Y_t$$

The stationary series W_t obtained as the dth difference (Δ^d) of Y_t ,

$$W_t = \Delta^d Y_t = (1-B)^d Y_t$$

ARIMA (p, d, q) has the general form:

$$\phi_p(B)(1-B)^d Y_t = \mu + \theta_q(B)\varepsilon_t \text{ or } \phi_p(B)W_t = \mu + \theta_q(B)\varepsilon_t$$

1.2 Once a stationary series has been obtained, then identify the form ofthe model to be used.

STEP 2: ***Model Estimation***

Estimate the parameters after a tentative model has been selected.

STEP 3: ***Model Checking***

In this step, model must be checked for adequacy by considering theproperties of the residuals whether the residuals from an ARIMA modelmust has the normal distribution and should be random. There are a large number of tests of randomness (e.g., the runs tests). Autocorrelation plots are one common method test for randomness. The Ljung-Box test is based on the autocorrelation plot. However, instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags. For this reason, it is often referred to as a "portmanteau" test. An overall check ofmodel adequacy is provided by the Ljung-Box Q statistic. More formally, the Ljung-Box test can be defined as follows.

H_0: The data are random.

H_1: The data are not random.

The test statisticQ is

$$Q_m = n(n+2)\sum_{k=1}^{m}\frac{r_k^{\,2}(e)}{n-k} \sim \chi^2_{m-r};$$

where

$r_k(e) =$ the residual autocorrelation at lag k

$n =$ the number of residuals

$m =$ the number of time lags includes in the test

If the p-value associated with the Q statistic is small (p-value $< \alpha$), the model is considered inadequate. The Ljung-Box test is commonly used in ARIMA modeling. Note that it is applied to the residuals of a fitted ARIMA model, not the original series.

Moreover we can check the properties of the residual with the followinggraph:

1. We can check about the normality by considering the normal probability plot or the p-value from the One-Sample Kolmogorov – Smirnov Test.
2. We can check about the randomness of the residuals by considering the graph of ACF and PACF of the residual. The individual residual autocorrelation should be small and generally within $\pm 2/\sqrt{n}$.

STEP 4: ***Forecasting with the Model***

Forecasts for one period or several periods into the future with the parametersfor a tentative model have been selected. Forecast can be in-sample or out-sample. In this thesis we are mindful of the in-sample forecast.

3.9. Preliminary Testing

a. ***The Complementary ADF Test for Stationarity***

Ahamada (2004) wisely tailored the cumulative sum of square (CSS) procedure in Inclán and Tiao (1994) to formulate a complementary test for the KPSS testing procedure (hereafter, complementary KPSS test). In the vein of Ahamada (2004), this study extends the application of the same CSS procedure in the case of ADF, yielding to the so called complementary ADF test. For compatibility, the current study follows closely the definitions and notations in Ahamada (2004).

Consider the following time series { y_t }, which is stationary around the level r_0:

$$y_t = r_0 + \varepsilon_t \qquad\qquad t = 1 \ldots T,$$

where ε_t is independent and identically distributed (i.i.d.) with a zero mean and constant variance, denoted $\varepsilon_t \sim$ i.i.d. (0, σ^2_ε).

The stationarity of { y_t }may be tested by the augmented Dickey-Fuller (ADF) test. ADF is the improved version of Dickey-Fuller (DF) test of the framework $t\Delta y_{t-1} = t\ \partial y_{t-1} + \omega_t$, where ω_t ~i.i.d. (0, σ^2_ε). Here, the null hypothesis of $\partial = 1$ (unit root) is tested against the alternative hypothesis of $\partial < 1$ (no unit root).

$$\Delta y_t = \partial y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \eta_t$$

where η_t ~ i.i.d.(0, σ^2_η), p is the autoregressive lag length large enough to eliminate possible serial correlation in η_t and ∂ is the coefficient of interest. Conventionally, if $\partial = 0$, the series contains a unit root implying nonstationary, whereas if $\partial < 0$, there is no unit root implying stationarity. In the ADF test, the null hypothesis of unit root, i.e.

H_0^{ADF}: $\partial = 0$ is tested against the alternative hypothesis of no unit root,

Ha^{ADF}: $\partial < 0$ using the t - test of individual significance.

b. ***Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test for Unit Root***

The (KPSS) test is a test where the null is the other way round: it test to see if a series can reject stationarity. Kwiatkowski *et al*. (1992). It assumes that the process for y can be written

$$Y_t = \delta t + \xi_t + \varepsilon_t$$

With an auxiliary equation for ξ_t

$$\xi_t = \xi_{t-1} + \mu_t, \text{with} \mu_t \sim iid\ (0, \sigma_u^2)$$

Such that it follows a random walk. A test of $\sigma_u^2 = 0$ is a test for stationarity. In general there are still size and power problems with this test in common with the Phillips – Perron test. However it represents a useful alternative hypothesis and may conflict with tests that assume non-stationarity as the null, thus indicating that there may be real doubt as to the properties of the data.

The hypothesis of their test is given below:

H_0: Y_t is level or trend stationary, I (0)

H_1: Y_t has a unit root

A *p*-value greater than 0.05 from the result of the KPSS test is enough to accept the null hypothesis at 5% level of significance.

c. *Model selection by comparing forecasts: Diebold and Mariano (1985)*

The Diebold and Mariano (1995) test compares the statistical significance of MAE, RMSE and MSE differences of two competing forecasting methods. The Diebold-Mariano statistic is simply the t statistic of the square error difference mean of two competing forecasting alternatives whose covariance matrix is estimated consistently by accounting for the autocorrelation introduced in multi-step forecasts. The Diebold and Mariano (1995) statistic, S_1 formula is

$$S_1 = \frac{1}{T}\sum_{t=1}^{T}\frac{[(e_{it})^2 - (e_{jt})^2]}{\sqrt{\frac{2\pi f(0)}{T}}}$$

where $e_i = y_i - \hat{y}_i$ and $e_j = y_j - \hat{y}_j$ are the forecast errors for observation t in two alternative models i and j, T is the sample size and $f(0)$ is the spectral density of the difference of the square prediction errors at frequency zero.

d. *Variance Ratio Tests by Lo and MacKinlay*

The variance ratio tests by Lo and MacKinlay were first proposed to test for a random walk in case of homoscedasticity and later extended to the more general case of an uncorrelated random walk in case of heteroscedasticity. This test utilises data sampled at various frequencies. Lo and MacKinlay demonstrated that variance ratio tests are statistically more powerful than the Box-Pierce Q-statistics.

As an important property of a random walk, the variance of its increments is linear in the observed period. Specifically, the variance estimated from the q-periods returns should be q times as large as the variance estimated from one-period returns, or:

$$\frac{Var(r_t^{\,q})}{Var(r_t)} = q$$

where

$r_t^{\,q}$ = Returns of a sample t for a the period with a length of q

r_t = Returns of a sample t with one-period length

The variance ratio VR (q) can be defined as:

$$VR(q)=\frac{Var(r_t^{\,q})}{qVar(r_t)}$$

The null hypothesis is therefore:

$H_0 = VR(q) = 1$

Lo and MacKinlay derived asymptotic standard normal test statistics for their variance ratios. I will use two different test statistics: z(q) in case of homoscedasticity and z*(q) in case of heteroscedasticity. The first statistic z(q) assumes an i.i.d. error term. The standard normal z (q) test statistic can be computed as:

$$z(q)=\frac{VR(q)-1}{\sqrt{\varphi(q)}}\approx N(0,1)$$

Where

$$\varphi(q)=\frac{2(2q-1)(q-1)}{3q(nq)}$$

The heteroscedastic test statistic z*(q) allows us to relax the requirements of i.i.d. increments. Despite the presence of heteroscedasticity, the test statistic z*(q) is still asymptotically standard normal in case of a random walk. It can be defined as:

$$z^*(q)=\frac{VR(q)-1}{\sqrt{\varphi^*(q)}}\approx N(0,1)$$

where

$$\varphi(q)=\sum_{j=1}^{q-1}\left[\frac{2(q-j)}{q}\right]^2\hat{\delta}(j)$$

and

$$\hat{\delta}(j) = \frac{\sum_{k=j+1}^{nq} (P_k - P_{k-1} - \hat{\mu})^2}{\left[\sum_{k=1}^{nq} (P_k - P_{k-1} - \hat{\mu})^2 \right]^2}$$

$\hat{\mu} =$ Average return

I use both homoscedastic and heteroscedastic test statistics for aggregation values q of 2, 4, 8 and 16.

3.10Error Metrics Applied

Error metrics are the mathematical equations which are used to describe error between actual and predicted values. The difference between actual and predicted values shows how well the model has performed. The main idea of forecasting techniques is to minimize this value since this should influence the performance and reliability of the model. These error metrics have significant importance in forecasting of future sales, as the measurements taken by these metrics, highly influence the future planning of the organizations. According to Bryan,(2005)any metric which measures the error should have five basic qualities which are: validity, easy to interpret, reliability, presentable and have statistical equation (i.e. can be represented in the formof mathematical equation).

Here *validity* refers to the degree to which an error metric measures what it claims to measure. In other words validity refers to whether error metric really measures what it intend to measure. The metric should measure the results which can relate to the available data. For example, if the metric is indented to measure result in binary form, then only binary result will be consider as valid output, otherwise validity of the metric can be questioned. Validity also refers to the authenticity of the measurement.

Easy to Interpret refers to the simplicity of the metric. The metric should be easy to understand and avoid complexity. The practitioners normally avoid using complex metrics since especially financial forecasting is complex and there is a need to be able to explain and motivate decisions, which is easier if the metric is simple and easy to understand.

Reliability is the consistency of the error metric to measure error using the same measurement method on same subject. If repeated measurements are taken and every time the results are highly consistent or even identical then there is a high degree of reliability, but if the measurements have large variations then reliability is low. The error metric is reliable when it is evident that it will produce the same result every time it is given the same data.

Presentable refers to the ability of error metric and its measurement to be represented in a form which is easy to understand.

S*tatistical equation* suggests that the error metric can be represented in the form of mathematical equation. Mathematical equations are most common and easy to interpret form of error metrics.

Rummel (1976) has considered linear correlation as the workhorse of quantitative research and analysis. This motivates to consider linear correlation in comparing models.

On the basis of reliability, validity and wide use, the following performance (error) measuring metrics are recommended for evaluating models. Mentzer J.T. and Moon M.A., (2005); Barreto H. and Howland (2006) all elaborate the significance of these metrics.

a. MAPE
b. WMAPE
c. RMSE

However, in this study three different error metrics are considered for the evaluation of forecasting models. They are root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE) and linear correlation.

a. ***Root Mean Square Error (RMSE)***

Root mean Square Error (RMSE) is square root of average of sum-squared errors and is given by following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where

$\hat{y}_i$ = estimated value of y_i,

y_i = actual value

n = number of observations

There is one problem with RMSE and it is that they may be close to 0 if large positive and negative errors cancel out each other. RMSE gives high weight to the large errors and are generally useful where large errors are not of importance.

RMSE are more sensitive than other metrics to the infrequent large errors as the squaring process gives large weight to very large errors (Decision 411, 2010).

b. ***Mean Absolute Error (MAE)***

The problem of RMSE, canceling out of large positive and negative errors can be avoided by using Mean Absolute Errors. In average, MAE weights all the differences equally.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$$

where

$\hat{y}_i$ = estimated value of y_i,

y_i = actual value

n = number of observations

MAE and RMSE can be used together to study the variation in the errors in a set of forecasts. RMSE are always larger or equal than MAE.

c. ***Mean Square Error (MSE)***

The sum of the squared forecast errors for each of the observations divided by the number of observations. It is an alternative to the mean absolute deviation(MAE), except that more weight is placed on larger errors. While MSE is popular among statisticians, it is unreliable and difficult to interpret. Armstrong and Fildes (1995) found no empirical support for the use of the MSE or RMSE in forecasting. Fortunately, better measures are available as discussed in Armstrong (2001d).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where

$\hat{y}_i$ = estimated value of y_i ,

y_i = actual value

n = number of observations

3.11Linear Correlation (r)

The Linear correlation shows the strength of relationship between dependent and explanatory variable (MathsBit.com, 2010).

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

where,

*x*and y are two random variables.

Value of *r* lies between -1 and +1.

The value of *r* can be interpreted as follows:

a. Value of *r* nearly zero or zero implies that there is little or no relationship between *y* and *x*

b. Value of *r* closer to -1 or 1 implies that there is strong linear relation between *y* and *x*

c. Positive value of *r* means that *y* value increases as *x* value increases

d. Negative value of *r* means that *y* value decreases as *x* value increa**ses**.

One point must be remembered that the causal relationship does not imply causality. It is possible that increase/decrease in value of *y* may be caused by some other factors except *x*. Hence, it will be wise to conclude that when high sample correlation is seen means there may exist linear trend between *x* and *y* (McClave, Benson, Sincich, 1998).

Table 3.1: Absolute Value of Correlation Coefficient Interpretation

0.90 – 1.00	Very high correlation
0.70 – 0.89	High correlation
0.40 – 0.69	Moderate correlation
0.20 – 0.39	Low correlation
0.00 – 0.19	Very low correlation

Source:Guide to interpretation of the correlation coefficient. (Tersine, 1988)

3.12 Akaike Information Criterion (AIC)

One of the most commonly used information criteria is AIC. The idea of AIC (Akaike, 1973) is to select the model that minimises the negative likelihood penalised by the number of parameters as specified in the equation (9).

$$AIC = -2 \log p(L) + 2p \qquad (9)$$

Where L refers to the likelihood under the fitted model and p is the number of parameters in the model.

Specifically, AIC is aimed at finding the best approximating model to the unknown true data generating process and its applications draws from (Akaike, 1973; Bozdogan, 1987; Zucchini, 2000).

