INDEX RETURN PREDICTION

Table of Contents

[1 CHAPTER ONE: INTRODUCTION 3](#_Toc130304399)

[1.1 Background Information 3](#_Toc130304400)

[1.2 Objectives of the Study 4](#_Toc130304401)

[1.2.1 General Objective 4](#_Toc130304402)

[1.2.2 Specific Objectives 4](#_Toc130304403)

[2 CHAPTER TWO: LITERATURE REVIEW 5](#_Toc130304404)

[2.1 Introduction 5](#_Toc130304405)

[2.2 Empirical Review 5](#_Toc130304406)

[3 CHAPTER 3: METHODOLOGIES 8](#_Toc130304407)

[3.1 Introduction 8](#_Toc130304408)

[3.2 Data Collection 8](#_Toc130304409)

[3.3 Data Analysis 8](#_Toc130304410)

[3.3.1 ARIMA Model 9](#_Toc130304411)

[4 CHAPTER FOUR: DATA ANALYSIS 10](#_Toc130304412)

[4.1 Introduction 10](#_Toc130304413)

[4.2 Descriptive Statistics 10](#_Toc130304414)

[4.3 Data Visualization 10](#_Toc130304415)

[4.4 Autocorrelation Function and Partial Autocorrelation Function 13](#_Toc130304416)

[4.5 Model Fitting 14](#_Toc130304417)

[4.6 Monthly Index Return 15](#_Toc130304418)

[5 CHAPTER 5 : DISCUSIONS AND CONCLUSIONS 18](#_Toc130304419)

[5.1 Discussions 18](#_Toc130304420)

[5.2 Conclusions 18](#_Toc130304421)

# CHAPTER ONE: INTRODUCTION

## 1.1 Background Information

In finance, specifically in the stock market, return is defined as an alteration in the price of an asset, investment, or project over time. Returns are usually represented in terms of price change or percentage change. A positive return indicates a profit, while a negative return indicates a loss. Thus, an index is simply an asset that tracks the whole market of stocks or a specific part of the market. The index returns usually reflect the overall performance of the stocks or securities in the particular index. Most investors and people who have bought stocks or shares use these indices as a benchmark to compare the overall market performance and their investments. The index return is calculated from ratios of prices of different stocks or shares of the companies in the index, and this gives the price indices. From the price indices, the returns are calculated, and thus, the index returns. On the other hand, the returns generated by a specific market index that tracks the performance of a set or portfolio of shares, equities, or stocks are known as equity index returns.

There are many stock market indices developed to aid in tracking the prices of stocks. The Dow Jones Industrial Average (DJIA) is one of the oldest stock market indexes. This index spearheaded the use of index returns in the stock market. DJIA was created on May 26, 1896, by Charles H. Dow, founder of Dow Jones and Co. It consisted of 12 companies, which were from the industrial sector. These industries were strategically chosen to represent significant areas of concern in the United States economy. The number of stocks in the index rose gradually; by 1972, the number of stocks in the index had crossed the 1000 mark. Initially, DJIA was created to track and gauge the well-being of the industrial sector in the US. So essentially, DJIA gave the foundation for the formation and calculation of the stock market indices and, thereby, the index returns. Today, many indices are created to track popular, if not all, stock markets. An example is the S&P 500, an equity index comprising the top 500 largest companies in the United States, which are traded on the New York Stock Exchange and others.

## 1.2 Objectives of the Study

The objectives of this study are as follows;

### 1.2.1 General Objective

The general Objective of this study is to create and develop a model for predicting future equity index returns.

### 1.2.2 Specific Objectives

The specific objectives of this study are as follows;

1. To determine whether the equity index returns predictability is elusive or not
2. To identify the best time series model for predicting equity index returns
3. To determine whether the predictability of equity index return depends on the frequency

# CHAPTER TWO: LITERATURE REVIEW

## 2.1 Introduction

This chapter discusses the previously done literature on predicting stock indices as well as equity index returns. This chapter also criticizes various literature materials.

