IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 13, 2019, accepted January 17, 2019, date of publication January 25, 2019, date of current version February 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2895252

A Prediction Approach for Stock Market Volatility Based on Time Series Data

SHEIKH MOHAMMAD IDREES, M. AFSHAR ALAM, AND PARUL AGARWAL
Department of Computer Science and Engineering, Jamia Hamdard, New Delhi 110062, India

Corresponding author: Parul Agarwal (pagarwal@jamiahamdard.ac.in)

This work was supported by the Hamdard National Foundation, India.

ABSTRACT Time series analysis and forecasting is of vital significance, owing to its widespread use in various practical domains. Time series data refers to an ordered sequence or a set of data points that a variable takes at equal time intervals. The stock market is considered to be one of the most highly complex financial systems which consist of various components or stocks, the price of which fluctuates greatly with respect to time. Stock market forecasting involves uncovering the market trends with respect to time. All the stock market investors aim to maximize the returns over their investments and minimize the risks associated. Stock markets being highly sensitive and susceptible to quick changes, the main aim of stock-trend prediction is to develop new innovative approaches to foresee the stocks that result in high profits. This research tries to analyze the time series data of the Indian stock market and build a statistical model that could efficiently predict the future stocks.

INDEX TERMS ARIMA model, forecasting, stock market forecasts, time series analysis, Box-Jenkins method.

I. INTRODUCTION

Future being a mystery is always a challenging task to predict. From the ages, human nature has always been more curious about the future. Forecasting refers to an approach of predicting what is likely to occur in the future by observing what has happened earlier in the past and what is occurring at present. In other words, it is just similar to driving a car in forward direction by keeping an eye on the rear-view mirror of a car. Forecasting is an important problem but with vital importance in all areas of real world like business and industry, medicine, social science, politics, finance, government, economics, environmental sciences and others. In recent years, with the rise of social media and other promising applications, stock market forecasting has attracted huge interest from people in general and business in particular. Advances in financial sectors are responsible for growth and stability of overall economy [1]. In business domain, forecasting is considered as one of the difficult tasks owing to the various complexities of the market [2]–[4]. But it is important since it helps to plan for future by providing a solid idea about how to allocate the resources and plan for foreseen costs in the forthcoming period of time. Investors always try to monitor the risks in real time so that the return on investments could be higher. Forecasting helps in safeguarding the trade of securities among the buyers and the sellers as well as elimination of the risks involved.

This paper discusses an ARIMA (Auto Regressive Integrated Moving Average) model for prediction of stock market movement. An ARIMA model is a vibrant uni-variate forecasting method to project the future values of a time series. The remaining of the paper is arranged as follows. Section II describes the forecasting process. Section III discusses the forecasting techniques, while section IV discusses the financial forecasting. Section V presents the Time Series Analysis. In section VI, we try to explain in detail, the various statistical models for forecasting. Data collection and methodology are discussed in section VII, while the final section of this paper provides a brief conclusion.

II. THE FORECASTING PROCESS

A Forecast refers to a scientifically calculated guess. Forecasting is an activity of vital importance for every business organization or even the government. This is mainly due to the reason that based on the forecasts the future strategies are developed. A good forecast needs to be accurate, reliable, time efficient, easy to understand, cost efficient and as simple as possible. There are basically three categories into which the forecasting problems are characterized. These are

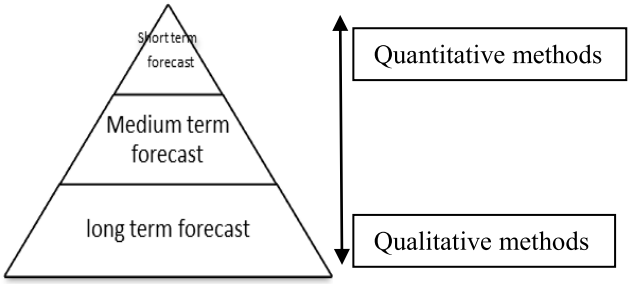

FIGURE 1. Categories of forecasts.

