MODELLING STOCK MARKET EXCHANGE BY AUTOREGRESSIVE INTEGRATED MOVING AVERAGE, MULTIPLE LINEAR REGRESSION AND NEURAL NETWORK

Mohamad Firdaus\textsuperscript{a}, Nur Arina Bazilah Kamisan\textsuperscript{a*}, Nur Arina Bazilah Aziz\textsuperscript{a}, Chan Weng Howe\textsuperscript{b}

\textsuperscript{a}Department of Science Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
\textsuperscript{b}School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Abstract

Stocks, sometimes known as equities, are fractional ownership shares in a firm, and the stock market is a venue where investors may purchase and sell these investible assets. Because it allows enterprises to quickly get funds from the public, a well-functioning stock market is critical to economic progress. The purpose of this study is to model Bursa Malaysia using autoregressive integrated moving average (ARIMA), multiple linear regression (MLR), and neural network (NN) model. To compare the modelling accuracy of these models for intraday trading, root mean square error (RMSE) and mean absolute percentage error (MAPE) as well as graphical plot will be used. From the results obtained from these three methods, the NN model provides the best trade signal.

Keywords: ARIMA, MLR, multilayer perceptron, modelling, neural network, stock market

Abstrak

Saham, kadangkala dikenali sebagai ekuiti, ialah saham pemilikan pecahan dalam firma, dan pasaran saham ialah tempat pelabur boleh membeli dan menjual aset boleh labur ini. Kerana ia membolehkan perusahaan mendapatkan dana daripada orang ramai dengan cepat, pasaran saham yang berfungsi dengan baik adalah penting untuk kemajuan ekonomi. Tujuan kajian ini adalah untuk meramalkan Bursa Malaysia menggunakan purata pergerakan bersepadu autoregresif (ARIMA), regresi linear berganda (MLR), dan model rangkaian saraf (NN). Untuk membandingkan ketepatan pemodelan model ini untuk dagangan intrahari, ralat purata kuasa dua (RMSE) dan ralat peratusan mutlak (MAPE) serta plot grafik akan digunakan. Keputusan yang diperolehi daripada kajian ini, model NN memberikan isyarat dagangan terbaik.

Kata kunci: ARIMA, MLR, perseptron berbilang lapisan, pemodelan, model rangkaian saraf, pasaran saham

© 2022 Penerbit UTM Press. All rights reserved
1.0 INTRODUCTION

The stock market is a collective term that refer to public market for the purpose of issuing, selling and buying stocks that trade on a stock exchange or over the counter. Stocks, sometimes known as equities, represent fractional ownership in a corporation, and the stock market is a marketplace for investors to buy and sell such investible assets. A well-functioning stock market is crucial to economic development because it allow businesses to swiftly acquire fund from the public. In additionally, a well-functioning stock market, according to one dictionary definition, is “a gathering of people for purpose of trading” [1]. Two or more persons meet, either in person or via computer link, and attempt to strike a bargain. It is possible for items to be present at market or it is possible for them not to be. A housing market is different from a food shop in several respects. Apples at the grocery store vary in terms of size, price and quality, yet when compared to houses, they are nearly identical. House differ far more than apples in terms of age and size. While these elements significantly alter the structure of the markets for apples and houses, the true reality is that these markets remain nothing more than gatherings of the people for the purpose of trade. The same is true of stock market, these are financial networks that enable individuals to trade stock.

Bursa Malaysia is a stock market holding company founded 1976 and listed on the Bursa Malaysia in 2005. It assists over 900 companies raising capital across 50 economic sectors, including SMEs (Small and Medium Sized Enterprises) [2]. The main market is reserved for established large cap companies. To ensure that a fair and orderly market is maintained for the securities and derivatives that are trade through its trading facilities. Then, Bursa Malaysia is responsible for ensuring orderly dealings with the securities deposited with exchange.

Generally, the country’s stock exchange is Bursa Malaysia and serves as a venue for corporations, organizations and governments to sell assets to the public contains 30 major companies on the main board [3]. Bursa Malaysia offers a diverse range of products, including bonds, clearing, settlement, equities, trading, derivatives and depository services. Investors often retain securities for a longer period of the time and trade stocks, exchange-traded fund and other securities.