CHAPTER 4

ANALYSIS AND RESULTS

4.0 Introduction

In this chapter the results of the performance concerning both forecasting methods and error metrics are analysed. First the relationships between the considered errors matrices(i.e. MAE, RMSE and MSE) are examined considering the linear correlation between the competing stocks and the forecasting models under consideration. This is followed by the results and analysis of the Box Jenkin's Method with/without an exogenous variable. The chapter ends with a model forecasting accuracy comparison where the two methods are compared given their error metrics scores.

Fundamental measures that will be very instrumental in the forecastability of the candidate methods are the Akaike Information Criterion (AIC) and the linear correlation between the considered stocks given the respective history of data. These stocks are: BUNGE LIMITED (BG GRP), BP, CAINE ENERGY (CNE) and TULLOW OIL (TLW). The linear correlation between the considered stocks given the six year history of monthly adjusted close price data is given below in Table 4.1.

Table 4.1: Correlation Matrix –First Time Horizon (2005 - 2010)

	BG	**BP**	**CNE**	**TLW**
BG	1.0000	-0.2490	0.5511	0.9114
BP	-0.2490	1.0000	0.0043	-0.2619
CNE	0.5511	0.0043	1.0000	0.6301
TLW	0.9114	0.2619	0.6301	1.0000

Subsequent to table 4.1 is the linear correlation between the considered stocks given the four year history of monthly adjusted close price data is given below in Table 4.2.

Table 4.2: Correlation Matrix –Second Time Horizon (2007 - 2010)

	BG	BP	CNE	TLW
BG	1.0000	-0.0678	0.6543	0.8106
BP	-0.0678	1.0000	-0.0789	-0.0601
CNE	0.6543	-0.0789	1.0000	0.755
TLW	0.8106	-0.0601	0.755	1.0000

None of the random walk test of all the considered stocks in the Oil and Gas Industry was significant both with homoskedastic and heteroskedastic errors as shown in appendix C.

4.1.1 Results for ARIMA Models Using BG as A UnivariateVariable (2005 - 2010)

Bunge Limited is a stock in the Oil and Gas Industry in the London Stock Exchange. With reference to the above correlation matrix in the first time regime, it has the lowest correlation with BP of (-0.2490) and the highest with TLW which is (+ 0.9114). Likewise, data in the second time regime also have the lowest and highest correlation with BP and TLW as (-0.0678) and (+0.8106) respectively.

Stocks with the lowest and highest correlation with BG served as the exogenous variables to the arima model using BG as the univariate variable in the case of the two time regimes shown in Tables 8. Unlike the second time regime, in first time regime,arima with/without the exogenous variable had the same model but significantly with different Akaike's Information Criterion (AICs'). Even though the AICs' for the respective models in the two time regimes are not far from each other, it is evident that the univariate arima model with an exogenous variable had smaller AIC's, exogenous variables with the highest correlation with the univariate variables (BG) had the lowest AIC followed by the exogenous variable with the lowest correlation.The AIC of the models considered in both regimes are arranged in ascending order as (DBG/DTLW, DBG/DBP, DBG) and (DBG*/DTLW*, DBG*/DBP*, DBG*) respectively. In both regimes, arima had the largest AIC.

Table 4.3: Selected Models

Ticker	Model Type	Selected Model	AIC
DBG	ARIMA	(0,1,2)	781.65
DBG/DBP	ARIMAX	(0,1,2)	765.72
DBG/DTLW	ARIMAX	(0,1,2)	757.26
DBG*	ARIMA	(0,1,2)	533.38
DBG*/DBP*	ARIMAX	(0,1,2)	525.53
DBG*/DTLW*	ARIMAX	(2,1,0)	517.59

The considered risk metric (i.e. MAE, RMSE and MSE) had smaller values for highly correlated exogenous variables with DBG. The linear correlation amongst the variables seems to have a significant impact on both the AIC and risk metrics.Likewise, the AIC is having some level of impact on the error metrics. This is evident in table 9.

Table 4.4: Error Metrics

	Test Type		
Ticker	**MAE**	**RMSE**	**MSE**
BG	44.7725	56.5525	3198.1830
BG/BP	36.80373	49.8227	2482.3040
BG/TLW	36.4139	46.9652	2205.7260
BG*	53.4383	65.2898	4262.7570
BG*/BP*	45.7858	58.7686	3453.753
BG*/TLW*	42.6022	54.0129	2917.3950

The Diebold and Mariano test for comparing models using the in – sample errors supports our earlier observations. The three models under the respective time regimes rate an exogenous variable with the univariate variable DBG very high as compared to the univariate. Implying that, in terms of comparing all the models, Tullow Oil (TLW), an exogenous variable which has the highest linear correlation with the univariate variable, Bunge (BG), and BG/BP in both time regimes emerges as the best forecasting model.The Diebold and Mariano test results can be seen in Table 10.A*p* – value less than or equal to 0.05, pegs the two competing models as having the same forecasting ability. On the otherhand, p-value more than 0.05 supports the second competing forecasting model as against the first model. Hence BG/BP and BG/TLW areequally good forecasting model as compared to the univariate model, BP.

In the second regime, BG*/BP* and BG*/TLW*equally were also the best as compared to the univariate model, BG*.

Table 4.5: Diebold Test for Comparing Models

Ticker	Test Type	p – value
BG VRS BG/BP	DM	< 0.0001
BG VRS BG/TLW	DM	0.0018
BG/BP VRS BG/TLW	DM	0.4370
BG* VRS BG*/BP*	DM	< 0.0001
BG* VRS BG*/TLW*	DM	0.0049
BG*/BP* VRS BP*/TLW*	DM	0.1824

The parameter estimates in both time regimes is presented in Table 4.6

Estimates	**BG**	**BG/BP**	**BG/TLW**	**BG***	**BG*/BP**	**BG*/TLW**
β	-	0.6785	0.4282	-	0.6366	0.4206
AR (1)	-	-	-	-	-	-0.0439
AR (2)	-	-	-	-	-	0.3265
AR (3)	-	-	-	-	-	-
MA (1)	-0.0199	-0.0297	0.0179	-0.022	-0.033	-
MA (2)	0.2916	0.3322	0.2779	0.3103	0.3367	-
MA (3)	-	-	-	-	-	-

Table 4.6: Parameter Estimates (BG)

Figure 4.1 below shows the in sample time plots of the selected models in the first time regime (2005 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (BG).

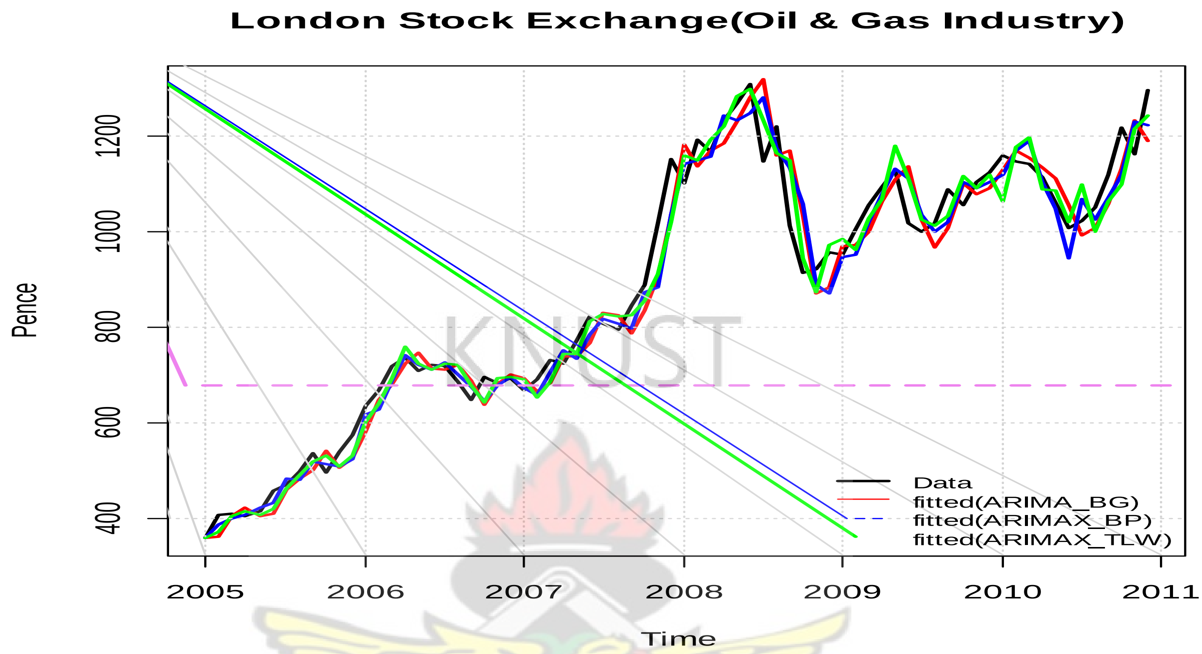

Figure 4.1: Time Plot for Performing Models (BP) - (2005 - 2010)

On the hand Figure 4.2 below shows the in sample time plots of the selected models in the second time regime (2007 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (BG). However, in the first regime, the candidate models followed the observed data (BG) very as compared to that of the second regime (2007 – 2010).). Error metrics for the second time regime is large as compared to the first time regime as in table 4.4 above. The graph in the second time regime is not as compact as the one in the first time regime.

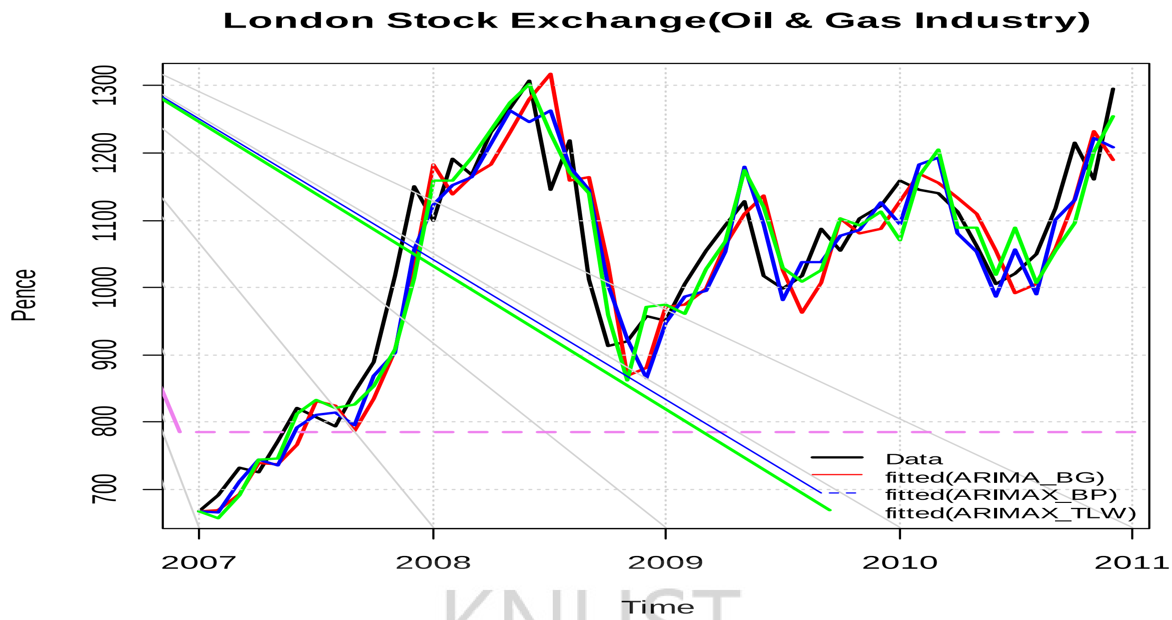

Figure 4.2: Time Plot for Performing models (BP) - (2007 - 2010)

The performance of the respective candidate models are ranked in Table 4.7 below. The Diebold and Mariano test guided the selection and ranking of the models.

Table 4.7: Results Summary (BG)

Rank	**1**	**1**	**3**	**1**	**1**	**3**
Model	**DBG/DTLW**	**DBG/DBP**	**DBG**	**DBG*/DTLW***	**DBG*/DBP***	**DBG***
L. COR	0.911	- 0.249	N/A	+ 0.811	- 0.068	N/A
AIC	757.26	765.72	781.65	517.57	525.53	533.38
MAE	36.41	36.80	44.77	42.60	45.79	53.44
RMSE	46.97	49.82	56.55	54.01	58.77	65.29
MSE	2205.73	2482.30	3198.18	2917.40	3453.75	4262.75

4.1.2Results for ARIMA Models Using BP as A Univariate Variable (2005 - 2010)

BP is another company listed in the Oil and Gas industry on the London Stock Exchange. With respect to the other considered stocks in the same industry, BP has the highest linear correlation of + 0.0043 with CNE and a lower correlation of - 0.2619 with TLW in the first time regime. However, with respect to the second time regime, there is a significant change in the linear correlation. Here, BP has the highest correlation of - 0.0601 with

CNE and lower correlation of- 0.0789 with TLW amongst the considered stocks. The following models,were selected by the Box Jenkins methodology in the two regimes, is presented below in Tables 13.The AIC of the models considered in both regimes are arranged in ascending order as (DBP/TLW, DBP/CNE, DBP) and (DBP*/TLW*, DBP*/CNE*, DBP*). Here there is a serious distortion, the stock with the highest correlation with the univariate variable CNE is rather having a higher AIC as compared to the exogenous variable with a lower correlation TLW. This is in sharp contrast with our observation in the previous models with BG.

Table 4.8: Selected Models

Ticker	Model Type	Selected Model	AIC
DBP	ARIMA	(3,1,0)	721.03
DBP/TLW	ARIMAX	(1,1,0)	700.84
DBP/CNE	ARIMAX	(1,1,0)	715.26
DBP*	ARIMA	(0,1,2)	494.73
DBP*/DCNE	ARIMAX	(1,1,3)	488.09
DBP*/DTLW	ARIMAX	(0,1,3)	478.71

The in-sample forecasting accuracy measurement by the considered error metrics is shown in Table 14. The results as in the first time regime show that BP/TLW is a good candidate for forecasting because of its smaller error metrics measurements as compared to the other competing candidate models namely BP/CNE and BP. The best forecasting model according to Table 14 are the exogenous variable with the lowest correlation (i.e. BP/TLW) and (BP/CNE) respectively.