## 2.2 Empirical Review

In 2003, Burton Malkiel in his article “The Efficient Market Hypothesis and Its Critics,” posited that stock prices follow a random walk, a concept central to the Efficient Market Hypothesis (EMH). Malkiel argues that a random walk is a mathematical concept in time series used to describe values determined by chance and random fluctuations and is the central assumption behind the EMH (Yalvaç 2011). According to the EMH, stock or asset prices reflect all publicly available information.

He concludes that predicting stock indices is essentially impossible and that the stock markets are efficient and less predictable. While the study does not aim to predict stock indices, it examines the predictability of their returns despite their unpredictability. In other words, the question is whether stock/equity index returns are predictable, given that the stock indices are unpredictable.

Moreover, the study also highlights the difference between an equity index and equity index returns. While the former is a static measure of a portfolio’s overall performance, the latter is a dynamic measure of the index’s value fluctuations over time. Therefore, understanding the difference between equity index and equity index returns is crucial when assessing the performance of a portfolio and when determining whether the returns of a particular stock or asset are predictable.

In 2000, Leung conducted a study examining the direction and signs of the stock index movement alongside other researchers (Leung, 2000). The study employed machine learning algorithms such as linear discriminant analysis, logit, and neural networks and compared them to level estimation models like exponential smoothing and the Kalman filter. The study concluded that classification algorithms generally outperformed level estimation models when predicting the direction of stock indices.

Similarly, in 2006, Manish Kumar conducted a study utilizing machine learning algorithms like Random Forest and Support Vector Machines (SVM) to predict the direction of stock indices from the S&P CNX NIFTY, which is a benchmark index that tracks over 40 of the largest equities in India (Kumar, 2006). The empirical study showed that support vector machines outperformed other classification algorithms when predicting the direction of stock indices. These two sources suggest that, aside from the commonly used time series forecasting of stock indices, there are other ways to predict future equity indices. However, this study focuses on the time series forecasting aspect.

In the year 2014, Dipankar Banerjee conducted a study aimed at forecasting the Indian stock market through the use of the Auto Regressive Integrated Moving Average (ARIMA) Model in time series (Banerjee, 2014). His findings revealed that the ARIMA model was a highly effective approach to forecasting the BSE index. Additionally, the same year, Manish Kumar and M. Thenmozhi conducted a research project to develop and identify a hybrid model that was most effective in predicting stock index returns (Kumar et al., 2014). To achieve this, they developed three different hybrid models by combining the linear ARIMA model with non-linear models such as Support Vector Machines, Neural Networks, and Random Forests. Upon analysis, the results indicated that the ARIMA-SVM model was the most efficient in forecasting the stock index returns.

# CHAPTER 3: METHODOLOGIES

## 3.1 Introduction

This chapter comprises a detailed analysis of the statistical models and methods used in the study. It also includes a detailed explanation of the process of data collection.

## 3.2 Data Collection

The data used in this study is the stock indices of the Dow Jones Industrial Average (DJIA) stock market index. The data is collected secondarily from the R-package known as **tsfe**. The data is named **indices and** contains data for daily stock price indices and significant currency exchange rates.

For the Dow Jones Industrial Average (DJIA) indices, the stock prices from 1960-01-29 to 2020-01-24 are recorded daily. This is the data set used in this study.

## 3.3 Data Analysis

Preliminary data preparation will be done. Also, the characteristics of the variable(s) in the data set will be examined to check whether there are any missing values and analyze the number of observations. This examination will also aid in handling the outliers if present.

After the data has been preprocessed, the data will be converted to a time series object through the *ts* function in R. Then, the data visualization will determine whether there is a trend and seasonality in the time series.

After preprocessing the data to meet the minimum requirements for modeling and predicting essentials such as stationarity, the following time series models will be used to try and see how well the time series fits the model. The best model will be used to predict the stock indices. The model is discussed below.

