National College of Ireland

Analyzing Historical Stock Market Data to Determine if a Correlation Exists Between Major Stock Market Indexes and if Time Series is Sufficient to Make Predictions

MSc Research Project
Data Analytics

Anicia Lafayette-Madden
x15006590

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Anicia Lafayette-Madden
Student ID:	x15006590
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Vikas Sahni
Submission Due Date:	22/08/2016
Project Title:	Analyzing Historical Stock Market Data to Determine if a Correlation Exists Between Major Stock Market Indexes and if Time Series is Sufficient to Make Predictions
Word Count:	5500

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	15th September 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analyzing Historical Stock Market Data to Determine if a Correlation Exists Between Major Stock Market Indexes and if Time Series is Sufficient to Make Predictions

Anicia Lafayette-Madden
x15006590
MSc Research Project in Data Analytics

15th September 2016

Abstract

Time series for forecasting stock market prices has become extremely popular over the last few decades. Both stand-alone time series algorithms and hybrids have been successfully implemented, with several papers documenting their positive results. Correlation and volatility across the major markets have also seen quite a number of works conducted with positive results being documented. This research has examined the existence of a correlation between major stock market prices and the effectiveness of time series for forecasting future prices. A cross-correlation matrix and ARIMA time series algorithm for forecasting were implemented. The results, while proving correlation exists between the major stock markets, was inconclusive when examining the percentage change in US GDP growth and its possible effect on the change in correlation. Time series results were unexpected, however, this researcher strongly believes it is indeed effective in making stock market price predictions.

Contents

1 Introduction

This research aims to determine the level of correlation and volatility existing between historic stock market price data, by examining indexes from major global stock markets. It also aims to determine whether time series is sufficient to make predictions. This research has utilized semi-annual Gross Domestic Product (GDP) growth data from 1991 to 2015, and has examined its effect on the change in correlation between stock market prices internationally. Many investors depend on economic news in order to decide whether to invest in the stock market or not. If volatility is expected of the market, it will be affected by particular economic data releases, as investors will be keen on waiting for this release before investing. This lead to investor speculation affecting the market either positively or negatively. Stock market correlation and volatility across major stock markets have received a growing level of interest, however, not as much work is available on the subject matter. Historic stock market price volatility is normally derived using time series applied to historical stock price data. This research has used cross-correlation analysis, along with time series model Autoregressive Integrated Moving Average (ARIMA), which is ideal when working with stock market data considered to be time-based and for serially correlated.

A healthy economy means a more stable and strong stock market, thus different economic factor will have some effect on the stock market and possibly other related markets. Economic factors such as unemployment rates, interest rates, exchange rate and economic growth are good determinants in examining market relationship. This can aid in the efforts to enhance long-term predictive abilities. Technological advancements in the last decade have seen a rapid migration to the use of artificial intelligence as a means of analyzing the vast amounts of stock data now available. While there are many types of research available on time series forecast of stock market prices, the same cannot be said for studies done on the correlation of global stock market indexes. For that reason, this research has focused more of the correlation aspect while testing the effectiveness of time series in forecasting. This research seeks to successfully apply time series analysis to find patterns in the stock market data and accurately forecasting stock market prices to gain valuable insight to more effectively optimize business decision making. This research is also motivated by building on previous work done in the area of stock market price daily correlation dynamics and volatility across the global market.

2 Background

An important part of the world's free market economy, the stock market holds shares for many companies that are traded publicly through exchanges daily. With continued globalization and increasing international trading, a massive amount of capital is being traded on the stock market around the world. Investors buy and sell shares thus making it an important way for companies to raise money or financial capital in order to expand their businesses. Undoubtedly one of the most analyzed field in the financial sector, the stock market is referred to as unpredictable by most who have attempted its forecasting.

2.1 Stock market price correlation and volatility

Correlation refers to the movement/trend of stock market index prices in relation to each other, while volatility speaks of the variation over time which is measured using

the standard deviation. Economic factors such as unemployment rates, interest rate, currency revaluation, GDP growth and government spending have significant effects on stock market prices. Economic growth makes companies more profitable as the demand for goods and services increase, leading to increases in stock price inflation (Tanning et al., 2013). Higher interest rates lead to people putting more of their earnings into paying off debt. This affects the amount of money available for investing in the stock market with investors opting to leave that market for better opportunities in other markets having better interest rates. A lower unemployment rate indicates when an economy is healthy and strong. More jobs and higher employment results in people having more money to make investments. If these variables can be taken into account when analyzing the dynamics of correlation and volatility across major stock markets, this will bring valuable insight and knowledge.

2.2 Time Series for Stock Market forecasting

In the analysis of historical stock market data, many different techniques have been put forward along with their accompanying success stories. The ability to create algorithms that will revolutionize stock market forecasting is always being attempted. This research is interested in the effectiveness of time series for the forecasting stock market data. Investors and analyst are convinced that by applying artificial intelligence to stock market data, they will be able to predict stock market prices, therefore, only investing when it is most profitable. Some of the most popular time series algorithms that have been utilized for stock market analysis and predictions include Artificial Neural Networks (ANN), Support Vector machines (SVM), ARIMA and Hidden Markov Model (HMM).

2.3 Scope and Objective

2.3.1 Scope

The scope of this research is to analyze historic stock market data from major stock markets to identify correlation that exists, and to determine if time series forecasting is sufficient to make adequate future stock market price predictions.

2.3.2 Objectives

The research will seek to identify times of US economic data releases (semi-annual) to determine if this affected the change in the correlation of major stock prices indexes. This analysis utilizes data over the time period 1991 to 2015.

2.4 Research Question

This research aims to:

To determine if a correlation exists between major stock market indexes.

To determine the effectiveness of time series analysis is in making predictions.

2.5 Data Description

Six sets of data covering the time period January 1991 to December 2015 were extracted from Yahoo Finance to CSV and includes 6519 rows of data. Stock market index

information was downloaded on the following:

Nikkei 225 - This is the Tokyo Stock Exchange (TSE) most respected stock market index. It is a price-weighted index which includes the top 225 firms in Japan (Hamao et al., 1990).

FTSE 100 index - Refers to the London Stock exchange listing for the Financial Times Stock Exchange 100 index. It represents the majority of the market capitalization and is an equity value weighted index (Hamao et al., 1990).

DJIA - Refers to the Dow Jones Industrial Average is a price-weighted average that widely used as a stock market indicator.

NDX - The Nasdaq 100 is a stock market index listed on the Nasdaq 100 (National Association of Securities Dealers Automated Quotations). It is a capitalization-weighted index and is composed of the 100 largest most actively companies listed on the Nasdaq 100 stock exchange.

S and P 500 - Refers to the American stock market index Standard and Poor's 500 which is considered the leading equities indicator. It is market value weighted and one of the most used benchmark for the U.S. stock market.

SSE - Refers to the Shanghai Stock Exchange composite index and is a market capitalization-weighted index.

Each dataset provided data in that countrys currency with columns containing the following:

Date = The trading date.

Open = The price of stock at the beginning of that day.

High = The highest price of the stock for that trading day.

Low = The lowest price of the stock for that trading day.

Closing = The price of the stock at the end of that days trading.