On the other hand, BP*/TLW*in the second time regimes also gives the best results with respect to the error metrics measurements. This seems to go in line with the subsequent results so far except that the linear correlation is inconsistent.

Table 4.9:Error Metrics (BP)

	Test Type		
Ticker	**MAE**	**RMSE**	**MSE**
BP	27.4674	36.3502	1321.3400
BP/TLW	22.4925	32.0509	1027.2600
BP/CNE	24.08444	35.47503	1258.4780
BP*	30.9007	43.3619	1880.2520
BP*/CNE*	26.0165	37.5155	1407.4110
BP*/TLW*	24.5750	34.9835	1223.8430

The Diebold and Mariano test for the model comparison in both regimes as presented below in Table15 supports the error metrics results discussed previously. In both cases, a model of the univariate variable (BP) with an exogenous variable (CNE or TLW) by the Diebold and Mariano testput them as having the same forecasting ability. Diebold test again supports models of BP with either CNE or TLW as exogenous variable than the singular BP model. This is evident in the results below in table 15.In the first regime, BP/CNE and BP/TLW were the best forecasting models followed BP. Subsequently, BP*/TLW*and BP*/CNE* in the second time regime are equally the best as against the univariate BP* model.

Table 4.10: Diebold Test for Comparing Models (BP)

Ticker	Test Type	p – value
BP VRS BP/TLW	DM	0.0154
BP VRS BP/CNE	DM	0.0314
BP/TLW VRS BP/CNE	DM	0.7667
BP* VRS BP*/CNE*	DM	<0.0001
BP* VRS BP*/TLW*	DM	0.0023
BP*/CNE* VRS BP*/TLW*	DM	0.2576

Table 4.10gives the model estimates for the performing forecasting models.

Table 4:11: Parameter Estimates (BP)

Estimates	BP	BP/TLW	BP/CNE	BP*	BP*/CNE	BP*/TLW
β	-	0.3043	0.4903	-	0.3591	0.2378
AR (1)	-0.101	0.0026	-0.0609	-	0.6029	-
AR (2)	0.0766	-	-	-	-	-
AR (3)	-0.291	-	-	-	-	-
MA (1)	-	-	-	-0.078	-0.6766	-0.0221
MA (2)	-	-	-	0.0759	0.1348	-0.0035
MA (3)	-	-	-	-	-0.3715	-0.245

Figure 4.3 below shows the in sample time plots of the selected models in the first time regime (2005 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (BP).

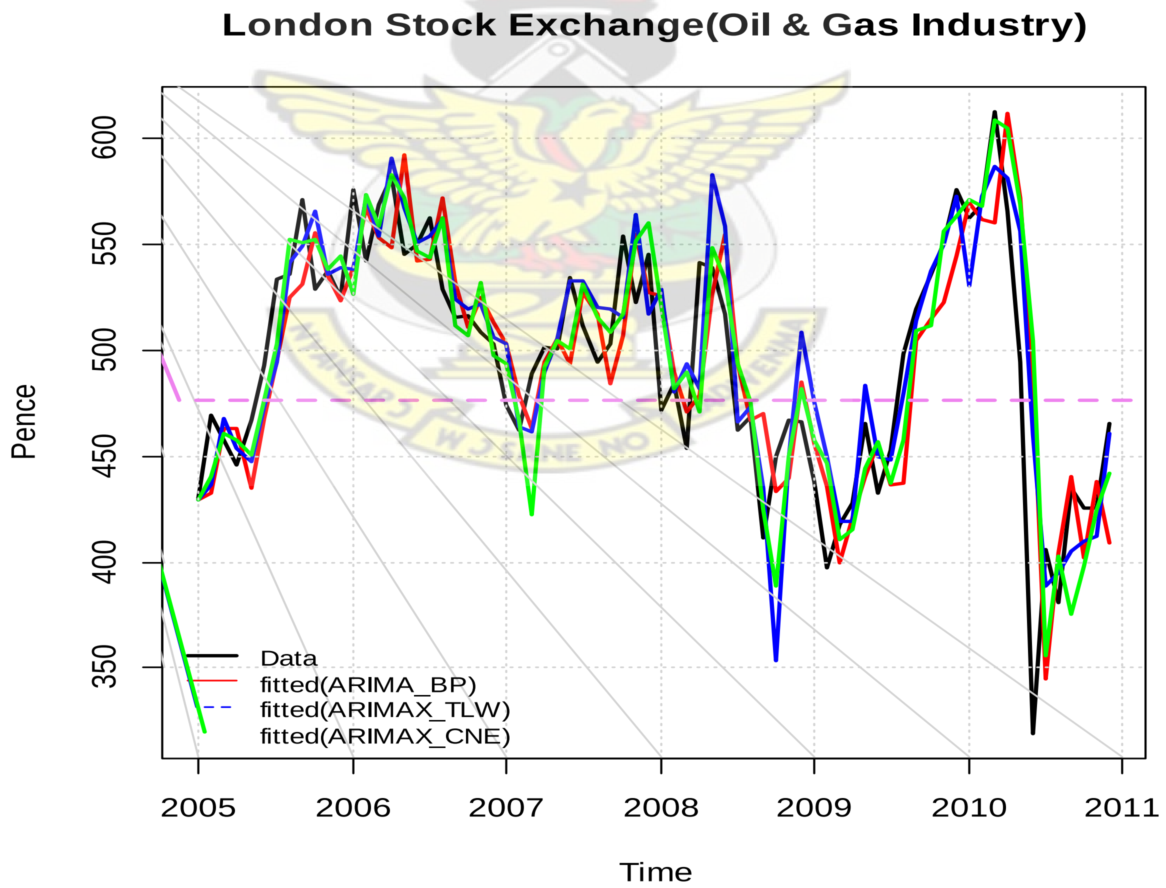

Figure 4.3: Time Plot for Performing Models (BP) - (2005 - 2010)

On the hand Figure 4.4 below shows the in sample time plots of the selected models in the second time regime (2007 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (BP). However, in the first regime, the candidate models followed the observed data (BP) very as compared to that of the second regime (2007 – 2010).

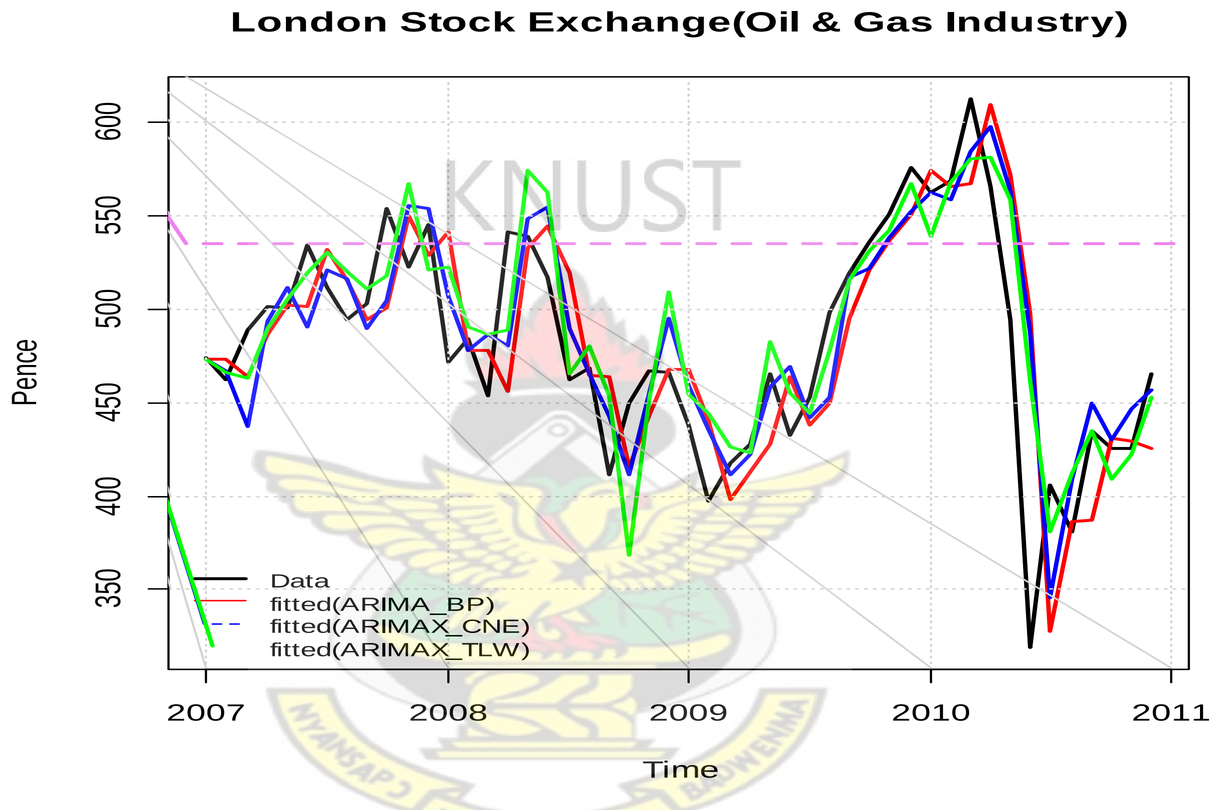

Figure 4.4: Time Plot for Performing Models (BP) - (2007 - 2010)

The performance of the respective candidate models are ranked in Table 4.11 below. The Diebold and Mariano test guided the selection and ranking of the models. Again it is evident from table 4.11 that the AICs affect the error metrics. Comparatively, the smaller the AIC the smaller the error metrics. Linear correlation has conflicting effect on the error metrics.

Table 4.12 Summary of Results (BP)

Rank	1	1	3	1	1	3
Model	**DBP/DCNE**	**DBP/DTLW**	**DBP**	**DBP*/DTLW***	**DBP*/DCNE***	**DBP***
L. COR	- 0.0043	- 0.2619	N/A	- 0.0601	- 0.0789	N/A
AIC	715.26	700.84	721.03	478.71	488.09	494.73
MAE	24.08	22.49	27.47	24.57	26.02	30.90
RMSE	35.48	32.05	36.35	34.98	37.52	43.36
MSE	1258.48	1027.26	1321.34	1223.84	1407.41	1880.25

4.1.3 Results for ARIMA Models Using CNE as A Univariate Variable (2005 - 2010)

Caine Energy (CNE) is the last but not the least considered company listed on the Oil and Gas industry in the London Stock Exchange. Caine Energy is highly correlated with BP and having the lowest correlation with TLWin both time regime among the considered stocks. The highest correlation between CNE and the considered stock in the first regime is (+ 0.6301 - TLW) and (+ 0.0043 – BP). On the other regime, CNE is highly correlated with (+0.755 – TLW*) and the lowest correlated stock among the considered stocks is (-0.0789 – BP*).Table 18 shows the best models in both regimes by the Box Jenkins' selection criteria. The AIC of the models considered in both regimes are arranged in ascending order as (DCNE/TLW, DCNE/BP, DCNE) and (DCNE*/TLW*, DCNE*/BP*, DCNE*).

Table 4.13: Selected Models (CNE)

Ticker	Model Type	Selected Model	AIC
DCNE	ARIMA	(0,1,1)	692.36
DCNE/BP	ARIMAX	(0,1,1)	680.95
DCNE/TLW	ARIMAX	(1,1,0)	662.52
DCNE*	ARIMA	(1,1,0)	469.66
DCNE*/BP*	ARIMAX	(0,1,3)	465.77
DCNE*/TLW*	ARIMAX	(0,1,3)	447.22

In – sample error measures to differentiate one model from the other in both regimes are shown in table 19. The best forecasting model with respect to the error metric is CNE/TLW and CNE*/TLW* in the first and the second time regime. These two models in the two time regimes do have the highest correlation among the considered stocks with the univariate variable. Subsequent to these models are CNE/BP and CNE*/BP* with the lowest linear correlation in both regimes. The ARIMA model with CNE is the worse forecasting model in terms of the error metric measures. The AIC is consistently affecting the error metrics measurements. A lower AIC for a candidate model also performs creditably well with the error metric.

Table 4.14: Risk Metrics (CNE)

	Test Type		
Ticker	**MAE**	**RMSE**	**MSE**
CNE	22.7411	30.6196	937.5589
CNE/BP	20.7456	27.8590	776.1221
CNE/TLW	17.3169	24.4696	598.7617
CNE*	24.5394	33.9308	1151.2960
CNE*/BP*	21.7394	30.4750	928.7241
CNE*/TLW*	17.44367	24.84734	617.3904

The Diebold and Mariano test gives p – values with respect to candidate models in both regimes. The results as in table 20 also support our conclusion in the previous models. In the first regime, the p – value of 0.0685 failed to reject the null hypothesis that the two forecasting models (i.e.CNE versus CNE/BP) have the same in-sample accuracy levels. Again, for modelsCNE versus CNE/TLW, p – value of (< 0.0001) renders the two models unequal with respect to its in-sample forecasting abilities. Here CNE/TLW according to the Diebold and Mariano test is the best model as compared to CNE. The last comparing candidate model in the first regime (i.e. CNE/BP VRS CNE/TLW) reject the null hypothesis of the Diebold and Mariano test. A p – value of 0.0045 rejects the null hypothesis and hencemaking CNE/TLWmodel more efficient than CNE/BP.

The models in the second time regimes are again in a similar fashion as the models in the first time regime. This is as a result of the AIC patterns within the candidate models in the first time regime which repeats itself in a similar way. The AIC consistently is dictating the pace with respect to the error metrics measurements.