Normally, traders will purchase and sell stocks and decide to hold it for a shorter amount of time and encounter in a bigger transactions. Therefore, big impact on market impulsiveness causes traders wrong and to lose confidence in their ability to make a choice [4]. Hence, using trading strategy indicator to assist in making best decisions and minimize market impact. This contributes to the expansion of the highest possible accuracy rate for investors. It is an extremely important intraday trading tool that assist in entering and exiting trading positions as well as identifying probable entry and exit locations, regardless of its strength and benefits, it cannot be utilised in isolation. As result, technical analysis and specialists emphasise the importance of employing the best indicator as trustworthy conformation indications [5].

The ARIMA model, which assumes that the time series are linear and follows a well know distribution, such as Gaussian. These models are constrained by the assumption of a linear time series, which can produce inadequate results in a wide variety of practical scenarios [6]. Then, the ARIMA model predicts the dependent variable only on the basis of its lagged or previous values. [7]. Moreover, ARIMA has been extensively utilised to create a variety of forecasts. When compared to Simple Exponential Smoothing (SES) and Two Parameters Holt Exponential Smoothing (HES), ARIMA forecasting has a higher prediction accuracy [8]. Furthermore, the Auto ARIMA approach conducts a grid search for all possible parameter combinations at a specified interval and chooses the model with the lowest AIC (Akaike Information Criterion) value. When applying Auto ARIMA to their model, a seasonal decomposition was performed to determine whether the dataset exhibited seasonality. After decomposition, it is discovered that the graph exhibits clear seasonal patterns. As a result, they set the seasonal parameter in Auto ARIMA to true, so establishing a seasonal cycle of 12 months [8].

Multiple Linear Regression (MLR) is a mathematical model that is used to describe the behaviour of a random variable of interest. This variable could be the financial market’s stock price, the growth of a biological species, or the possibility of detecting a gravitational wave. The dependent variable is referred to as the dependent variable [9]. Moreover, the MLR is a linear regression model that takes use of the causative relationship between target and input variables by estimating their corresponding regression coefficients using the data’s extreme deviations. Another significant advantage of Regression is a linear regression is its ability to decrease changes caused by unexplained “noise” [10].

In addition, MLR models are easily updated as newer chunks of input and target data arrive, and therefore their transparency in terms of the relationship between inputs and outputs, as well as their flexibility in implementation, have contributed to their increased acceptance in real-world application [10]. Hassan and Mohd-Saleh (2010) used MLR to inspect the value relevance of financial instrument disclosure in Malaysia because the model could explain how to estimate firm worth using accounting data rather than the market prices and it allow researchers to describe tests for perceived share mispricing and bridging the gap between financial analysis and valuation [11].

Artificial Neural Network (ANN) models include a multi-layered neural network architecture that enables them to efficiently collect price movement patterns using deep learning techniques. Then, using deep learning techniques on a training dataset, our ANN forecasts the following period’s price direction (up or down) based on the input data. As a result, we
use simple investment tactics to determine when to purchase or sell when the price is expected to increase or decrease [12].

Besides that, price forecasting and financial trading Neural Network model, where empirical examination of futures trading confirmed the suggested model's practice in terms of risked return. Thus, Neural Network models with their sequential learning capabilities are most frequently utilized to forecast financial time series. However, previous research has not adequately addressed the features of high noise and substantial volatility in financial time series, which jeopardize the viability of prediction models [13]. Moreover, to create a neural network that performs well in-sample. After identifying such a network, the parameters that support it are effectively frozen for out-of-sample testing. At that moment, the degree to which the ANN has been curve-fit will become apparent [14].

In this paper, the stock market data from Bursa Malaysia will be fitted with ARIMA, MLR and NN model. The scope of the study in this paper, will be limited to the modelling the close price of the stock market and not forecasting. This is because, the limitation number of the data makes it not suitable for forecasting. Thus, modelling will be a better alternative to find a model that could fit with the data. These three models are selected since the data is recorded half-hourly for a day. Three models will be used for this purpose to represent the conventional and advanced models in time series. MLR model will represent the conventional model to be fitted with the stock market. ARIMA model will be a benchmark model for data fitting and NN model will represent the advanced model in this study.