Volume = The number of shares traded on that day.

Adjusted Close = Represents the difference open and closing after trading hours.

Economic data relating to GDP growth in the US between 1991 and 2015 was extracted from the FRED Economic Data at https://fred.stlouisfed.org/series/GDP/downloaddata. The file downloaded contained semi-annual data with 50 rows of data and includes:

Date = Date of data release.

Value = Percentage increase in GDP growth

2.6 Research Overview

This research paper consists of the follow:

Background - Research specification including research question and definition of research variables.

Literature Review An examination of previous and recent work done in the field regarding the analysis and forecasting of stock market data, the techniques used and recommendations for improvements.

Solution and development - Information on techniques used for analysis along with a description of test implementation.

Results and Evaluation Presentation of results and findings with the summary discussion.

Conclusion - Provides the overall argument for the research project and justification for further research in the field.

3 Literature Review

Different time series techniques have been applied to stock market prices to analyze its movement in relation to other major markets, and to forecast future prices. The literature review presents the findings.

3.1 Correlation and volatility

International stock market index correlation and volatility have been studied over the last few decades, with researchers finding evidence of correlation and volatility existing during different conditions. The assessments have seen different versions of the family of Autoregressive Conditional Heteroskedastic (ARCH) statistical model used to explore the relationships and volatility existing between international stock markets. Short term risk management have led to the increase in the modelling of the day-to-day correlation of stock market prices. However, the difference in trading time has made it difficult to observe these correlations directly (Martens and Poon, 2001). ARCH models have been the most popular tools for dealing with time series heteroskedastic models and estimating conditional variance (Ledoit, et al., 2003). Generalize Autoregressive Conditional Heteroskedastic (GARCH) models are the most used of all the ARCH models. This popularity is due to its ability to provide volatility measures such as standard deviation, useful in risk analysis for areas such as financial decision making, portfolio selection and derivative pricing (Engle, 2001). It is considered the most robust and simple of all the other ARCH models and can be extended or modified. GARCH's ability to include real-world context, as it relates to financial modelling when attempting to make financial price predictions, also makes it the most popular.

In one research studying the interdependence of stock prices, and the volatility of major stock markets, the authors focused on three stock markets indexes (Nikkei 225 index, FTSE index, and the S and P 500 Composite index). By applying the GARCH model, the research was able to capture the effect of the change in volatility in the time-series market data (Hamao et al., 1990). This was done by examining daily and intraday stock price activities from 1985 to 1988. Building on previous research, a different approach was taken by Ramchand and Susmel (1998). Here univariate and bivariate SWARCH techniques were used to test the hypothesis that correlations across major markets are dependent on the endogenously determined variance regime. Daily correlation dynamics between international markets was examined by employing both a non-synchronous and synchronous procedures utilizing the Rickmetrics and GARCH methods (Martens and Poon, 2001).

The transmission of volatility has been an important and major issue of contention, as its minimization seems now to be important. Previous research findings report that volatility appeared to be contagious across markets (Solnik et al., 1996), with strong volatility spillover reported from the US market. Results obtained by Poon and Martens (2001) were similar to that of Hamao et al. (1990), however, reverse spillovers to the US market was better explained by Poon and Martens (2001). They determined that under extremely adverse conditions large negative returns were registered, the only time an increase in daily correlation was obvious. Studying the dynamics of price and volatility spillovers between G-7 countries, Yan and Doong (2004) revealed that movements in stock prices had a greater impact on future exchange rates movements than vice versa. They raised the important point that the examination of volatility spillover could increase

knowledge and understanding about information transmission between the stock prices and exchange rate. Yan and Doong (2004) stated that economic financial globalization and integration was significantly affected by the advancements in information technology, which in turn affected the transmission of volatility across the market. This view was taken into account when this research chose to evaluate the change in correlation between the major stock market indexes in relation to the changes in semi-annual US GDP growth.

While the results of previous research provided valuable insights and raised important avenues for exploration, the issue of non-synchronous trading time across international markets had to be addressed. Trying to align the stock price data based on daily trading time, resulted in the S and P 500, and the Nikkei's previous day closing price being used as a substitute for stocks of the index that were not opened (Hamao et al., 1990) (Martens and Poon, 2001). To avoid this issue, along with the issue of having noise associated with daily data, weekly data were used instead (Ramchand and Susmel, 1998) (Martens and Poon, 2001). This paper dealt with the issue of non-synchronous trading time across international markets by focusing on the end of day stock closing prices.

3.2 Time Series for Stock Market forecasting

3.2.1 Standalone Time Series Algorithms

In both investigating and forecasting stock market returns, different time series algorithms have been applied to historic stock market data. Support Vector Machines has been one of the most popular used algorithms for forecasting stock market prices (Shen et al., 2012). They are viewed as promising to the field (Kim, 2003) due to their ability to use risk/decision function (Huang et al., 2005), which uses regularization for structural risk minimization principle (Kim, 2003). HMM has been regularly used for analyzing and predicting time series phenomena due to its strong mathematical structure and forecasting abilities (Hassan, and Nath, 2005) (Hassan et al, 2007).

In a 2005 study, the predictive power of financial and economic variables using ANN was proposed and implemented by Enke and Thawornwong. This tested ANN's effectiveness for level estimation and classification, where the algorithm was chosen because of its ability to select variables when non-linearity exists within the data, and the variable's usefulness unknown. Because financial data is considered to be unpredictable, unstructured and noisy (Huang et al., 2005) (Choudhry and Garg, 2008), it is important to select a technique capable of dealing with these issues. The technique proved successful and was recommended for future use. Support Vector Machines was used to predict the next-day stock trend of the NASDAQ, S and P 500, and DJIA, resulting in accuracy levels of 74.4, 76, and 77.6 percent respectively (Shen et al., 2012). In addition, SVM performed very well for forecasting the weekly movement of the Nikkei 225 index (Huang et al., 2005). Both HMM and ANN were applied to predict same day closing prices and search for variables with interesting behavioural data patterns (Hassan, and Nath, 2005). The experiment revealed that both ANN and HMM predictive accuracy were similar, however, due to ANN's black box characteristic, the inability to properly explain the model was a limitation (Hassan, and Nath, 2005). HMM is explainable which makes it a more favourable choice over ANN for time series forecasting (Hassan, and Nath, 2005). However, Hasan and Nath thought that efficiency and accuracy could be improved if made a part of a hybrid system, which was successfully implemented two years later.

3.2.2 Hybrid Time Series Algorithms

A few different hybrid time series algorithms have been implemented in the forecasting of financial stock market movement. The hybridized system including HMM, ANN and GA were proposed and implemented using stock market data from the IT sector (Hassan et al, 2007), while the PCA, neuro-fuzzy and ANN hybrid utilized data from the Nasdaq 100 to predict next day stock prices. By applying each algorithm to different aspects of the analysis, fusing the time series algorithms into a hybrid produced a higher performance compared to basic models (Hassan, et al, 2007) (Abraham et al., 2001). GA and PCA are used for removing irrelevant instances from the training data and to optimize the initial parameters (Hassan et at., 2007) (Kim, 2006) (Abraham et al, 2001); ANN was used for transforming values and fine tune the parameters for input into the neuro-fuzzy system; while the HMM was used to find patterns within the data (Abraham et al., 2001) (Kim, 2006). Time series algorithms SVM and ARIMA were combined to forecast stock prices using ten stocks along with their daily closing prices (Pai and Lin, 2005). The hybrid algorithm was thought to be much better than the single model but still needed improving. Just simply mashing two algorithms together for the sake of creating a hybrid is not enough and does not necessarily produce the best outcomes (Pai and Lin, 2005). The overall conclusion was that exploring creative and effective methods can yield better time series forecasting performance (Shen et al., 2012).