Table 4.15: Diebold Test for Comparing Models (CNE)

Ticker	Test Type	p - value
CNE VRS CNE/BP	DM	0.0685
CNE VRS CNE/TLW	DM	< 0.0001
CNE/BP VRS CNE/TLW	DM	0.0045
CNE* VRS CNE*/BP*	DM	0.0708
CNE* VRS CNE*/TLW*	DM	0.0003
CNE*/BP* VRS CNE*/TLW*	DM	0.0131

Estimates of the selected candidate models used for the model comparison are shown in table 4.15.

Table 4.16: Parameter Estimates (CNE)

Estimates	CNE	CNE/BP	CNE/TLW	CNE*	CNE*/BP	CNE*/TLW
β	-	0.3303	0.2573	-	0.3711	0.281
AR (1)	-	-	-0.0764	-0.021	-	-
AR (2)	-	-	-	-	-	-
AR (3)	-	-	-	-	-	-
MA (1)	0.0025	0.1153	-	-	0.1656	-0.2696
MA (2)	-	-	-	-	0.0479	0.2129
MA (3)	-	-	-	-	0.2432	0.379

Figure 4.5 below shows the in sample time plots of the selected models in the first time regime (2005 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (CNE).

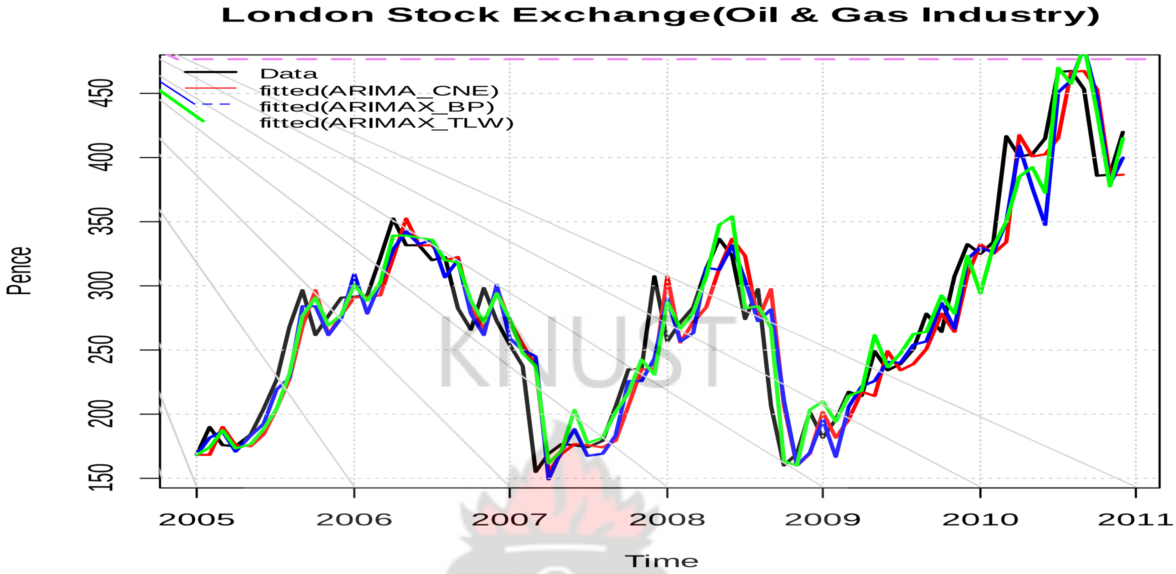

Figure 4.5: Time Plot for Performing Models (CNE) - (2005 - 2010)

On the hand Figure 4.6 below shows the in sample time plots of the selected models in the second time regime (2007 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (CNE). However, in the first regime, the candidate models followed the observed data (CNE) very as compared to that of the second regime (2007 – 2010).). Error metrics for this second time regime is large as compared to the first time regime as in table 4.13 above as the graph in the second time regime is not as compact as the one in the first time regime.

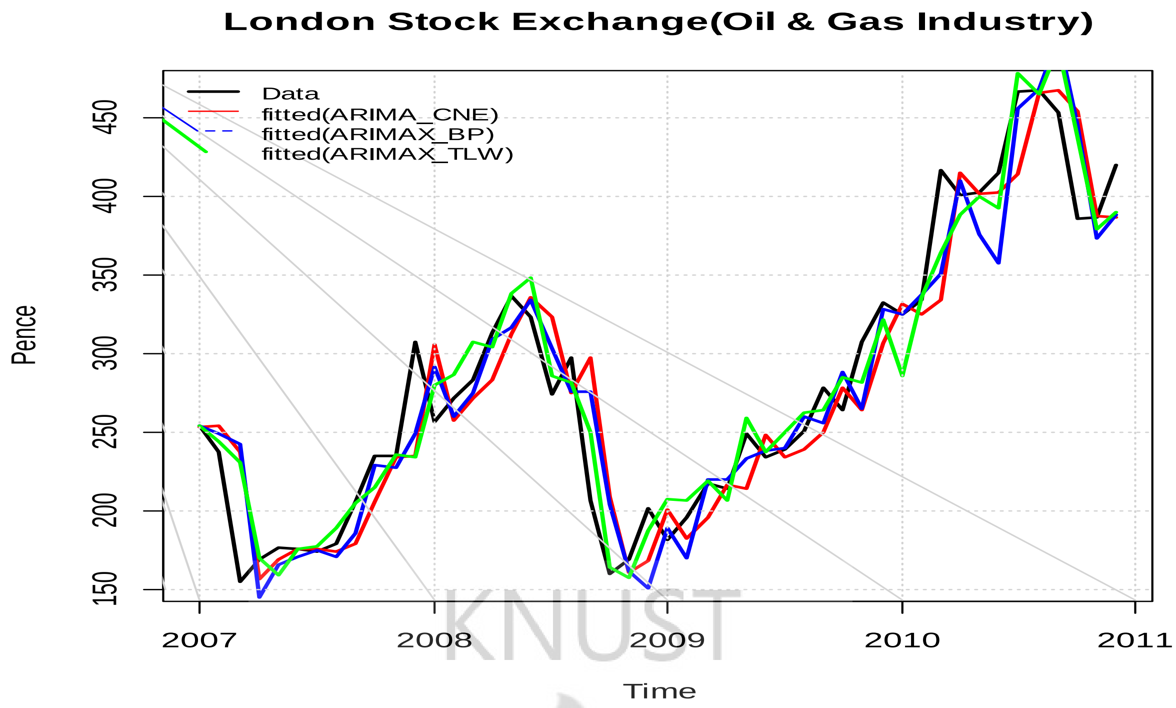

Figure 4.6: Time Plot for Performing Models (CNE) - (2007 - 2010)

The performance of the respective candidate models are ranked in Table 4.16 below. The Diebold and Mariano test guided the selection and ranking of the models.

Table 4.17: Summary of Results

Rank	**1**	**2**	**3**	**1**	**2**	**2**
Model	**DCNE/DTLW**	**DCNE/DBP**	**DCNE**	**DCNE*/DTLW***	**DCNE*/DBP***	**DCNE***
L. COR	+ 0.630	+ 0.004	N/A	0.755	- 0.079	N/A
AIC	662.52	680.95	692.36	447.22	465.77	469.66
MAE	17.32	20.75	22.74	17.44	21.74	24.54
RMSE	24.47	27.86	30.62	24.85	30.48	33.93
MSE	598.76	776.12	937.56	617.39	928.72	1151.30

4.1.4 Results for ARIMA Models Using TLW as A Univariate Variable (2005 - 2010)

Tullow Oil (TLW) is the last considered company listed under the Oil and Gas industry in the London Stock Exchange. Tullow Oil is highly correlated with BG and having the lowest correlation with BPin both time regime among the considered stocks. The highest correlation between TLW and the considered stock in the first regime is (+0.9114- BG) and (- 0.2619– BP). On the other regime, TLW is highly correlated with (+0.8106 - BG) and the lowest correlated stock among the considered stocks is (-0.0601- BP). Table 23 shows the best models in both regimes by the Box Jenkins' selection criteria. The AIC of the models considered in both regimes are arranged in ascending order as (DTLW/BP, DTLW/BG, DTLW) and (DTLW*/BP*, DTLW*/BG*, DTLW*).

Table 4.18: Selected Models (TLW)

Ticker	Model Type	Selected Model	AIC
DTLW	ARIMA	(0,1,1)	813.91
DTLW/BP	ARIMAX	(1,1,0)	787.62
DTLW/BG	ARIMAX	(1,1,0)	789.57
DTLW*	ARIMA	(0,1,1)	557.36
DTLW*/BP*	ARIMAX	(0,1,1)	540.36
DTLW*/BG*	ARIMAX	(1,1,0)	542.44

In – sample error measurement to differentiate one model from the other in both regimes are shown in table 24. In both regimes, is significantly evident that the error metrics are positioned according to their AIC scores. It is evident that DTLW/BP and DTLW*/BP* in both regimes were the models with minimal AIC. This is reflected in the error metrics as shown in table 24 except for TLW/BP in the first regime which failed this order.

Table 4.19: Error Metrics (TLW)

	Test Type		
Ticker	**MAE**	**RMSE**	**MSE**
TLW	51.9178	72.0734	5194.5800
TLW/BP	42.1243	59.0466	3486.4990
TLW/BG	41.7744	59.8640	3583.6970
TLW*	67.8227	86.2577	7440.3970
TLW*/BP*	54.4490	70.4521	4963.5030
TLW*/BG*	55.2465	72.04212	5190.0671

The best forecasting model by the error metric is TLW/BPand TLW*/BP* in the first and the second time regime respectively. These two models in the various regimes do have the lowest correlation among the considered stocks with the univariate variable (TLW). Subsequent to these models are TLW/BG and TLW*/BG* with the highest linearcorrelationin both regimes. The arima model,TLW is the worse forecasting model in terms of the error metric measures.ARIMAX models (TLW/BP, TLW/BG) and (TLW*/BP*, TLW*/BG) in both time regime were equally the best models whereas TLW and TLW* were the worst performing models.

Table 4.20: Diebold Test for Comparing Models (TLW)

Ticker	Test Type	p - value
TLW VRS TLW/BP	DM	0.0004
TLW VRS TLW/BG	DM	0.0032
TLW/BP VRS TLW/BG	DM	0.4617
TLW* VRS TLW*/BP*	DM	0.0006
TLW* VRS TLW*/BG*	DM	0.0103
TLW*/BP* VRS TLW*/BG*	DM	0.5576

Estimates of the selected candidate models used for the model comparison are shown in Table 4.20.

Table 4.21: Parameter Estimates (TLW)

Estimates	TLW	TLW/BP	TLW/BG	TLW*	TLW*/BP	TLW*/BG
β	-	1.0752	0.6824	-	1.1522	0.6966
AR (1)	-	0.1269	-0.113	-	-	-0.1068
AR (2)	-	-	-	-	-	-
AR (3)	-	-	-	-	-	-
	-					
MA (1)	0.031	-	-	-0.027	0.1593	-
MA (2)	-	-	-	-	-	-
MA (3)	-	-	-	-	-	-

Figure 4.7 below shows the in sample time plots of the selected models in the first time regime (2005 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (TLW).

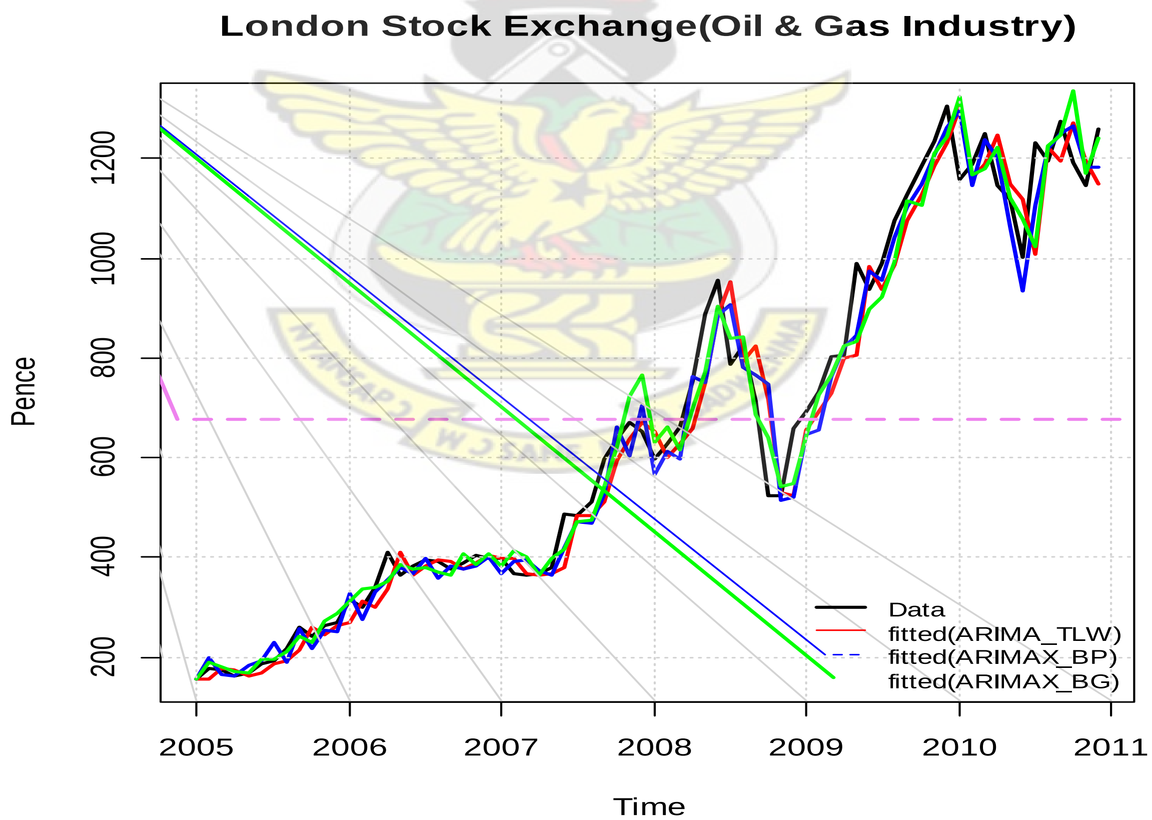

Figure 4.1: Time Plot of Performing Models (TLW) - (2005 - 2010)

On the hand Figure 4.8 below also shows the in sample time plots of the selected models in the second time regime (2007 – 2010). It is evident from the graph that all the candidate models performed well as each model followed the time plot of the observed data (TLW). However, in the first regime, the candidate models followed the observed data (TLW) very as compared to that of the second regime (2007 – 2010). Error metrics for the second time regime is large as compared to the first time regime as in table 4.18 above. The graph in the second time regime is not as compact as the one in the first time regime.