### 3.3.1 ARIMA Model

The Autoregressive Integrated Moving Average is a model used in time series to forecast the future values of a time series based on the time series’ past values and errors in past predictions. The model has three parameters, p. D, and q. P represents the autoregressive part of the model. This parameter enables the modeling of the dependence of the current value on past matters. d, conversely, represents the integrated part of the model. q Represents the moving average part of the model. These parameters represent the different ARIMA models. Hence, other ARIMA models based on these parameters will be tested, and the best will be selected for forecasting the stock price indices.

The ARIMA model will also be estimated at different frequencies of the data set: daily, weekly, monthly, and yearly. This will aid in determining whether the frequency of the equity index affects its predictability.

# CHAPTER FOUR: DATA ANALYSIS

## 4.1 Introduction

This chapter entails a detailed data analysis and statistical data set modeling.

## 4.2 Descriptive Statistics

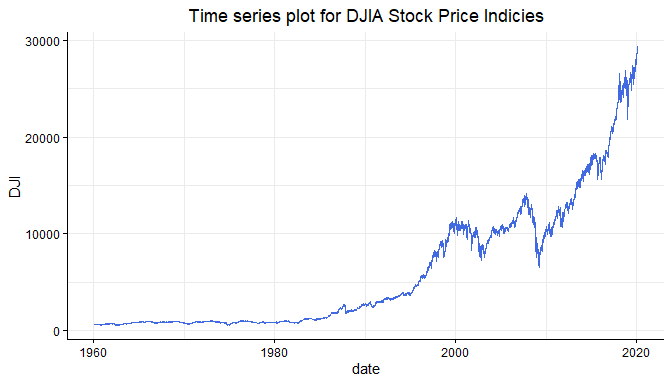
The descriptive statistics of the data were as follows. This study only considered the DOW JONES INDUSTRIALS - PRICE INDEX` variable in the data set as the study’s data set.

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 535.8 891.2 2635.6 6122.4 10522.1 29348.1

The minimum index was 535.8, while the maximum was 29348.1. The mean index of all the prices across the years was 6122.4. The variance of the indices was r, var (data$DJI).

## 4.3 Data Visualization

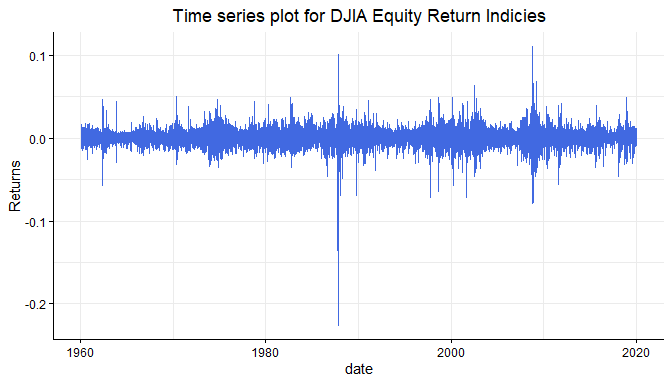
The data was converted into time series, and the time series plot obtained was as follows;



Daily Stock Price Indices

From the plot above, there is an evident upward trend of the daily stock price indices, and also, there is seasonality evidenced by regular troughs across the years. The detrending and deseasonalizing can be done using first and second-order differencing, respectively. However, this study was primarily interested in the index returns. Hence the index returns were calculated using the following formula.

After using the formula, the daily equity index returns were obtained, and the plot is as follows;



Daily Equity Return Indices

The plot above suggests no seasonality and trend in the index returns. The above seems like a noise that remains after decomposing or differencing. Thus since there was no evidence of seasonality or movement, the next step was to check whether the series was stationary. This was done using the Augmented Dickey-Fuller (ADF) test. The null hypothesis of the test is that;

The Series is a random walk and hence not stationary

: The series is stationary.