4 Solution development

4.1 Research methodology and specification

This research implemented an algorithm capable of forecasting stock market price. Research has shown that approaches taken previously have been fruitful, however, with a continued change in the types of data, techniques/technologies availability and their advancements, better analysis and predictions can be achieved.

4.1.1 Methodology

This paper has followed a number of steps in the analysis and forecasting of time series data. As financial time series data has become one of the most analyzed in today's data mining driven world, the ARIMA algorithm has become very popular. This experiment utilizes closing stock price data between the period 1991 and 2015 and includes six stock market indexes from four major stock markets around the world. It is also important to note that correlation calculations will be performed across stock market price indexes as it relates to US GDP growth rate.

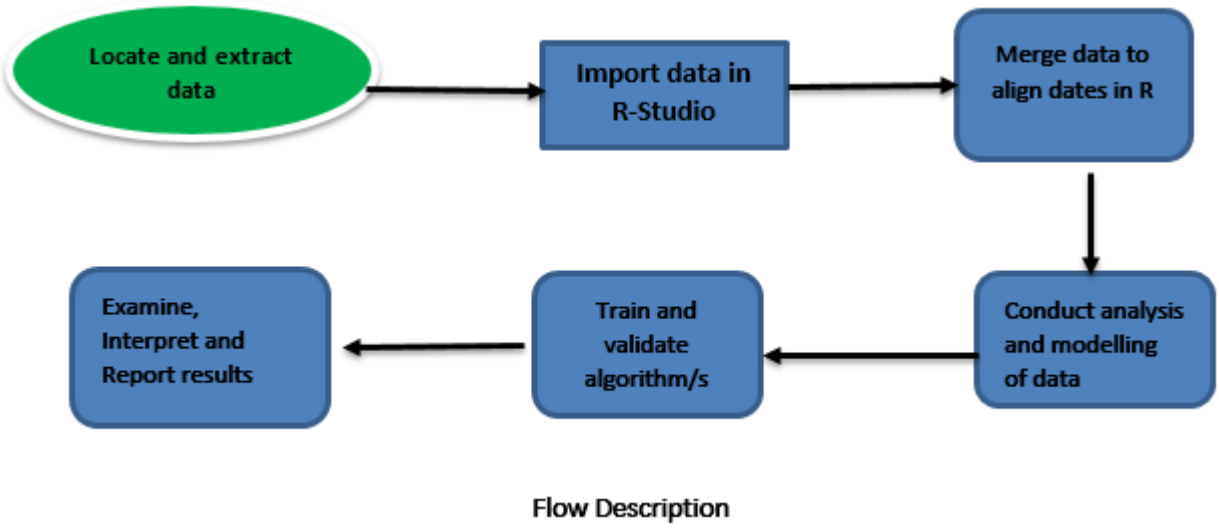


Diagram 1 -Design Work Flow

4.1.2 Correlation Matrix

The correlation matrix is used to determine how dependent multiple variables are on each other on the same table. The degree of the relationship is calculated for all the variables at the same time. Here the sample variance of the variables is used, which is displayed on a table. Correlation in this paper was not only used to examine long-term relationships between prices, but also the medium-term relationships (by examining the data semi-annually based on US GDP growth rate).

4.1.3 ARIMA

One of the most used time series forecasting algorithms (Pai and Lin, 2005), the ARIMA was introduced by George Box and Gwilym Jenkins in the early 1970s and is normally referred to as the Box-Jenkins approach. The model's linear capability to deal with non-stationary characteristics normally associated with stock prices (Box et al., 2015), makes it an ideal and popular algorithm used for forecasting stock market prices. The ARIMA model does not utilize independent variables to make predictions, but rather it reflects the linear combination of past values and error as the future value (Pai and Lin, 2005). This, however, limits the ARIMAs abilities for predicting future values but its ease of use and implementation have made it very popular over the last three decades (Zhang, 2003). One of its dependents is on the existence of autocorrelation within the data and employs the ACF (Autocorrelation Function) in order to identify viable candidates for the model, while also visually showing if the data is stationary or non-stationary.

ARIMA (p, d, q) is a version of the ARMA model where:
p = autoregressive terms
d = number of non-seasonal differences needed for stationarity
q = the moving average lags
The ARIMA can be denoted by:
yt = + 1 yt-1 + 2 yt-2 +. + p yt-p + et 1 et-1 2 et-2 qe-0t-q, where:
yt = the predicted value
j 1 = the coefficients

et = the random error at time t

4.2 Implementation Steps

The approach in regards to analysis and research conducted for this project includes graphs and statistical calculations to determine correlated trends within the data. For time series analysis and forecasting, this approach is important as it allows the researcher to better identify trends that exist within historical stock market prices. This was done by:

Gathering all the necessary data- Stock market data was downloaded from Yahoo Finance website, which includes daily closing stock prices going back 24 years (1991 to 2015). Data on the percentage change in US GDP growth rate was downloaded from Federal Reserve Economic Data website.

Cleaning and transforming data- Each CSV file was then transformed to only include the closing price and trading date for each stock market index. These were then merged, aligning prices based on the DowJones' trading dates. After this, missing values created by the merger were replaced using the average of the price of the day above and below each missing price. This was done based on the assumption taken that prices were linear in nature.

Data Normalization- The average for each day was taken and then used as a baseline value to scale all the data points, converting each into percentages of that baseline value. This provided a more even distribution in order to help reduce the variability within the data. A correlation matrix was then generated from the now normalized data, after which the data was then divided into and training set (20 years) and a validation set (4 years).

Apply analysis and algorithms to data- Correlation analysis was conducted on normalized data. After data was divided, it was then fed to the ARIMA algorithm.

Evaluation results

Reporting on result obtained from the analysis

4.2.1 Correlation Implementation overview

1. Overall stock market index correlation matrix was generated from normalized data.

2. Correlation in relation to percentage growth in US GDP rate was generated with data taken on semi-annual US GDP growth.

Create 50 (January and July) excel files and load into R for correlation matrix to be generated.

Visual diagrams were then created to identify trends/patterns that exist.

4.2.2 ARIMA Implementation overview

1. Visualize the time series Plot to identify and understand trends/patterns.

2. Use Dickey-Fuller test, Autocorrelation Function (ACF) plots and Partial Autocorrelation Function (PACF) plots to check stationarity. The Dickey-Fuller test tests the null hypothesis that the data is non-stationary against the alternative it is stationary (Cheung and Lai, 1995). If the results show data to be non-stationary, differencing must be applied to make it stationary. Another Dickey-Fuller test is reapplied and ACF and PACF plots are regenerated to confirm the data is now stationary.