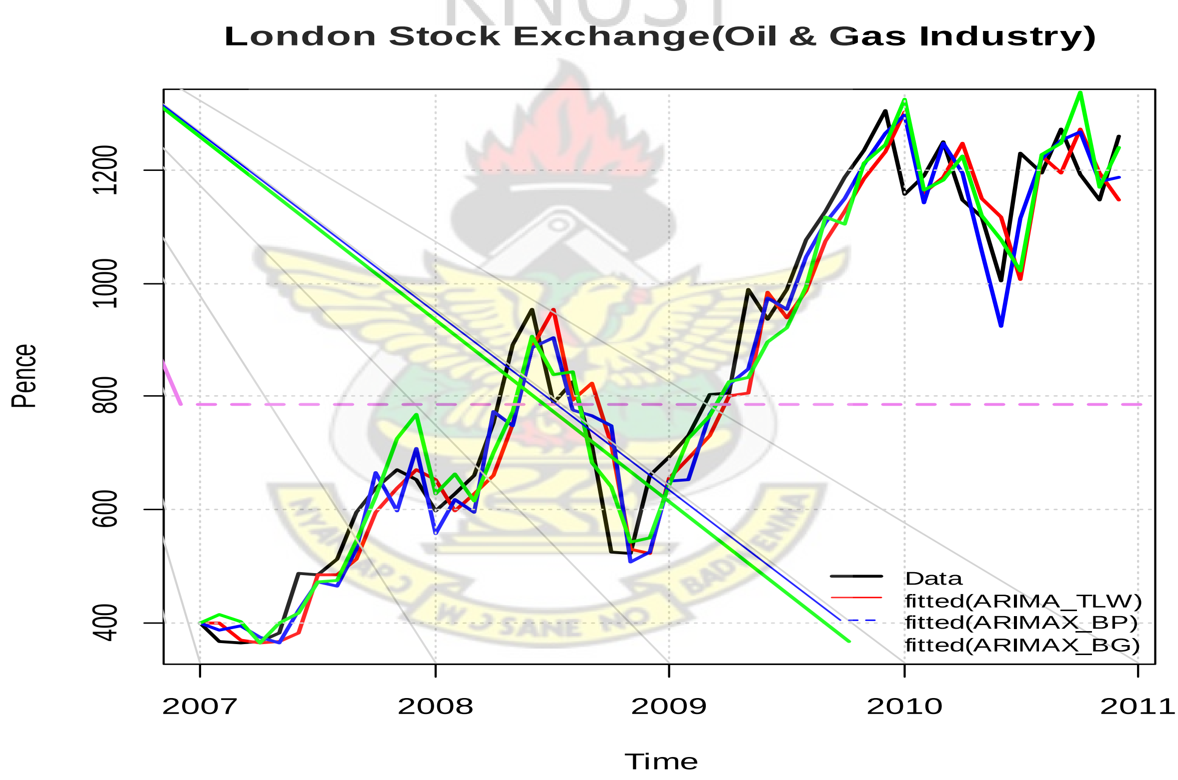

Figure 4.8: Time Plot of Performing Models (TLW) - (2007 - 2010)

The performance of the respective candidate models are ranked in Table 4.16 below. The Dieboldand Mariano test guided the selection and ranking of the models. The performance of the respective candidate models are ranked in Table 4.11 below. The Diebold and Mariano test guided the selection and ranking of the models. Again it is evident from table 4.21 that the AICs affect the error metrics. Comparatively, the smaller the AIC the smaller the error metric of the candidate model. Linear correlation has conflicting effect on the error metrics.

Table 4.22: Summary of Results (TLW)

Rank	**1**	**1**	**3**	**1**	**1**	**3**
Model	**DTLW/DBG**	**DTLW/DBP**	**DTLW**	**DTLW*/DBG***	**DTLW*/DBP***	**DTLW***
L. COR	+ 0.911	+ 0.262	N/A	+ 0.811	- 0.060	N/A
AIC	789.57	787.62	813.91	542.44	540.36	557.36
MAE	41.7744	42.1243	51.9178	55.2465	54.449	67.823
RMSE	59.864	59.0466	72.0734	72.04212	70.4521	86.258
MSE	3583.697	3486.499	5194.58	5190.0671	4963.503	7440.4

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

The summary of the findings as well as the conclusion are presented in this chapter. Recommendations suggested by the researcher have also been included in this section that provides a frame work of how stakeholders in the financial industry can improve upon in-sample stock forecasting accuracy.

5.1Discussion

From the background to the problem, the objectives of the research study and the data used and analysed, the researcher has established the status and how to improve the in-sample forecasting accuracy of stocks using the ARIMA models with/without an exogenous variable

With reference to the first objective of this thesis, it is empirically evident that ARIMA model with an exogenous variable (ARIMAX) performed creditably well in all cases and scenarios as outlined in chapter four. This emphasises that, when improving the in – sample forecasting accuracy of a stock price using the Box – Jenkins model, it is in order to incorporate an exogenous variable to further augment the accuracy of the in – sample forecast. In this thesis, historical adjusted close stock prices of four considered stocks in the Oil and Gas Industry in the London Stock Exchange were use as possible exogenous variable or as public information.

On the other hand, linear correlation between the ARIMA model with exogenous variable did very little to improve the in-sample forecasting accuracy of all the considered scenarios in this thesis. In most cases, the high and low linear correlation between stocks of candidate models only gave signal to the corresponding Akaike Information Criterion

(AIC) value. High correlation in most cases gave a lower value of the AIC and vice-versa. However this assertion was not consistent. Evidently, the Diebold and Mariano test of accuracy is dependent AIC of the candidate models.However, in most cases smaller AIC values turn to minimise the considered error metrics (i.e. MAE, RMSE and MSE) and vice versa. This is evident throughout the results. The linear correlation on the hand had little or no impact on the performing models.

The Box-Jenkins Method with/without an exogenous variable supports the semi – strong form of EMH. Thus, the information, Ω_t set comprising of the past and current asset prices (as well as possibly dividends and variables such as trading volume) and all publicly available information supports the Efficient Market Hypothesis (EMH) in its semi-strong form. Timmermann and Granger, (2004) in their paper "Efficient market hypothesis and forecasting" argued that traditional time seriesforecasting methods relying on individual forecastingmodels or stable combinations of these are notlikely to be useful. This in one way or the other confirms our findings that even though ARIMAX model is an improvement of an Arima model in most cases, however the two forecasting models are not useful since it cannot guarantee any investor to gain undue advantage in making economic gains.

5.2 Conclusion

Overall, it is clear from the findings that again, the Efficient Market Hypothesis is still very relevant since none of the considered model was able to violate the semi – strong form of EMH. Evidently, among the considered models, no single successful forecasting model was sufficient to demonstrate violation of the EMH. However, once model uncertainty is accounted for in all the considered models, this confirms the random nature of stock prices. This behaviour of the stock market is much needed otherwise there would exist a 'money-machine' producing unlimited wealth, which cannot occur in a stable economy.

5.3 Recommendations

In this thesis, model uncertainty was accounted for the three considered models namely the Box-Jenkins Method with/without an exogenous variable in all time regimes. The Box-Jenkins Method with an exogenous variable given the lowest/highest correlated stocks with each considered univariate variables (i.e. BG, BP, CNE and TLW). Hence stock market participants should not rely solely on these predictive models for any sort of leads in the stock market because the market follows a random walk according to the Variance RatioTest as shown in appendix C.

REFERENCES

Armstrong, J. S. (2001d), "Evaluating forecasting methods," in J. S. Armstrong (ed.), Principles of Forecasting. Norwell, MA: Kluwer Academic Press

Armstrong, J. S. &Fildes R. (1995), "On the selection of error measures for comparisons among forecasting methods," Journal of Forecasting, 14, 67-71

Bachelier, L. (1964). Theory of speculation. In Cootner, P. (Ed.), The Random Character of Stock Market Prices. Cambridge: Massachusetts Institute of Technology Press.

Barreto H., Howland F.M., 2006, Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel, Cambridge University Press.

Bianchi, L., Jarrett, J., &Hanumara, T. C. (1998).Improving forecasting for telemarketing centers by ARIMA modeling with interventions.International Journal of Forecasting, 14, 497– 504.

Birgul E., Melten O. and Bertan B., Stock Market Prediction Using Artificial Neural Networks.www.hicbusiness.org/biz2003/proceedings/Birgul%Egeli.pdf (2003).

Chaudhuri, S, Madigan, D and Fayyad, U eds. Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD, San Diego, California, August 11 – 18, 1999). New York: ACM Press, 1999, pp. 53 – 62

Cirna G., Ajith A., Sang Y.H. and Ramos V. (2010) Stock Market Prediction Using Multi Expression Programming. www.softcomputing.net/alea.pdf

Cootner, P. (Ed.), (1964). The Random Character of Stock Market Prices. Cambridge: Massachusetts Institute of Technology Press.

Cummins, J. D., &Griepentrog, G. L. (1985). Forecasting automobile insurance paid claims using econometric and ARIMA models. International Journal of Forecasting, 1,203– 215.

De Gooijer J.G. and R.J. Hyndman, 25 years of time series forecasting: International Journal of Forecasting, (2006) vol. 22 443– 473

Dhrymes, P. J., &Peristiani, S. C. (1988).A comparison of the forecasting performance of WEFA and ARIMA time series methods.International Journal of Forecasting, 4, 81– 101.

Di Caprio, U., Genesio, R., Pozzi, S., &Vicino, A. (1983).Shortterm load forecasting in electric power systems: A comparison of ARMA models and extended Wiener filtering. Journal of Forecasting, 2, 59– 76.

Downs, G. W., &Rocke, D. M. (1983).Municipal budget forecasting with multivariate ARMA models.Journal of Forecasting, 2, 377– 387.

duPreez, J., & Witt, S. F. (2003). Univariate versus multivariate time series forecasting: An application to international tourism demand. International Journal of Forecasting, 19,435– 451.

Edlund, P. -O., &Karlsson, S. (1993). Forecasting the Swedish unemployment rate.VAR vs. transfer function modeling. International Journal of Forecasting, 9, 61– 76.

Eriksson, L., Wiedersheim-Paul, F., (2001) Attutreda, forskaochrapportera, Liber ekonomi, Malmö.

Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. Journal of Finance, 25, 383– 417.

Fama, E. F. (1991). Efficient capital markets II. Journal of Finance, 46, 1575–1617.

Fawcett, T., and Provost, F, 1999. Activity Monitoring: Noticing Interested Changes in Behavior. Fung G.P.C., Yu, J.X., and Lam, W., 2003. Stock Prediction: Integrating Text Mining Approach Using Real – time News. In *Proceedings of the 7th IEEE International Conference on Computational Intelligence for Financial Engineering (CIFEr) (Hong Kong, China, March 20 – 23, 2003)*, IEEE Press, pp. 395 – 402.

Fung G.P.C., Yu, J.X., and Lu, H., 2005.The Predicting Power of Textual Information on Financial Markets.*IEEE Intelligent Informatics Bulletin*, 5(1), pp. 1 – 10.

Fung G.P.C., Yu, J. X., and Lam, W., 2002.News Sensitive Stock Trend Prediction.

GebhardKirchgässner and JürgenWolters (2007).Introduction to Modern Time Series Analysis. Springer Berlin Heidelberg, New York, pp.178

Geurts, M. D., & Kelly, J. P. (1986).Forecasting retail sales using alternative models.International Journal of Forecasting, 2,261–272.

Geurts, M. D., & Kelly, J. P. (1990). Comments on: In defense of ARIMA modeling by D.J. Pack. International Journal of Forecasting, 6, 497– 499.

Gidofalvi, G., 2001. Using News Articles to Predict Stock Price Movements. San Diego: Department of Computer Science and Engineering, University of California. [Online] Available from: http://www.cs.aau.dk/~gyg/docs/financial-prediction.pdf

Grambsch, P., &Stahel, W. A. (1990). Forecasting demand for special telephone services: A case study. International Journal of Forecasting, 6, 53–64.

Harris, J. L., & Liu, L. -M. (1993).Dynamic structural analysis and forecasting of residential electricity consumption.International Journal of Forecasting, 9, 437– 455.

Hein, S., &Spudeck, R. E. (1988).Forecasting the daily federal funds rate.International Journal of Forecasting, 4, 581– 591.

Heuts, R. M. J., &Bronckers, J. H. J. M. (1988).Forecasting the Dutch heavy truck market: A multivariate approach.International Journal of Forecasting, 4, 57– 59.

Hillmer, S. C., Larcker, D. F., & Schroeder, D. A. (1983).Forecasting accounting data: A multiple time-series analysis.Journal of Forecasting, 2, 389– 404.

Hutchinson J.M., A radial Basis Function Approach to Financial Time Series Analysis, PhD dissertation, Massachusetts Institute of Technology, (1994), USA

Information and Knowledge management (CIKM) (McLean, Virginia, November 06 – 11, 2000). New York: ACM Press, 2000, pp 389 – 396.