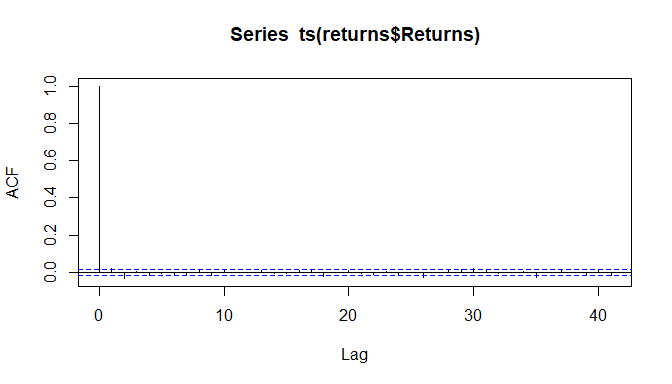
The null hypothesis would be rejected if the p-value was less than. The test results were as follows;

Augmented Dickey-Fuller Test  
  
data: returns$Returns  
Dickey-Fuller = -26.009, Lag order = 25, p-value = 0.01  
alternative hypothesis: stationary

Since the p-value was less than 0.05, the null hypothesis was rejected, and hence it was concluded that the series was stationary.

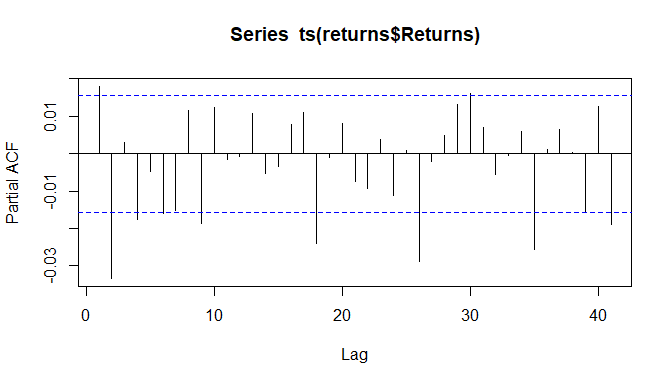
## 4.4 Autocorrelation Function and Partial Autocorrelation Function

After checking for this requirement, the autocorrelation and partial autocorrelation function were plotted to aid in identifying the best model for the series. The Autocorrelation function plot of the series was as shown below;



ACF Plot

The ACF plot shows a very significant spike at lag one since it is the correlation between the observation and itself. Also, there is a significant spike at lag 2. The PACF plot is as shown below;



PACF plot

Apart from the first significant spike, there are other significant spikes at lag 2, 4, then at lag six, and so on.

From both the ACF and PACF, the significant spike at lag 2 suggests an Autoregressive model of order 2 (AR (2)), an AR (4), or an AR (6) would be appropriate an appropriate model for this time series.

## 4.5 Model Fitting

Before fitting the various ARIMA models, the last ten values of the series were removed. These were aimed at leaving them for the model to predict and test the Accuracy.

The **auto. Arima ()** function in R was used to select the best model for the data. This function runs a series of step-wise algorithms to search for the best model for the time series. The results were as follows;

Series: series\_train   
ARIMA (2,0,0) with non-zero mean   
  
Coefficients:  
 ar1 ar2 mean  
 0.0186 -0.0334 3e-04  
s.e. 0.0080 0.0080 1e-04  
  
sigma^2 = 9.454e-05: log likelihood = 50272.96  
AIC=-100537.9 AICc=-100537.9 BIC=-100507.3

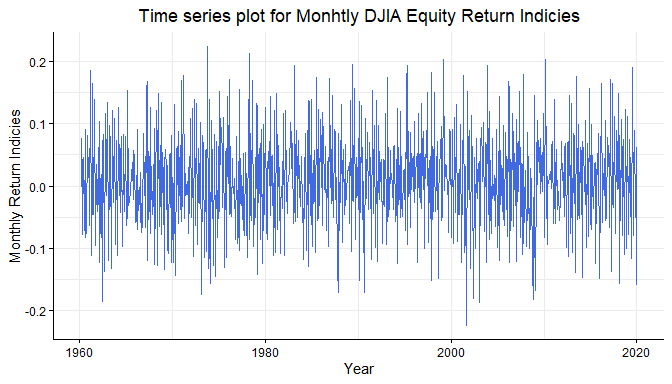
Thus the algorithm selected the **ARIMA (2,0,0)** as the best model for the time series data. The Forecasting for the ten values was done, and below are the accuracy results

ME RMSE MAE MPE MAPE MASE  
The training set 2.628525e-08 0.009722402 0.006568813 NaN Inf 0.7015741  
Test set 3.058948e-04 0.004100668 0.003049770 -Inf Inf 0.3257269  
 ACF1  
Training set 9.436269e-05  
Test set NA

The accuracy test resulted in shallow values for the Mean Error (ME), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). This indicates that the model was a good fit for the data. However, these results were to be used to compare the model obtained using daily and monthly frequency.