3. Use ACF and PACF to identify viable ARIMA models

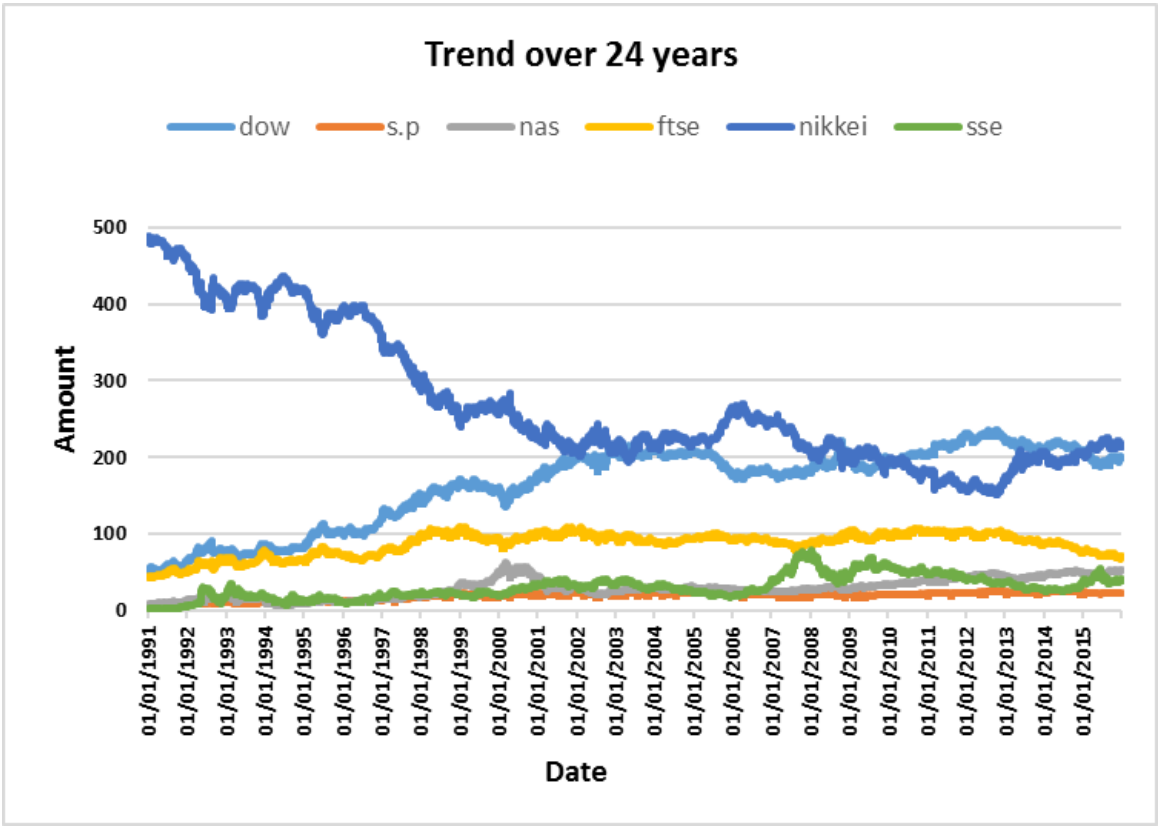
4. Test the different model fits and choose the optimal one based on the AIC values.

5. Test Prediction

6. Evaluation of results will be done by accessing the results from the forecast plots and by comparing forecast figures with validation set figures.

5 Results and Evaluation

5.1 Correlation and volatility



Graph 1 - Visual descriptive of trend

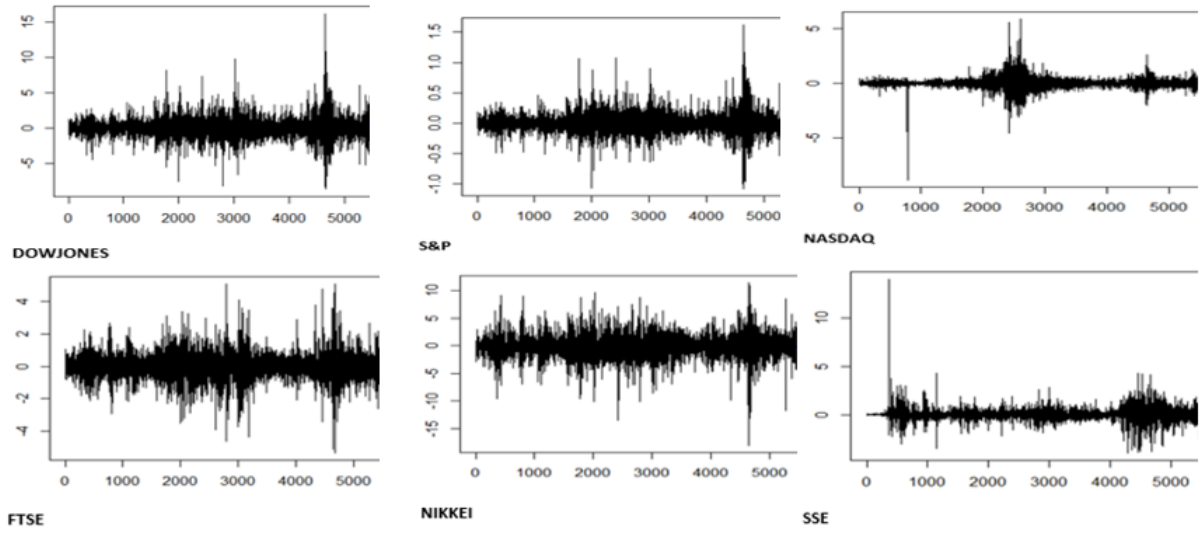
Graph 1 above shows the Nikkei moving downward compared to the other stock market indexes between the period 1991 and 2002. It shows the Nikkei becoming more stabilized and less negative for the remaining time period. This is evident in the correlation matrix table below in Table 1, which records the overall correlation between the Nikkei and the other indexes as being a strong negative correlation ranging between -0.799 and -0.999. The graph shows the DowJones, S and P 500, FTSE and SSE all moving in the same direction, however, by just examining the graph alone, the assumption that the correlation is a strong one cannot be made. This can be proven with the cross-correlation matrix in Table 1, which shows that while there is a positive correlation, the SSE results were moderate while the others had a mixture of moderate and strong.

	dow	s.p	nas	ftse	nikkei	sse
dow	1	0.978681	0.766	0.828	-0.985	0.67
s.p	0.978681	1	0.826	0.841	-0.974	0.608
nas	0.766352	0.826342	1	0.568	-0.804	0.498
ftse	0.827658	0.840752	0.568	1	-0.862	0.566
nikkei	-0.98461	-0.97418	-0.804	-0.862	1	-0.749
sse	0.67015	0.60798	0.498	0.566	-0.749	1