Jensen, M. (1978).Some anomalous evidence regarding market efficiency. Journal of Financial Economics, 6, 95– 101

Khare, R, Pathak, N., S.K., and Sohi, S., 2004. Stock Broker P – Sentiment Extraction for the Stock Market. In A. Zanasi, N.F.F. Ebecken, and C.A. Brebbia eds. *Data Mining V, Proceedings of the Fifth International Conference on data Mining, Text Mining and their Business Applications (Malaga, Spain, Sep. 15 – 17, 2004).* Southampton, Boston: WIT Press, 2004, Vol. 33, pp. 43 – 52.

Kim K. and Lee W., (2004) Stock Market Prediction using ANN with Optimal Feature Transformation, Neural Computing & Applications, Springer. (2004) Vol. 13, No.3, 225-260 (6).

Kwiatkowski, D., Phillips P., and Shin Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics 54 (1-3), 159-78

Lapedes A. and Farber R., Nonlinear signal processing using neural network: prediction and system modeling, Los Alamos National Laboratory (1987)

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Alla J., (2000).Language Models for Financial News Recommendation. In *Proceedings of the 9th International Conference on*

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Alla J., 2000.Mining of Concurrent Text and Time Series. In *Proceedings of the Workshop on text Mining at the Sixth International Conference on Knowledge Discovery and Data Mining*, (Boston, MA, August 20 – 23, 2000). New York: ACM Press, 2000, pp. 37 – 44.

Layton, A., Defris, L. V., &Zehnwirth, B. (1986).An international comparison of economic leading indicators of telecommunication traffic.International Journal of Forecasting, 2,413– 425.

Leone, R. P. (1987). Forecasting the effect of an environmental change on market performance: An intervention time-series. International Journal of Forecasting, 3, 463–478.

Lin, W. T. (1989). Modeling and forecasting hospital patient movements: Univariate and multiple time series approaches. International Journal of Forecasting, 5, 195–208.

Liu, L. -M., & Lin, M. -W. (1991).Forecasting residential consumption of natural gas using monthly and quarterly time series.International Journal of Forecasting, 7, 3– 16.

Lo A. W., &| MacKinlay A. C. Stock market prices do not follow random walks: evidence from a simple specification test. Rev Financ Stud 1988; 1: 41-66.

Lo AW, MacKinlay AC. The size and the power of the variance ratio test in finite samples: a monte carlo investigation. J Econom 1989; 40: 203-38.

Maddala, G. S. and I. – M. Kim (1998).Unit Roots, Cointegration and Structural Change. Cambridge University Press.

Mentzer J.T., Moon M.A., 2005, Sales Forecasting Management, A Demand Management Approach, Sage Publications Inc.

Mittermayer, M.A., 2004. Forecasting Intraday Stock Price Trends with Text Mining Techniques. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS) (Big Island, Hawaii, January 05 – 08, 2004).* Washington, DC: IEEE Computer Society, 2004, 3(3), pp. 30064.2.

Mun J., 2010, Modeling Risk, 2nd Edition, John Wiley & Sons Inc.

Oller, L. -E. (1985). Macroeconomic forecasting with a vector ARIMA model. International Journal of Forecasting, 1, 143–150.

Pack, D. J. (1990). Rejoinder to: Comments on: In defense of ARIMA modeling by M.D. Geurts and J.P. Kelly. International Journal of Forecasting, 6, 501– 502.

PegahFalinouss 2007 Stock Trend Prediction Using News Articles: A Text Mining Approach Master's Thesis

Pflaumer, P. (1992). Forecasting US population totals with the Box–Jenkins approach. International Journal of Forecasting, 8,329– 338.

Phung, Y.C., 2005. Text mining for Stock Movement Predictions: a Malaysian Perspective.

Remenyi, D., Williams, B., Money, A., Swartz E., (1998) Doing Research in Business and Management. Sage publications, London.

Ripley B.D., (1994). Neural Network and related methods for classification: Journal of the royal statistical society, (1994) series B, vol. 50, 409-556.

Schumaker, R.P., and Chen, H., (2006). Textual Analysis of Stock Market Prediction Using Financial News Articles.On *the 12th Americas Conference on Information Systems (AMCIS, Acapulco, Mexico, August 4 -6, 2006).*

Sejnowski T. J., Yuhas B. P., Goldstein M.H and Jenkins R.E., Combine visual and acoustic speech signals with a neural network improves intelligibility, Advances in Neural Information Processing Systems, (1990) vol. 2, 232-239 case of Bank Failure Predictions, Management Science, (1992) vol. 38, 926-947.

Sykes A.O., (1992), An Introduction to Regression Analysis, Available at: http://www.law.uchicago.edu/files/files/20.Sykes_.Regression. pdf [Accessed 15 Nov 2010]

Tam K.Y. and Kiang M.Y., Managerial Application of Neural Networks: The case of Bank Failure Predictions, Management Science, (1992) vol. 38, 926-947.

Tersine, R.J., Principles of Inventory and Materials Management. Elsevier Science Publishing Co, New York.

Timmermann, A., Granger C.W.J. International Journal of Forecasting 20 (2004) 15–27

Yu, P. S and Liu, B eds. *Proceeding of the 6th Pacific-Asia Conference (PAKDD) on Advances in Knowledge Discovering and Data Mining (Taipei, Taiwan, May 06 – 08, 2002), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2002. Vol. 2336, pp. 481 – 493.

Wang G.C.S., Jain C.L., 2003, Regression Analysis, Modeling and Forecasting, Graceway Publishing Company Inc.

Weller, B. R. (1989). National indicator series as quantitative predictors of small region monthly employment levels. International Journal of Forecasting, 5, 241–247.

Wu S.I. and Zheng H., Modeling Stock Market using Neural Networks.Proceedings of Software Engineering and Applications, ACTA Press, USA, 2003.

Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J., and Lam, W., 1998.Daily Stock Market Forecast from Textual Web Data.In IEEE International Conference on Systems, man, and Cybernetics (San Dirgo, California, October 11 – 14, 1998). IEEE Press, Vol 3, pp. 2720 – 2725.

Zanasi, A., Brebbia, C.A and Ebecken N.F.F eds. Data Mining VI, Proceeding of the 6th international Conference on Data Mining, Text Mining and Their Business Applications (Skiathos, Greece, may 25 – 27, 2004). Southampton, Boston: WIT Press, 2005, vol. 35, pp. 103 – 111.

Zopounidis C. and Pardalos P.M., 1998, Managing in uncertainty: Theory and Practice, Kluwer Academic Publishers.

www.econsult.worldbank.org/suite/public/populateViewPublicNotification.bg

APPENDIX A1: Unit Root Test for Considered Stocks

Table A1.1: Unit Root Test for BG (2005 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
BG	DF	-1.2115 (-3.51)	-2.0607 (-4.04)
	KPSS	1.5786 (0.463)	0.1348 (0.146)
BG/BP	DF	-2.7542 (-3.51)	-3.105 (-4.04)
	KPSS	0.3278 (0.463)	0.065 (0.146)
BG/TLW	DF	-0.4930 (-3.51)	-2.9644 (-4.04)
	KPSS	1.7826 (0.463)	0.1044 (0.146)

Table 5Unit Root Test for Difference BG (2005 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
DBG	DF	-4.4722 (-3.51)	-4.4294 (-4.04)
	KPSS	0.0709 (0.463)	0.0737(0.146)
DBG/DBP	DF	-5.4502 (-3.51)	-5.4191 (-4.04)
	KPSS	0.0735 (0.463)	0.0849 (0.146)
DBG/DTLW	DF	-5.8975 (-3.51)	-5.8596 (-4.04)
	KPSS	0.0447 (0.463)	0.0606 (0.146)

Table 6 Unit Root Test for BG* (2007 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
BG*	DF	-1.9590 (-3.58)	-2.1025 (-4.15)
	KPSS	0.5399 (0.463)	0.1047 (0.146)
BG*/BP*	DF	-2.5038 (-3.58)	-2.6080 (-4.15)
	KPSS	0.1277 (0.463)	0.0668 (0.146)
BG*/TLW*	DF	-1.1559 (-3.58)	-2.5342 (-4.15)
	KPSS	1.1572(0.463)	0.0629 (0.146)

Table A1.4: Unit Root Test for Difference BG* (2007 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
DBG*	DF	-3.5136 (-3.58)	-3.4458 (-4.15)
	KPSS	0.1064 (0.463)	0.1136 (0.146)
DBG*/DBP*	DF	-4.2737 (-3.58)	-4.2094(-3.50)
	KPSS	0.0514 (0.463)	0.0768 (0.146)
DBG*/DTLW*	DF	-4.7198 (-3.58)	-4.6814 (-4.15)
	KPSS	0.0468 (0.463)	0.0779 (0.146)

Table A1.5: Unit Root Test for BP (2005 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
BP	DF	-2.7542 (-3.51)	-3.1050 (-4.04)
	KPSS	0.3278 (0.463)	0.0650 (0.146)
BP/TLW	DF	-0.4930 (-3.51)	-2.9644 (-4.04)
	KPSS	1.7826 (0.463)	0.1044 (0.146)
BP/CNE	DF	-1.3385(-3.51)	-1.7030 (-4.04)
	KPSS	0.5570 (0.463)	0.1136 (0.146)

Table A1.6: Unit Root Test for Difference BP (2005 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
DBP	DF	-5.4502 (-3.51)	-5.4191(-4.04)
	KPSS	0.0735 (0.463)	0.0849 (0.146)
DBP/DTLW	DF	-5.8975 (-3.51)	-5.8596 (-4.04)
	KPSS	0.0447 (0.463)	0.0606 (0.146)
DBP/DCNE	DF	-5.7784 (-3.51)	-5.7417 (-4.04)
	KPSS	0.0951 (0.463)	0.0829 (0.146)

Table A1.7: Unit Root Test for BP* (2007 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
BP*	DF	-2.5038 (-3.58)	-2.6080 (-4.15)
	KPSS	0.1277 (0.463)	0.0668 (0.146)
BP*/CNE*	DF	-0.8400 (-3.58)	-2.1333 (-4.15)
	KPSS	0.7795 (0.463)	0.1097 (0.146)
BP*/TLW*	DF	-1.1559 (-3.58)	-2.5342 (-4.15)
	KPSS	1.1572 (0.463)	0.0629 (0.146)

Table A1.8: Unit Root Test for Difference BP* (2007 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
DBP*	DF	-4.2737 (-3.58)	-4.2094 (-4.15)
	KPSS	0.0514 (0.463)	0.0768 (0.146)
DBP*/DCNE*	DF	-4.7198 (-3.58)	-4.6814 (-4.15)
	KPSS	0.0468 (0.463)	0.0779 (0.146)
DBP*/DTLW*	DF	-5.2842 (-3.58)	-5.1975 (-4.15)
	KPSS	0.1417 (0.463)	0.0731 (0.146)

Table A1.9: Unit Root Test for CNE (2005 - 2010)

Ticker	Test Type	Test Statistics	
		Constant	Constant + Trend
CNE	DF	-1.3385 (-3.51)	-1.7030 (-4.04)
	KPSS	0.5570 (0.463)	0.1136 (0.146)
CNE/BP	DF	-2.7542 (-3.51)	-3.1050 (-4.04)
	KPSS	0.3278 (0.463)	0.0650 (0.146)
CNE/TLW	DF	-0.4930 (-3.51)	-2.9644 (-4.04)
	KPSS	1.7826 (0.463)	0.1044 (0.146)

Table A1.10: Unit Root Test for Difference CNE (2005 - 2010)

Ticker	Test Type	Test Statistics	
		Constant	Constant + Trend
DCNE	DF	-5.4502 (-3.51)	-5.4191 (-4.04)
	KPSS	0.0735 (0.463)	0.0849 (0.146)
DCNE/DBP	DF	-5.4502 (-3.51)	-5.4191 (-4.04)
	KPSS	0.0735 (0.463)	0.0849 (0.146)
DCNE/DTLW	DF	-5.8975 (-3.51)	-5.8596 (-4.04)
	KPSS	0.0447 (0.463)	0.0606 (0.146)

Table A1.11: Unit Root Test for CNE* (2007 - 2010)

Ticker	Test Type	Test Statistics	
		Constant	Constant + Trend
CNE*	DF	-0.8400 (-3.58)	-2.1333 (-4.15)
	KPSS	0.7795 (0.463)	0.1097 (0.146)
CNE*/BP*	DF	-2.5038 (-3.58)	-2.6080 (-4.15)
	KPSS	0.1277 (0.463)	0.0668 (0.146)
CNE*/TLW*	DF	-1.1559 (-3.58)	-2.5342 (-4.15)
	KPSS	1.1572 (0.463)	0.0629 (0.146)

Table A1.12: Unit Root Test for Difference CNE* (2007 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
DCNE*	DF	-5.2842 (-3.58)	-5.1975 (-4.15)
	KPSS	0.1417 (0.463)	0.0731(0.146)
DCNE*/DBP*	DF	-4.2737 (-3.58)	-4.2094 (-4.15)
	KPSS	0.0514 (0.463)	0.0768 (0.146)
DCNE*/DTLW*	DF	-4.7198 (-3.58)	-4.6814 (-4.15)
	KPSS	0.0468 (0.463)	0.0779 (0.146)

Table A1.13: Unit Root Test for TLW (2005 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
TLW	DF	-0.4930 (-3.51)	-2.9644 (-4.04)
	KPSS	1.7826 (-0.463)	0.1044 (0.146)
TLW/BP	DF	-2.7542 (-3.51)	-3.1050 (-4.04)
	KPSS	0.3278 (0.463)	0.0650 (0.146)
TLW/BG	DF	-1.2115 (-3.51)	-2.0607 (-4.04)
	KPSS	1.5786 (0.463)	0.1348 (0.146)

Table A1.14: Unit Root Test for Difference TLW (2005 - 2010)

Ticker	Test Type	Test Statistics Constant	Constant + Trend
DTLW	DF	-4.4722 (-3.51)	-4.4294 (-4.04)
	KPSS	0.0709 (0.463)	0.0737 (0.146)
DTLW/DBP	DF	-5.4502 (-3.51)	-5.4191 (-4.04)
	KPSS	0.0735 (0.463)	0.0849 (0.146)
DTLW/DBG	DF	-5.8975 (-3.51)	-5.8596 (-4.04)
	KPSS	0.0447 (0.463)	0.0606 (0.146)