2.0 METHODOLOGY

2.1 ARIMA Model

ARIMA models are the most frequent employed class of models for forecasting time series that can be transformed to be stationary using techniques such as differencing and logging [15]. The ARIMA model is a linear model that can show both stationary and non-stationary time series. Most academics use this approach to model univariate time series data.

ARIMA is an abbreviation that stands for autoregressive integrated moving average. Autoregressive terms are used to refer to the lags of differentiated series appearing in the modelling equation, whereas moving average terms are used to refer to the delays of forecast errors. In general, any time series that requires differentiation to become stationary is referred to as an integrated form of a stationary series. A non-seasonal ARIMA model is denoted by the notation ARIMA (p, d, q), where p denotes the number of autoregressive components, d denotes the number of non-seasonal differences, and q is the number of lagged forecast errors in the prediction equation. ARIMA’s generalized form is as follows.

\[
(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \epsilon_t
\]

or

\[
\phi_p(B) \nabla^d y_t = c + \theta_q(B) \epsilon_t
\]

where

\[
\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p
\]

is the non-seasonal autoregressive operator of order p

\[
\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q
\]

is the non-seasonal autoregressive operator of order q

\[
\nabla^d = (1 - B)^d
\]

is the difference operator d degree

\[
c\text{ is the constant.}
\]

The ARIMA model is a three-step process for determining the best model for the stochastic component defining a repeated time series. The first step is to identify the characteristics of the data such as trend, seasonality or irregular pattern. Next is to identify the parameters p, d, q of the model. This part can be determined by using the ACF and PACF plots. Lastly, after the model is determined from the ACF and PACF, few recommended models will be chosen, and the model's goodness of fit can be determined by utilizing the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). As previously stated, early identification, or simply specification, is a critical component of the model development process.

However, by mixing the various orders p, d, and q, there are an infinite number of viable models for any set of longitudinal data. As a result, these processes are repeated until a model is developed that is appropriate for the given data set. A variant of the Hyndman and Khandakar (2008) algorithm was utilized in this study to accelerate the model identification process. This approach combines unit root testing with the minimization of the Akaike’s Information Criterion (AIC) and Ljung-Box test statistics for ARIMA models [16].

2.2 Multiple Linear Regression Model

Multiple linear regression (MLR) is a mathematical technique for calculating the connection between a dependent variable and one or more independent variables [17]. A Multiple Linear Regression (MLR) assume that Y is a quantitative response variable and that x_i is a quantitative predictor variable. The MLR model is then frequently a very useful model. For the MLR model,

\[
Y_i = \alpha + x_{i,1} \beta_1 + x_{i,2} \beta_2 + \cdots + x_{i,p} \beta_p + e_i
\]

For i = 1, …, n. The Y_i is the response variable x_i is a p x 1 vector of nontrivial predictors. In this study the x_i will be based on the variables of open price, high price, low price, and volume of the stock market. The \( \alpha \) is unknown coefficients, and \( e_i \) is error term that follows
a normal distribution with a mean of 0. The parameter \( \beta_1, \beta_2, \ldots, \beta_p \) coefficient values will be automatically generated by the software Minitab.

The Gaussian or normal MLR models makes the additional assumption that the errors \( e_i \) are \( N(0, \sigma^2) \) random variable. This model can also be written as \( Y_i = \alpha + \beta^T x_i + e_i \) where \( e_i \sim N(0, \sigma^2) \), or \( Y | x \sim N(\alpha + \beta^T x, \sigma^2) \). The normal MLR model is a parametric model since, given \( x \), the family of conditional distributions is completely specified by the parameters or \( \alpha, \beta \) and \( \sigma^2 \). Since \( Y | SP \sim N(SP, \sigma^2) \), the conditional mean function \( E(Y | SP) \equiv M(SP) = \mu(SP) = SP = \alpha + \beta^T x \).