Ranges	
above zero - 0.399	weak
0.4 - 0.699	moderate
0.7 - 0.999	strong

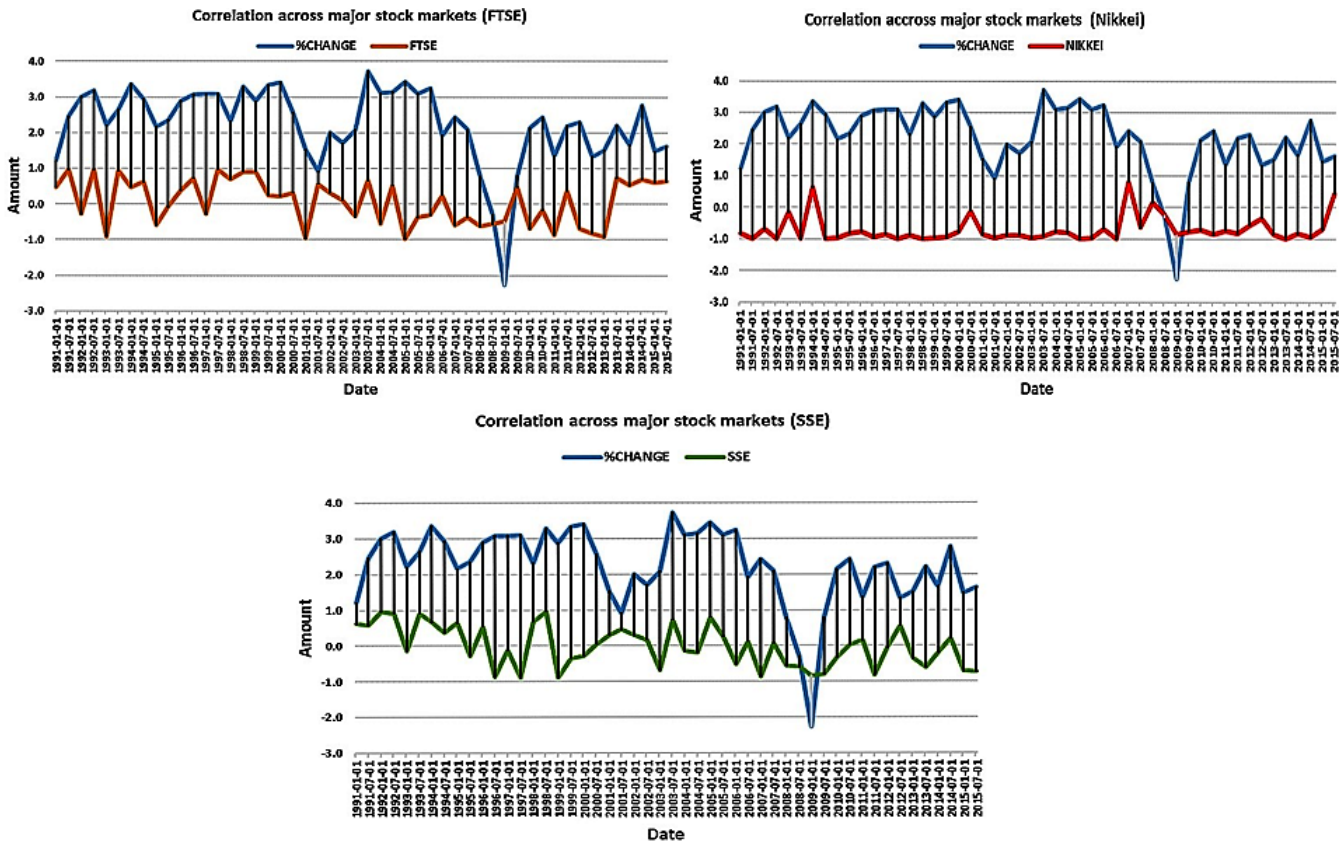
Table 1 - Cross-correlation matrix of stock market closing prices

A Cross-correlation matrix was done on the normalized stock price data which shows the Nikkei having a strong negative relationship with all the 5 other indexes. This can also be seen in the line plot in Graph 1 above, which shows the Nikkei trending downwards while the other are trending upwards throughout most of the time period being observed. However, the Nikkei shows signs of increase since 2013. All other indexes recorded moderate to strong positive correlation with each other. The resulting correlation between the indexes is not surprising as globalization has allowed many companies to have businesses in other countries while trading on their stock markets.



Graph 2 - Graph showing price volatility

Graph 2 above shows all the indexes experiencing high and low volatility at different times. It shows the high volatility are always followed by low volatility, with the DowJones, S and P 500, FTSE and Nikkei seeing relatively high volatility in the earlier period being studied and almost all experiencing the highest volatility peaking between 2008/2009.