Table A1.15: Unit Root Test for TLW* (2007 - 2010)

Ticker	Test Type	Test Statistics	
		Constant	Constant + Trend
TLW*	DF	-1.1559 (-3.58)	-2.5342 (-4.15)
	KPSS	1.1572 (0.463)	0.0629 (0.146)
TLW*/BP*	DF	-2.5038 (-3.58)	-2.608 (-4.15)
	KPSS	0.1277 (0.463)	0.0668 (0.146)
TLW*/BG*	DF	-1.959 (-3.58)	-2.1025 (-4.15)
	KPSS	0.5399 (0.463)	0.1047 (0.146)

Table A1.16: Unit Root Test for Difference TLW*(2007 - 2010)

Ticker	Test Type	Test Statistics	
		Constant	Constant + Trend
DTLW*	DF	-4.7198 (-3.58)	-4.6814 (-4.15)
	KPSS	0.0468 (0.463)	0.0779 (0.146)
DTLW*/DBP*	DF	-4.2737 (-3.58)	-4.2094 (-4.15)
	KPSS	0.0514 (0.463)	0.0768 (0.146)
DTLW*/DBG*	DF	-3.5136 (-3.58)	-3.4458 (-4.15)
	KPSS	0.1064 (0.463)	0.1136 (0.146)

APPENDIX A2: Time Plot for Considered Stocks with its Exogenous Variables

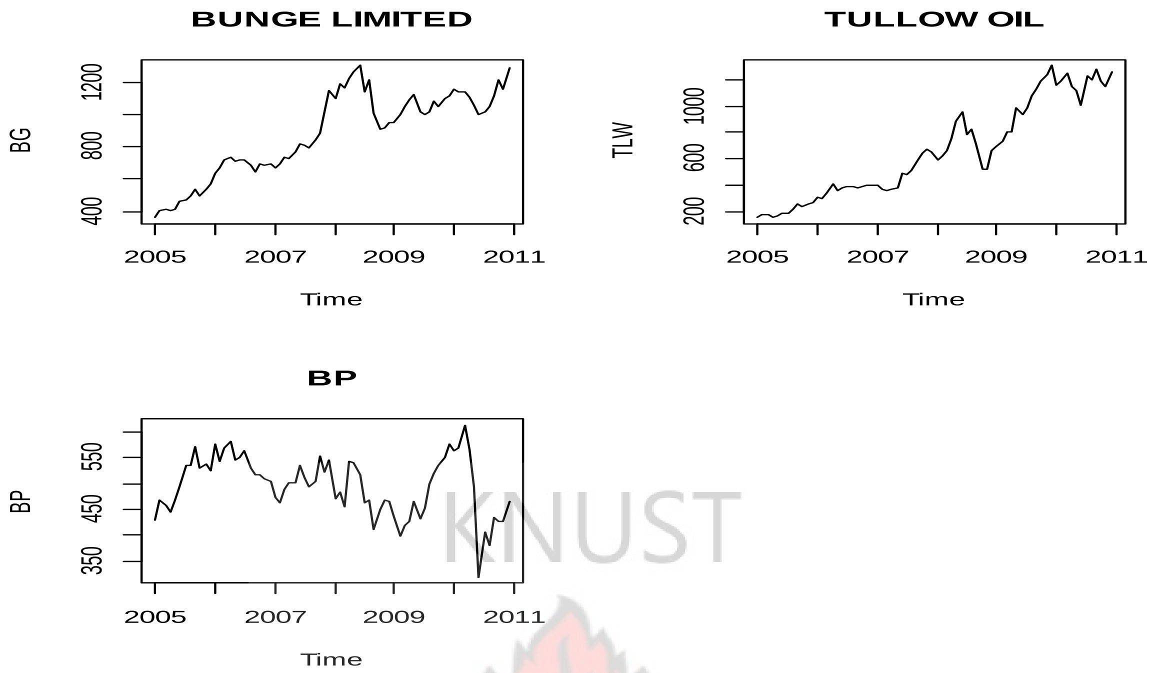

Figure A2.1: Time Plot for BG and its Exogenous Variables (2005 - 2010)

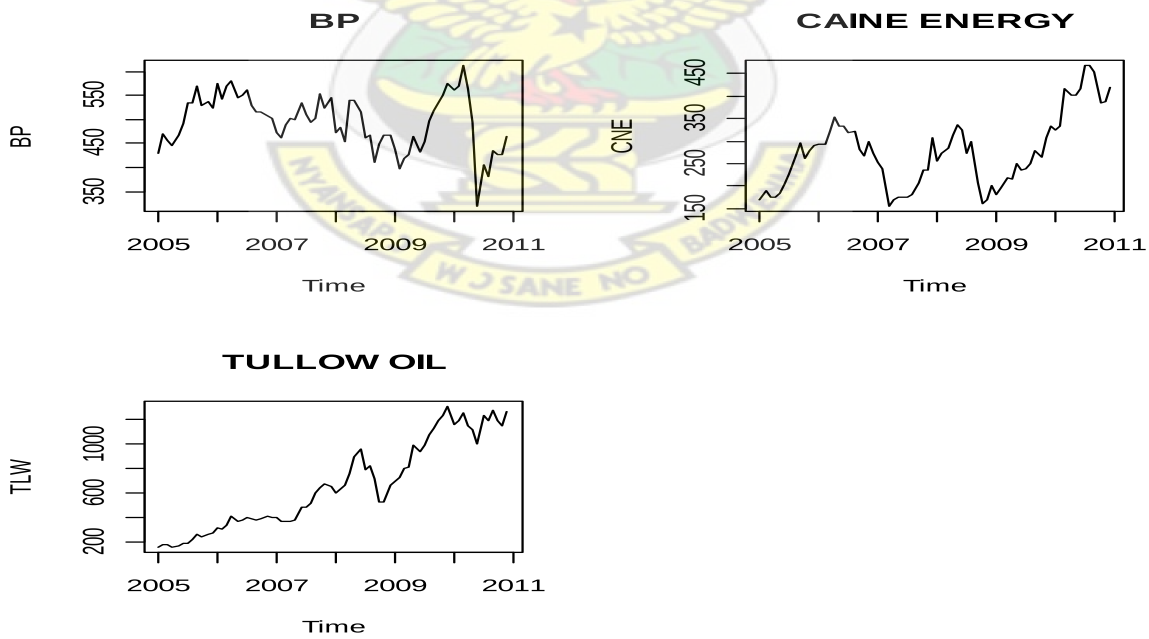

Figure A2.2: Time Plot for BP and its Exogenous Variables (2005 - 2010)

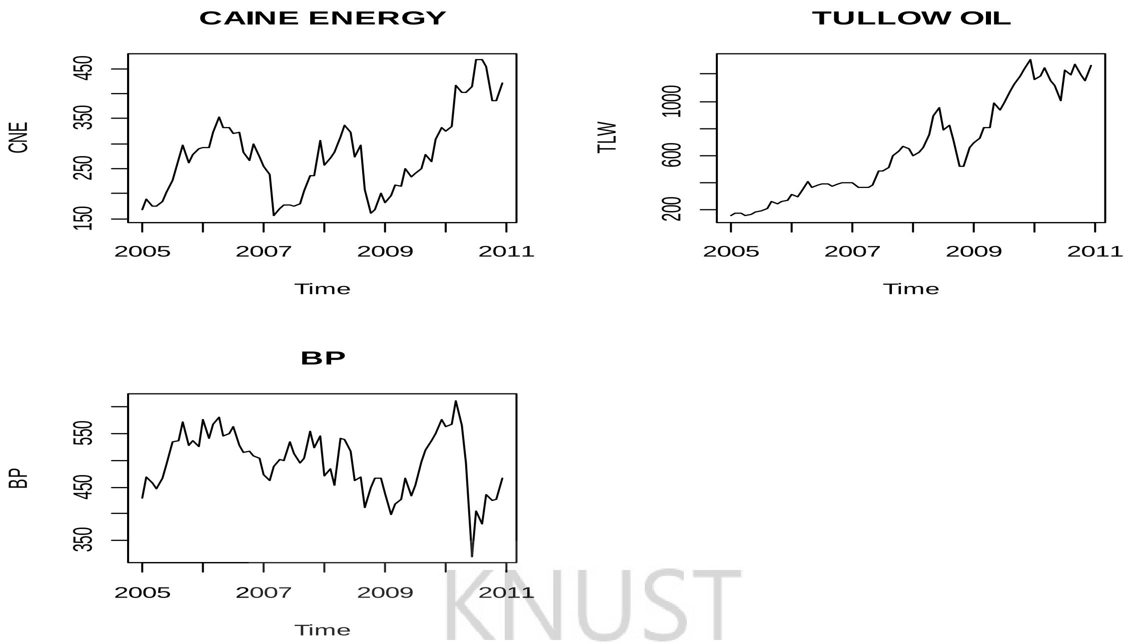

Figure A2.3: Time Plot for CNE and its Exogenous Variables (2005 - 2010)

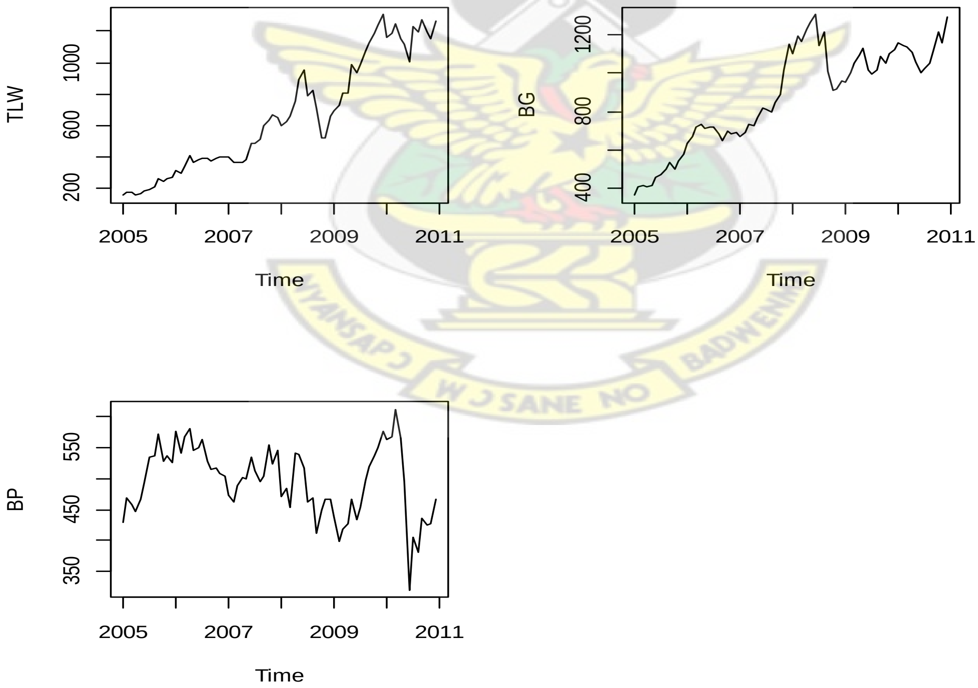

Figure A2.4: Time Plot for CNE and its Exogenous Variables (2007 - 2010)

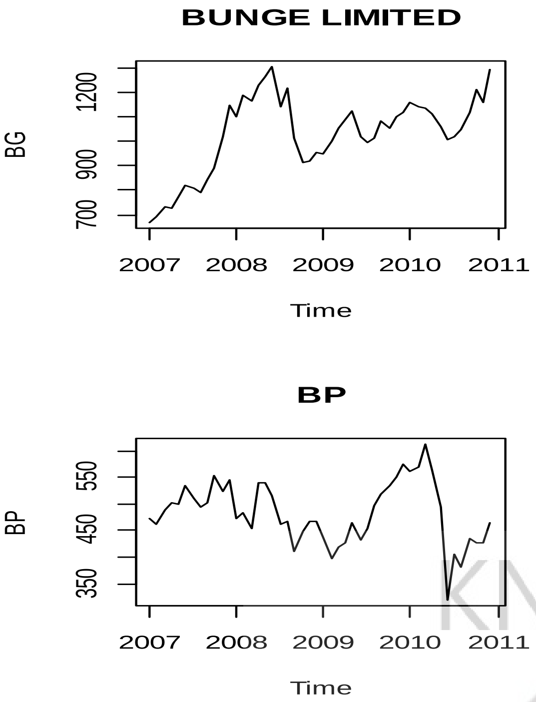

Figure A2.5: Time Plot for TLW and its Exogenous Variables (2005 - 2010)