2.3 Multilayer Perceptron Neural Network

Soft computing techniques has been said to improve the forecast [18]. Neural networks have been proposed as an alternative model in time series forecasting because they are versatile in collecting nonlinear time series data [19]. Therefore, a multi-layer perceptron neural network was used for this study to observe whether it could improve the forecasting.

Figure 1 shows how neural networks is made up of a series of interconnected nodes known as neurons. Typically, these are stacked in layers. Then, a typical neural network will include an input layer, a hidden layer, and an output layer.

![Figure 1 Neural network structure](image)

The amounts of features or qualities you want to inject into the neural network is represented by the input layer nodes. The number of output nodes is proportional to the number of items you want to model or categorize. Since the data has open price, close price, high price, low price, and volume of the stock market, the close price will be the output and other variables will be treated as attributes in the input layers. In most cases, the hidden layer nodes are utilized to execute non-linear transformations on the original input attributes. In their most basic form, neural networks transport attribute input through the network to create a prediction, with a continuous or discrete output for regression or classification.

3.0 RESULTS AND DISCUSSION

In this research, the data used in this study is a three-and-half-hour data being recorded for every 5 minutes on 20th January 2022 from Bursa Malaysia. Total data used for the modelling process are 42. As illustrated in Figure 2, the original series has an irregular pattern. The autocorrelation function (ACF) of original series slowly decrease as can be seen from Figure 3, it shows that the series is not stationary.

![Figure 2 Time series plot of the Bursa Malaysia](image)

![Figure 3 ACF plot of the data](image)

Before using the ARIMA model, it is important to have a stationary data. Since the ACF plot shows that the data is not stationary, we differenced the data to make the data stationary. The time series plot and ACF plot after being differenced is shown as below.
From Figure 4 above, the time series plot has become stationary, and the trend also has been eliminated from the graph after the data being differenced with respect to 1.

As illustrated in Figure 6, the PACF of the first difference of series fluctuates around mean zero and the lag 1 has spike outside the red line which indicate the possible value for $p$ is 1. Then from the ACF plot, we also have a spike at lag 1 which can indicate that the possible value for $q$ is 1. And since we also differenced the data with respect to 1 to make it stationary, then $d$ value will be 1. From this decision possible ARIMA model is ARIMA(1,1,1) but we could also include two possible ARIMA model from this model. The three ARIMA models that can be identified using the ACF and PACF are.

As can be seen from Figure 5, the ACF plot also has becomes stationary after being differenced with respect to 1. Since the data has become stationary after being differenced, we could then proceed to modelling the ARIMA model by plotting the partial autocorrelation function (PACF) to determine the $p$, $d$ and $q$ lags for the ARIMA model. The $p$ value will be determined from PACF plot, the $d$ is from the differenced value that we choose, and the $q$ value will be determined from the ACF plot.

From these three models, we must check the adequacy of the models by using Ljung-Box statistics and AIC. For the Ljung-Box statistics, to obtain the adequate model, the p-value must be larger than the critical value, $\alpha = 0.05$ and if all the models satisfy the Ljung-Box statistics, then to find the best model for modelling id by referring to the model with the smallest AIC value.

By referring to Table 1, all of the ARIMA models are adequate and the smallest AIC is belong to ARIMA(0,1,1). Hence, we will use this model to model the data. The model of the ARIMA model can be seen as below in Figure 7.
The blue plot represents the original data, and the red line represents the model of the ARIMA model.

After model the data with ARIMA model, we model the data with MLR model. The closing price or actual price is the response variable for modelling in this study, as for the predictor variables we will use the open price, high price, low price, and volume. The fitting is illustrated in Figure 8. The equation of the MLR can be written as:

\[ Y_{close,t} = 0.013 - 0.4132x_{open} + 0.63x_{high} + 0.773x_{low} + 0x_{volume} \] (4)

From the equation (4), the volume variable does not give any meaning to the equation, therefore we could exclude the variable from the fitting process. The model of the MLR looks nearer to original data but if we look closely, ARIMA has better estimation compared to MLR. Most of the pattern of the data can be read by the model.