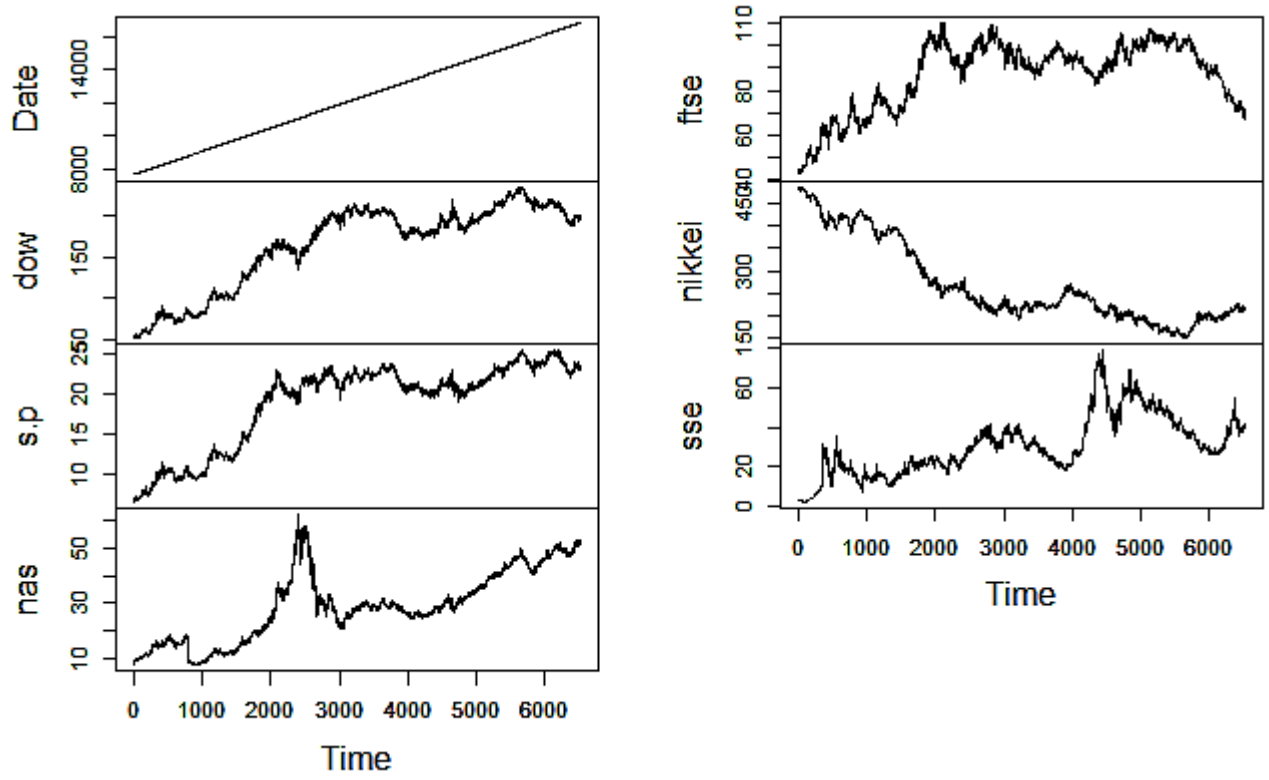


Graph 3- Change in US GDP growth against change in Correlation

One analysis of this research examined the change in the correlation of one US-based stock market index correlation to the other three non-US based indexes. This was done based on the change in percentage US GDP growth over the time period being observed. Fifty correlation matrices were generated based on the release of semi-annual US GDP growth data. The results of which shows the medium term relationship between the DowJones and the other non-US stock markets over the 24-year period. The DowJones has been randomly chosen for this task as the US-based stock market index for making comparisons. Graph 3 illustrates the change in correlation between the DowJones and 3 non-US stock market prices based on the change in percentage GDP growth. For this period, the medium term relationships do not seem to show any discernible trends or patterns to suggest there will be an increasing or decrease in stock market prices in relation to the DowJones based on percentage change in GDP growth or the data's release. Investors may not opt to speculate or take their money from one stock market to another in the long term only based on this factor.

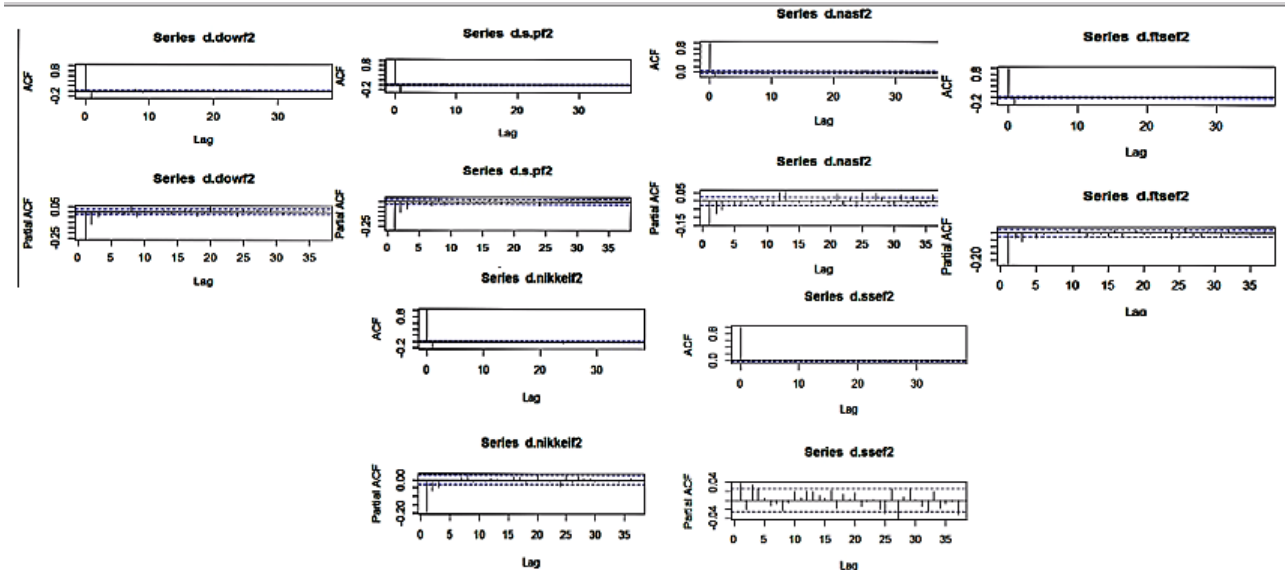
Graph 3 also shows the interconnectedness of the DowJones to the non-US stock market indexes, with there being positive to negative correlation swings over the time period. However, the correlation overall shows the DowJones having a positive correlation with the FTSE (strong) and the SSE (moderate) and strong negative correlation with the Nikkei. While a specific trend in the movement was not identifiable, the results show that the relationship between the stock market indexes is not stationary, but swings both in the same and opposite directions at times. The result shows correlation existing which is also proven in the cross-correlation matrix in Table 1.

5.2 Time series analysis



Graph 4 - Visualization of Time series

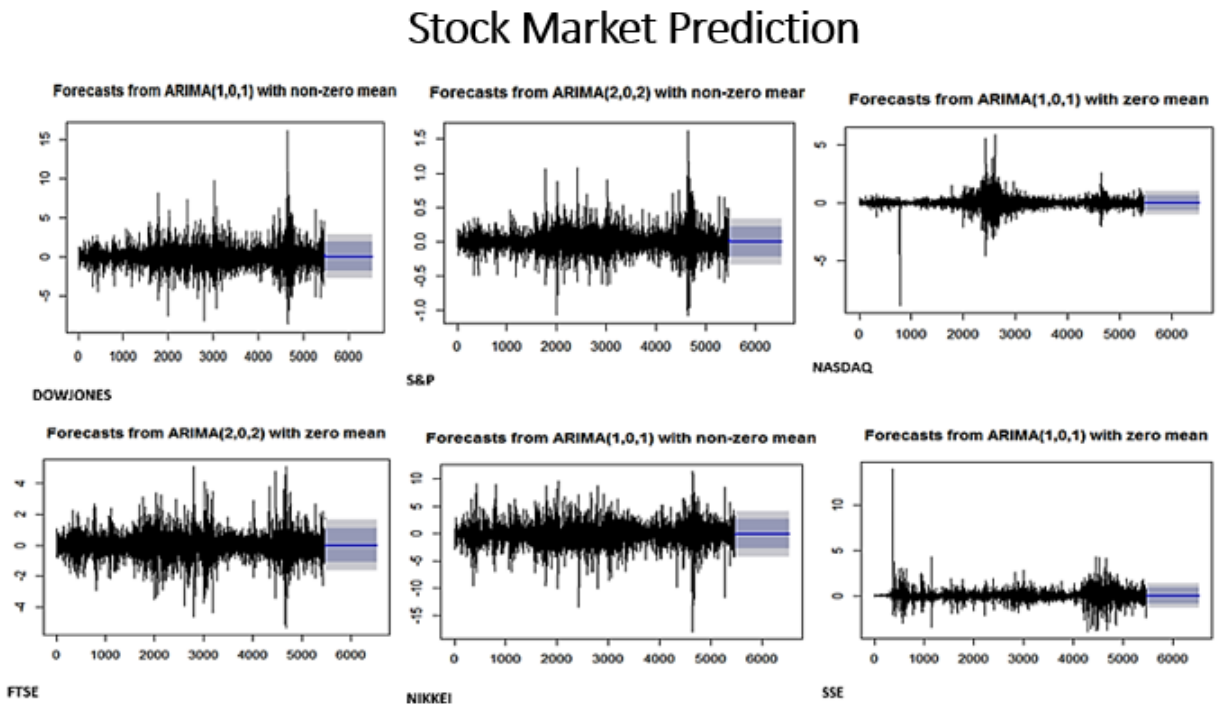
The above plots were used to identify and gain an understanding of the directions of the stock market price trends. These are captured as upward, downward or sideways trends. The Nikkei experienced a relatively strong downward trend until around 2002, then began to show some sign of sideways movement. The other indexes were positively sloped showing an upward trend, however, the Nasdaq 100 also started showing some signs of a sideways trend. They all show minor spikes illustrative of the day to day changes in prices.