## 4.6 Monthly Index Return

The frequency of the time series was changed from daily to monthly to check if the frequency of the time series affects predictability. The time series plot for the monthly equity index is shown below;



Monthly Equity Return Indices

The plot above shows that the monthly equity return indices are random and fluctuate frequently. The test for stationary was as follows;

Augmented Dickey-Fuller Test  
  
data: monthlyreturns$`Monthly Return Indices  
Dickey-Fuller = -8.5697, Lag order = 8, p-value = 0.01  
alternative hypothesis: stationary

The ADF test indicated that the series was stationary. Then the auto. Rima function was used to obtain the best model for the data. The results were as follows;

Series: series\_train2   
ARIMA(2,0,2) with non-zero mean   
  
Coefficients:  
 ar1 ar2 ma1 ma2 mean  
 -1.0042 -0.6195 0.4009 0.1848 0.0088  
s.e. 0.0635 0.0513 0.0731 0.0811 0.0015  
  
sigma^2 = 0.004627: log likelihood = 914.56  
AIC=-1817.11 AICc=-1816.99 BIC=-1789.64

The results indicated that the best model for the time series was an **ARIMA (2,0,2)**. This model was used to forecast ten future values of the data, and the Accuracy obtained was as follows;

ME RMSE MAE MPE MAPE MASE  
the training set 2.936679e-05 0.06778572 0.05338412 92.07662 165.38601 0.4330150  
Test set -1.347478e-02 0.08309789 0.06523873 72.78082 75.18131 0.5291713  
 ACF1  
Training set -0.003147707  
Test set NA

Also, the ME, RMSE, and MAE were relatively low in this model, indicating a good model for the time series.

# CHAPTER 5: DISCUSSIONS AND CONCLUSIONS

## 5.1 Discussions

While Malkiel, in his study, concluded that the stock indices are less unpredictable, the data analysis in this study has indicated that the index return indices are predictable compared to the static indices. The data analysis shows that the ARIMA models are the best fit for predicting or forecasting the equity return index. The study has also found that the equity return index is not entirely elusive in predictability and hence can be expected. There was a difference in the best models between the daily index return and the monthly index return. The daily index return suggested that the best model for the time series is the ARIMA (2,0,0). In contrast, the monthly index returns have indicated that the best model for the time series is the ARIMA (2, 0,2).

The ARIMA(2,0,2) had larger values for ME, RMSE, and MAE as compared to the importance of ARIMA(2,0,0), indicating that the ARIMA(2,0,0) is a better model.

## 5.2 Conclusions

Therefore, this study concluded that the predictability of equity index returns is not elusive or predictable. It shows that the predictability of the equity index returns depends on the frequency and tends to change as the frequency changes. Finally, this study concluded that the ARIMA (2,0,0) is the best model for predicting equity index returns.

# REFERENCES

Banerjee, D. (2014, January). Forecasting of the Indian stock market using the time-series ARIMA model. In *2014 2nd international conference on Business and information management (ICBIM)* (pp. 131-135). IEEE.

Kumar, M., & Thenmozhi, M. (2014). Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. *International Journal of Banking, Accounting, and Finance*, *5*(3), 284-308.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. Journal of economic perspectives, 17(1), 59-82.

Leung, M. T., Daouk, H., & Chen, A. S. (2000). Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of Forecasting*, *16*(2), 173-190.

Yalvaç, E. F. (2011). *Event Studies and the Efficient Market Hypothesis-Evidence from Turkish Merger Annocements* (Doctoral dissertation, Marmara Universities (Turkey)).