After model with MLR, we proceed to model the data with the neural network. Similar with MLR, the open price, high price, low price, and volume of the stock market will be the input layers. As for the hidden layer for neural network modelling, we choose 1, 3, 5 and 7 as the hidden layer and by using the residual minimum value, residual maximum values, and residual standard error, we will choose the best hidden layer from the smallest values among these values. The results of the residual values are as shown in below.

<table>
<thead>
<tr>
<th>H</th>
<th>Residual min</th>
<th>Residual max</th>
<th>Residual Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.00279</td>
<td>0.00323</td>
<td>0.002211</td>
</tr>
<tr>
<td>3</td>
<td>-0.003636</td>
<td>0.002744</td>
<td>0.002309</td>
</tr>
<tr>
<td>5</td>
<td>-0.005111</td>
<td>0.004167</td>
<td>0.002785</td>
</tr>
<tr>
<td>7</td>
<td>-0.005071</td>
<td>0.004669</td>
<td>0.002914</td>
</tr>
</tbody>
</table>

The data from Bursa Malaysia is more accurate when represented by a neural network with one hidden layer, according to the residual calculations. The reason for this is that the maximum residual value for this layer is the smallest and the residual standard error is quite small compared to the other H. Based on this result, the fitting plot can be seen as below in Figure 9.

From the plot comparison between MLR, ARIMA and NN model in Figure 7 until 9, NN model seem to give the best fitting. NN is known to be a model that could model based on the information that it has and train it to follow the original data as possible. In comparison to previous probability-based models, NN make no assumptions about the underlying probability density functions or other probabilistic information about the pattern classes [20].

To support the comparison plot, the RMSE and MAPE for each model is calculated to select the best model to model the Bursa Malaysia data between the three selected models and the result obtained is as below in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>MLR</th>
<th>NN</th>
<th>ARIMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.0529</td>
<td>0.0477</td>
<td>0.0507</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.4790</td>
<td>0.4222</td>
<td>0.5454</td>
</tr>
</tbody>
</table>
Both performance indicators show that NN model is the best among the three. As for ARIMA, for RMSE it shows smaller value compared to MLR but for MAPE the value is larger than the MLR. However, RMSE is preferable to be used as the performance indicator because when lower residual values are preferred, RMSE is more useful, thus, the result from the RMSE will be used to rank the modelling the performance. The NN is the best, ARIMA as the second best and MLR is the least good in this modelling analysis.

To check whether the model is good for forecasting or not, we used the ACF plot of the residuals from each modelling to see whether it shows white noise or not. From the ACF plot below in Figure 10, all three models show white noise which means all of them are suitable to be used for forecasting.

![ACF plots](image)

**Figure 10** The ACF plots of residuals for each selected model

Often when performing forecasting, modelling process is neglected to check whether the model is a suitable for forecasting or not. From this modelling outcome, it could help the stockholder to use the information from the output in the future to predict the price of the stock market. Modelling could give indication in model selection for forecasting before the real forecasting is done.

Based on the study, these three models are suitable for forecasting. Depends on the analyst to choose which models as the forecasting model or to proceed all three models for forecasting and compare, based on the modelling process outcome and the residuals ACF plot above, all three models are reliable for forecasting.

### 4.0 CONCLUSION

Neural network has improved the modelling of the stock market and furthermore nowadays, it has been used widely and easier to be used from any software. MLR as the most conventional compared to the other two models, although it gives the largest value of performance indicator, but the difference is relatively small. Thus, we can say that MLR is still convenient to be used these days. As for ARIMA, it has always been a benchmark model for a comparison because of the practicality and convenience to be used as a conventional and advanced model. For future works, a hybrid model could be considered to improve the modelling. Hybrid model could help finding the balance between the advantages and disadvantages of two models.

### References

6. Gligor, Cristian-Dragos Dumitru and Adrian. 2019. Wind Energy Forecasting: A Comparative Study between a Stochastic Model (ARIMA) and a Model Based on Neural Network (FFANN). The 12th International Conferences Interdisciplinarity in Engineering.