Graph 5 - ACF and PACF plots (2nd set of plots)

To ensure that the data was stationary before developing and applying the ARIMA model, ACF and PACF plots were generated. The first set of plots showed that the data was, in fact, non-stationary, which was backed up with results from the Dickey-Fuller test. They visually show how stationary the data is and also any patterns (seasonal or cyclical) that may exist. After data was found to be non-stationary, differencing was applied to the data to make it stationary. After differencing was applied a regeneration of the ACF and PACF plots were done along with another Dickey-Fuller test.

ACF and PACF plots from Graph 5, and the visual time series plots from Graph 4, do not appear to show any structure or discernible patterns existing in the data. Here the ACF shows no existing spikes outside the insignificant zone of the any of the indexes. This may be as a result of the type time series (daily) being analyzed.



Graph 6- ARIMA Prediction Output

Graph 6 above illustrates outputs from the ARIMA prediction which shows a constant line and has also produced predictive values that are also constant for all the indexes. While this was unexpected, it does not infer that the ARIMA time series was unable to properly analyze the data to make proper predictions. One explanatory reason may be the use of time series data on a daily level, spanning over the period of 20 years. It is important when analyzing time series data that the time granularity is taken into consideration.

6 Conclusion and Future Work

Previous studies have shown that more volatility results in higher correlation across markets, which varies over time (Solnik et al., 1996) (Ramchand and Susmel, 1998). This

research can conclude that correlation across major markets does exist based on the report findings. This has also been concluded by the many types of research previously done, most of which have focused on times of major instability/crisis, with data focusing more on trading times being synchronous. They concluded that trading times being non-synchronous or synchronous is the reason for contagion, however, they do not mention smaller scale factors (economic factors) such as changes in a country's GDP, unemployment rate or exchange rates. This research has focused on the end of day closing stock prices, eliminating the issue of non-synchronous or synchronous trading times. This research has also examined the effect of the percentage change in US GDP growth on the change in correlation between the DowJones and three major non-US stock market indexes. This research can conclude that the release of US GDP growth data or its fluctuation does not show any impact or influence whatsoever on the change in the correlation of the DowJones, in relation to the non-US stock market indexes. However, more research can be done to obtain more meaningful results and make a more solid conclusion.

Stock market prices are considered hard to predict and a challenging task (Kim, 2003) (Pai and Lin, 2005) (Kara et al. 2011) (Kim and Shin, 2007). The nature of the financial market is characterized as a complex and non-linear system dynamic (Huang et al., 2005) (Choudhry and Garg, 2008), but this has not deterred this research from its task. By using historic data and data mining techniques, the information gained could be useful in pre-empting stock market price fluctuations. This can lead to greater benefits for the investors such as better decision making and return on investments. While results obtained from the ARIMA algorithm were unexpected, research on previous work conducted and their documented success have shown that ARIMA and time series have been effective in obtaining successful forecasting results. These successes have been recorded with both stand-alone and hybrid algorithms. This research can conclude that based on previous work reviewed, time series is effective in making short-term stock market forecasts. By conducting a future review of the use of the ARIMA algorithm utilized in this research, it is envisaged that a conclusive result would be obtained as to the use of time series for forecasting.

Hybridized time series algorithms were observed to be the most optimal choice, as it relates to data mining applications being applied to stock market predictions. These applications tend to be more successful when combined than when used separately, and also tend to increase in accuracy. However, adequate time for a hybrids development is important. Given the knowledge gained from the research of previous literature, and this research, additional improvements can be made in the field with the combined efforts of a hybrid algorithm, coupled with better market specific domain knowledge. Future works will include other factors (especially economic factors) as independent variables for the prediction of future stock prices. Future work will also seek to implement a hybrid model, utilizing the Kondratiev theory of Long Waves. This theory will be useful for identifying long-term stock market price trends as historic stock market data going back more than 100 years will be utilized. This, if successful, could lead to even greater benefits for the investors and also the field of financial data science.

Acknowledgements

I would like to extend my heartfelt gratitude and thanks to my project supervisor Mr. Vikas Sahni for his continued guidance, patience and encouragement throughout this

process. I was able to benefit immensely from your knowledge and value input on the subject matter that was researched. I would also like to thank my colleague Ken Lawlor for his continued support and input.

7 References

Abraham, A., Nath, B. and Mahanti, P.K., (2001) Hybrid intelligent systems for stock market analysis. In Computational science-ICCS 2001 (pp. 337-345). Springer Berlin Heidelberg.

Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015). Time series analysis: forecasting and control. John Wiley Sons.

Choudhry, R., and Garg, K. (2008) A hybrid machine learning system for stock market forecasting. World Academy of Science, Engineering, and Technology, 39(3): pp.315-318.

Cheung, Y.W. and Lai, K.S., (1995) Lag order and critical values of the augmented DickeyFuller test. Journal of Business Economic Statistics,13(3): pp.277-280.

Enke, D. and Thawornwong, S., (2005) The use of data mining and neural networks for forecasting stock market returns. Expert Systems with applications, 29(4): pp.927-940.

Hamao, Y., Masulis, R.W. and Ng, V., (1990) Correlations in price changes and volatility across international stock markets. Review of Financial studies, 3(2): pp.281-307.

Hassan, M.R., and Nath, B., (2005) September. Stock market forecasting using hidden Markov model: a new approach. In Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on (pp. 192-196). IEEE.

Hassan, M.R., Nath, B. and Kirley, M., (2007) A fusion model of HMM, ANN and GA for stock market forecasting. Expert Systems with Applications,33(1): pp.171-180.

Huang, W., Nakamori, Y. and Wang, S.Y., (2005) Forecasting stock market movement direction with support vector machine. Computers Operations Research, 32(10): pp.2513-2522.

Kara, Y., Boyacioglu, M.A. and Baykan, .K., (2011) Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert systems with Applications, 38(5): pp.5311-5319.

Kim, K.J., (2006) Artificial neural networks with evolutionary instance selection for financial forecasting. Expert Systems with Applications, 30(3): pp.519-526.

Kim, K.J., (2003) Financial time series forecasting using support vector machines. Neurocomputing, 55(1): pp.307-319.

Kim, H.J. and Shin, K.S., (2007) A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. Applied Soft Computing, 7(2): pp.569-576.

Ledoit, O., Santa-Clara, P. and Wolf, M., (2003) Flexible multivariate GARCH modelling with an application to international stock markets. Review of Economics and Statistics, 85(3): pp.735-747.Engle, R., (2001) GARCH 101: The use of ARCH/GARCH models in applied econometrics. The Journal of Economic Perspectives, 15(4), pp.157-168.

Martens, M. and Poon, S.H., (2001) Returns synchronization and daily correlation dynamics between international stock markets. Journal of Banking Finance, 25(10): pp.1805-1827.

Pai, P.F. and Lin, C.S., (2005) A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 33(6): pp.497-505.

Ramchand, L. and Susmel, R., (1998) Volatility and cross correlation across major stock markets. Journal of Empirical Finance, 5: pp.397-416.