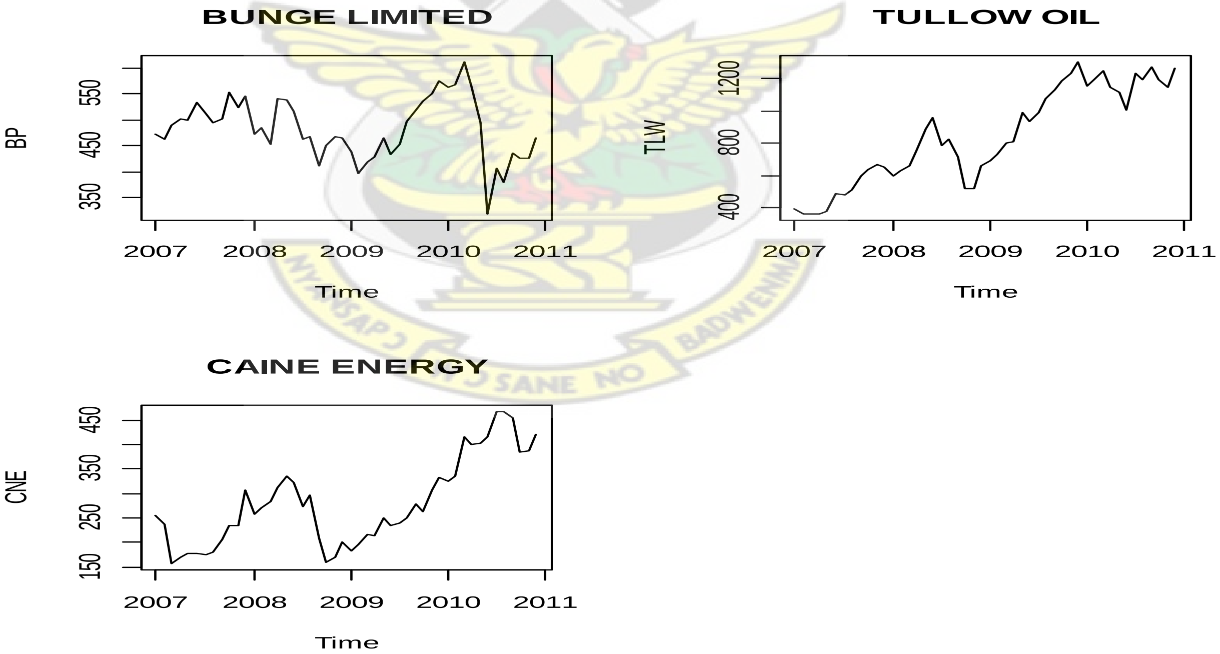

Figure A2.6: Time Plot of BP* and its Exogenous Variables (2007 - 2010)

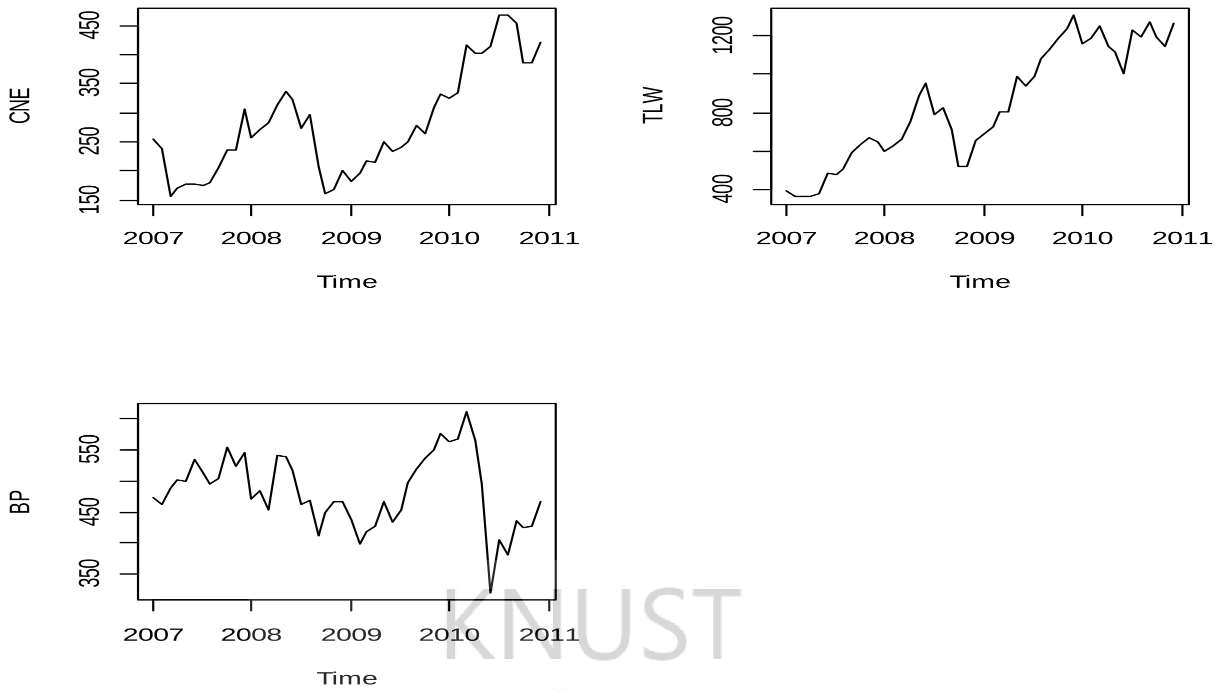

Figure A2.7: Time Plot of CNE* and its Exogenous Variables (2007 - 2010)

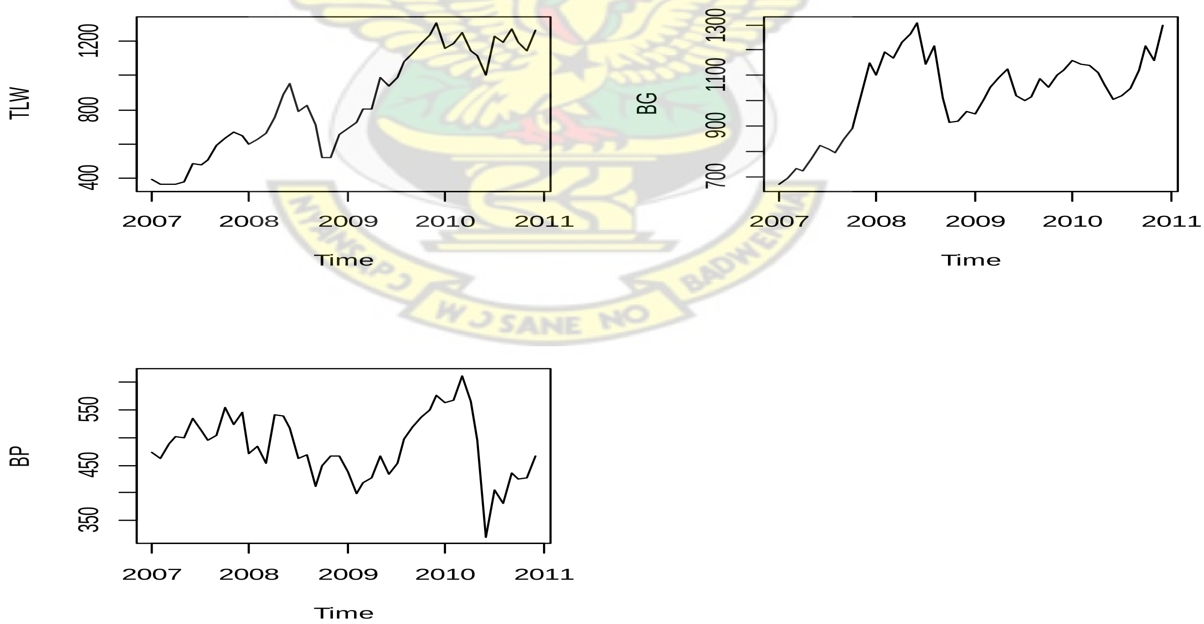

Figure A2.8: Time Plot of TLW* and its Exogenous Variables (2007 - 2010)

APPENDIX A3: **Test for Random Walk**.

1. **Variance Ratio Test - BUNGE LIMITED (BG)**

Null Hypothesis: random walk with homoskedastic errors

	var.ratio	**std.err**	**stat**
2	0.9855	0.1187	-0.121784
3	1.1731	0.1769	0.978493
4	1.2695	0.2220	1.214009
5	1.3224	0.2600	1.240097

Null Hypothesis: random walk with heteroskedastic errors

	var.ratio	**std.err**	**stat**
2	0.9855	0.1595	-0.090616
3	1.1731	0.2456	0.704762
4	1.2695	0.3083	0.874245
5	1.3224	0.3535	0.912249

* : significant at 5% level
** : significant at 1% level

2. **Variance Ratio Test - BP (BP)**

Null Hypothesis: random walk with homoskedastic errors

	var.ratio	**std.err**	**stat**
2	0.8551	0.1187	-1.2209
3	0.9208	0.1769	-0.4476
4	0.7974	0.2220	-0.9125
5	0.7584	0.2600	-0.9291

Null Hypothesis: random walk with heteroskedastic errors

	var.ratio	**std.err**	**stat**
2	0.8551	0.1947	-0.7440
3	0.9208	0.2740	-0.2890
4	0.7974	0.3295	-0.6148
5	0.7584	0.3721	-0.6493

* : significant at 5% level
** : significant at 1% level

3. **Variance Ratio Test - CAINE ENERGY (CNE)**

Null Hypothesis: random walk with homoskedastic errors

	var.ratio	**std.err**	**stat**
2	0.8551	0.1187	-1.2209
3	0.9208	0.1769	-0.4476
4	0.7974	0.2220	-0.9125
5	0.7584	0.2600	-0.9291

Null Hypothesis: random walk with heteroskedastic errors

	var.ratio	**std.err**	**stat**
2	0.8551	0.1947	-0.7440
3	0.9208	0.2740	-0.2890
4	0.7974	0.3295	-0.6148
5	0.7584	0.3721	-0.6493

* : significant at 5% level
** : significant at 1% level

4. **Variance Ratio Test – TULLOW OIL (TLW)**

Null Hypothesis: random walk with homoskedastic errors

	var.ratio	**std.err**	**stat**
2	1.0112	0.1187	0.09433
3	1.0309	0.1769	0.17454
4	1.0871	0.2220	0.39209
5	1.1732	0.2600	0.66601

Null Hypothesis: random walk with heteroskedastic errors

	var.ratio	**std.err**	**stat**
2	1.0112	0.1165	0.09608
3	1.0309	0.1714	0.18013
4	1.0871	0.2133	0.40821
5	1.1732	0.2483	0.69740

* : significant at 5% level
** : significant at 1% level

APPENDIX A4: WORKING DATA

BG	BP	CNE	TLW
361.75	429.91	168.44	156.75
408.25	469.21	189.96	178
411.25	457.94	176.17	173.75
405.25	445.83	175.56	160.25
416.75	467.3	184.05	167
459	492.74	204.19	186.5
471.25	533.87	227.22	191.25
499.5	535.99	268.27	215.5
538	571.19	297.05	260
496	529.21	261.6	242.5
540.5	537.69	277.21	264
574.5	524.97	290.84	270
635	576.26	291.9	313.5
667.5	541.96	292.36	298.5
719.5	568.63	321.89	339
737	581.96	352.34	411.5
710	545.14	332.19	363.5
722.5	549.94	332.19	382
720	562.58	319.77	394.25
686.5	529.18	322.95	392.25
649	515.89	282.51	377
695.5	516.77	265.85	389.5

684.5	508.8	298.41	405.25
693	503.04	272.51	398
668.5	473.78	254.03	398.5
692	462.7	237.52	367.25
733	489.3	155.5	364.75
725.5	501.26	169.6	366.5
772	500.38	176.5	380.75
821.5	534.5	176.4	488
808	511.9	174.6	483.5
794	494.17	179.7	512
846	503.04	206.4	596
889.5	554	235.3	638
1018	522.98	234.9	671
1150	545.14	307.4	651.5
1100	471.57	256.5	597
1192	483.98	271.7	627.5
1167	453.84	283.4	660.5
1231	541.59	313.4	754
1266	538.93	336.6	890
1307	517	323.5	955
1146	462.26	274	787.5
1219	468.69	297.8	825
1013	411.29	207.2	713
913	449.63	160.4	524
920	466.91	169	522.5

957	466.25	201.5	659.5
950.5	438.33	182	692
1005	397.33	196.1	731
1055	417.94	217.5	803
1093	427.69	214.3	806.5
1128	465.56	249	990
1018	432.86	234.2	937.5
999.4	452.99	239.6	988.5
1017	498.32	250.7	1077
1087	519.31	278.7	1128
1055	536.07	264.2	1187
1103	551.23	308	1234
1122	575.59	332.6	1305
1159.5	563.03	325	1157.05
1145	568.7	334.4	1189
1140.5	612.62	417	1250
1113	565.55	401.3	1147
1061	494.8	402.9	1117
1006	319.36	414.9	1003
1021.5	405.95	466.8	1231
1049.5	380.6	467.4	1193.78
1118.5	435.05	453.6	1274
1215.5	425.8	385.9	1191.74
1161.5	425.95	386.5	1147
1296	465.55	420	1261

APPENDIX A5: LISTED INDUSTRIES ON THE LONDON STOCK MARKET

1. AEROSPACE & DEFENSE
2. ALTERNATIVE ENERGY
3. AUTOMOBILES & PARTS
4. BANKS
5. BEVERAGES
6. CHEMICALS
7. CONSTRUCTION & MATERIALS
8. ELECTRICITY
9. ELECTRONIC & ELECTRICAL EQUIPMENT
10. EQUITY INVESTMENT INSTRUMENTS
11. FINANCIAL SERVICES
12. FIXED LINE TELECOMMUNICATION
13. FOOD PRODUCERS
14. FOOD & DRUG RETAILERS
15. FORESTRY & PAPER
16. GAS, WATER & MULTIUTILITIES
17. GENERAL INDUSTRIALS
18. GENERAL RETAILERS
19. HEALTH CARE EQUIPMENT & SERVICES
20. HOUSEHOLD GOODS & HOME CONSTRUCTION
21. INDUSTRIAL ENGINEERING
22. INDUSTRIAL METALS & MINING
23. INDUSTRIAL TRANSPOTATION
24. LEISURE GOODS

25. LIFE INSURANCE

26. MEDIA

27. MINING

28. MOBILE TELECOMMUNICATIONS

29. NONEQUITY INVESTMENT INSTRUMENTS

30. NONLIFE INSURANCE

31. OIL & GAS INDUSTRY

32. OIL EQUIPMENT & SERVICES

33. PERSONAL GOODS

34. PHARMACEUTICALS & BIOTECHNOLOGY

35. REAL ESTATE INVESTMENT & SERVICES

36. REAL ESTATE INVESTMENT & TRUST

37. SOFTWARE & COMPUTER SERVICES

38. SUPPORT SERVICES

39. TECHNOLOGY HARDWARE & EQUIPMENT

40. TOBACCO

41. TRAVEL & LEISURE