Shen, S., Jiang, H. and Zhang, T., (2012) Stock market forecasting using machine

learning algorithms. Computer Science and Engineering, SSN College of Engineering Chennai, India.

Solnik, B., Boucrelle, C. and Le Fur, Y., (1996) International market correlation and volatility. Financial analysts journal, 52(5), pp.17-34.

Tanning, T., Saat, M. and Tanning, L., (2013) Kondratiev wave: an overview of world economic cycles. Global Business and Economics Research Journal,2(2): pp.1-11.

Yang, S.Y. and Doong, S.C., (2004) Price and Volatility Spillovers between Stock Prices and Exchange Rates: Empirical Evidence from the G-7 Countries. International Journal of Business, 3(2): pp.139-153.

Zhang, G.P., (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50: pp.159-175.

Hyndman, R.J. and Athanasopoulos, G. (2013) Forecasting: principles and practice. OTexts: Melbourne, Australia. Section 8/7.Available from: http://otexts.org/fpp/8/37. [Accessed August 1st 2016].

Introduction to ARIMA: non-seasonal models. ARIMA models for time series forecasting, Available from: http://people.duke.edu/ rnau/411arim.htm[Accessed July 20th 2016].

Nasdaq 100 Index. Investopedia. Available from: http://www.investopedia.com/terms/n/nasdaq100 [Accessed July 1st 2016].

Sideways Trend. Investopedia. Available from: http://www.investopedia.com/terms/s/sidewaystren [Accessed July 1st 2016].

Srivastava, T. (2015) A complete tutorial on time series modelling in R. Analytics Vidhya, 16th December. Available from: https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/. [Accessed July 20th 2016].

Standard Poors 500 Index SP 500.Investopedia. Available from: http://www.investopedia.com/term [Accessed July 1st 2016].

Upadhyay, R. (2015) Step-by-step Graphic guide to forecasting through ARIMA modelling in R-Manufacturing case study example(Part4). YOU CANanalytics, 28th June. Available from: http://ucanalytics.com/blogs/step-by-step-graphic-guide-to-forecasting-through-arima-modeling-in-r-manufacturing-case-study-example/. [Accessed July 20th 2016].

Replacing NA with previous and next rows mean in R. Stackoverflow. Available from: http://stackoverflow.com/questions/22916525/replace-na-with-previous-and-next-rows-mean-in-r[Accessed July 20th 2016].

Reading and Writing .CSV Files in R-Studio. Reed College, Instructional Technology Services. Available from: http://www.reed.edu/data-at-reed/resources /R/reading$_a nd_w riting.html[Ac$

A First Appendix Section

A.1 Definitions, Acronyms, and Abbreviations

SP 500 - Standard and Poors 500 Index
DJIA - Dow Jones Industrial Average
NIKKEI - Nikkei 225 indexes
SSE - Shanghai Stock Exchange
FTSE 100 - Financial Times Stock Exchange 100 index
NDX - Nasdaq 100
GDP - Gross Domestic Product
HMM - Hidden Markov Model
ANN - Artificial Neural Networks
GA - Genetic Algorithm
SVM - Support Vector Machine
ARIMA- Autoregressive integrated moving Average
SWARCH - Switching Autoregressive Conditional Heteroscedasticity
GARCH - Generalized Autoregressive Conditional Heteroscedasticity

A.2 Change in Correlation vs change in US GDP growth

Results from correlation matrices generated based on the effect percentage change in US GDP growth has had on the change in correlation between the DowJones and non-US stock market indexes.

DATE	% CHANGE IN GDP GROWTH	FTSE	NIKKEI	SSE
1991-01-01	1.2	0.46861	-0.83677	0.62731
1991-07-01	2.5	0.96555	-0.97902	0.58811
1992-01-01	3.0	-0.28775	-0.68451	0.96032
1992-07-01	3.2	0.95978	-0.99082	0.91051
1993-01-01	2.2	-0.90923	-0.15766	-0.14514
1993-07-01	2.6	0.94053	-0.99587	0.9143
1994-01-01	3.4	0.47301	0.63199	0.66872
1994-07-01	2.9	0.61613	-0.99362	0.37995
1995-01-01	2.2	-0.60244	-0.96167	0.64123
1995-07-01	2.3	-0.08528	-0.81511	-0.27157
1996-01-01	2.9	0.3823	-0.75954	0.53938
1996-07-01	3.1	0.72445	-0.94707	-0.87478
1997-01-01	3.1	-0.28872	-0.84089	-0.12341
1997-07-01	3.1	0.97437	-0.97908	-0.88077
1998-01-01	2.3	0.69575	-0.86599	0.68102
1998-07-01	3.3	0.90776	-0.99546	0.96052
1999-01-01	2.9	0.89933	-0.95144	-0.8808
1999-07-01	3.3	0.24584	-0.93927	-0.35071
2000-01-01	3.4	0.23041	-0.77672	-0.28185
2000-07-01	2.6	0.31611	-0.10493	0.05466
2001-01-01	1.5	-0.96416	-0.86336	0.30063
2001-07-01	0.9	0.56887	-0.95607	0.4552
2002-01-01	2.0	0.32071	-0.88984	0.29631
2002-07-01	1.7	0.10926	-0.86962	0.16414
2003-01-01	2.1	-0.35776	-0.9734	-0.68944
2003-07-01	3.7	0.66566	-0.90948	0.72546
2004-01-01	3.1	-0.54876	-0.77692	-0.15294
2004-07-01	3.1	0.51943	-0.80785	-0.17997

DATE	% CHANGE IN GDP GROWTH	FTSE	NIKKEI	SSE
2005-01-01	3.4	-0.97426	-0.99547	0.79965
2005-07-01	3.1	-0.3761	-0.97243	0.26075
2006-01-01	3.2	-0.3105	-0.67761	-0.50681
2006-07-01	1.9	0.24615	-0.99465	0.12502
2007-01-01	2.4	-0.59641	0.81938	-0.85995
2007-07-01	2.1	-0.36418	-0.63687	0.0745
2008-01-01	0.8	-0.62614	0.1623	-0.5701
2008-07-01	-0.3	-0.55495	-0.23605	-0.58885
2009-01-01	-2.3	-0.48297	-0.86356	-0.8344
2009-07-01	0.8	0.47304	-0.75324	-0.78903
2010-01-01	2.1	-0.69252	-0.71068	-0.31848
2010-07-01	2.4	-0.17553	-0.85814	0.02561
2011-01-01	1.4	-0.86474	-0.73884	0.15313
2011-07-01	2.2	0.37638	-0.8125	-0.81972
2012-01-01	2.3	-0.68342	-0.56094	-0.0559
2012-07-01	1.3	-0.8131	-0.33258	0.56118
2013-01-01	1.5	-0.91689	-0.84012	-0.35841
2013-07-01	2.2	0.72882	-0.98716	-0.60342
2014-01-01	1.7	0.52922	-0.82544	-0.20936
2014-07-01	2.8	0.702	-0.95038	0.20538
2015-01-01	1.5	0.61442	-0.68774	-0.70426
2015-07-01	1.6	0.647	0.44861	-0.73729

...