as: - ''short-term, medium-term and long-term forecasts'' [5]. Short-term forecasting problems include predicting future events only up to a small time period. Such forecasts are particularly meant for a period ranging from few days, weeks, months or less than a year into future. Medium term forecasts are generally put into practice when we try to predict events that range for a period of one to two years into future. On the other hand, the long term forecasts can extend over the medium term forecasts by many years. Both short and medium term forecasting approaches are put into practice for predicting wide range of events that may include operations management and budgeting or may even include selection of new research and development projects. However, the long-term forecasts are mainly employed in events that require strategic planning. The short and medium-term forecasting tactics are normally focused around recognizing, modeling, and extrapolating or generalizing the patterns found in historical data. An illustration of these categories of forecasts is given in figure below:

A. STEPS IN FORECASTING PROCESS

Irrespective of the forecasting approach that is being employed while making a forecast, there are some elementary steps [6] that need to be followed. There are typically five general steps in any forecasting task

STEP 1: IDENTIFYING PROBLEM

Since the forecasts are being anticipated with the aim to predict and plan for the future events, so it is important to know who needs these predictions, in what way these forecasts will be used, and in what way the forecasting function fits inside the organization that requires the forecasts. In this phase an analyst has to devote some time to talk with every person involved in collection of data, maintenance of databases, and the ones who will use the forecasts for future planning.

STEP 2: GATHERING OF INFORMATION

In this phase, the forecaster determines the related variables that need to be considered and chooses how to assemble the data. This data can be either primary data or secondary [7]. The primary data does not have any previous existence and is collected directly from the respondents. This data is considered very important in contrast with all other forms of data but, its reliability may raise many questions. Secondary data is the data that has been collected previously at some time.

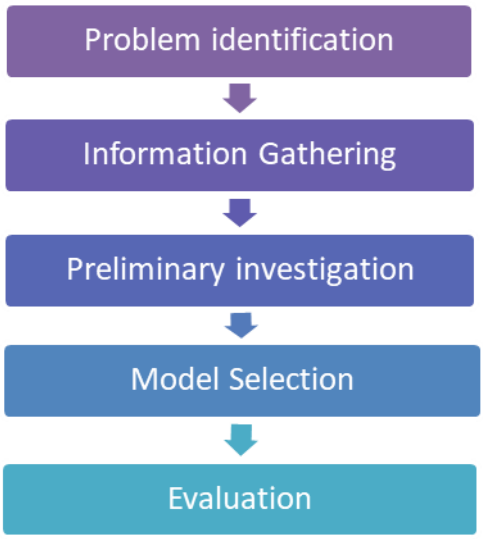

FIGURE 2. Steps in forecasting.

STEP 3: PRELIMINARY ANALYSIS

This phase performs an analysis of data and advocates whether the data collected is beneficial or not. Also, the preliminary analysis is helpful in revealing the patterns or trends and thus helps in choosing the models that best fit in. An important thing that is done in this phase is to check for the redundant data and to check it down. As such it simplifies the forecasting process by omitting the redundant data.

STEP 4: CHOOSING AND FITTING OF MODELS

Once all the information is collected and analyzed, it requires choosing the prediction model that could give the best prediction results. Selection of a model is largely governed by the availability and nature of data.

STEP 5: USAGE AND EVALUATION OF FORECASTING MODEL

Once the prediction model is selected and its various parameters have been estimated, the model is then put into practice to make predictions. The next step involves the evaluation of the forecasting model. The validation of this model can be precisely done only after the data for the forecasting time period have become available.

B. FORECASTING TECHNIQUES

Forecasts are being employed in a wide range of situations. In spite of all the numerous real life circumstances that involve forecasts, there exist only two types of forecasting techniques to be implemented [8] [9]:

a) Qualitative Forecasting models.
b) Quantitative Forecasting models.

The Qualitative forecasting models are generally subjective in nature and are mostly grounded on the opinions and judgments of experts. Such types of methods are generally used when there is little or no past data available that can be used to base the forecast. An example of the qualitative forecasting model is the Delphi method which engages

a board of experts who are supposed to be well-informed about the problem. Hence, the outcome of the forecast is based upon the knowledge of the experts regarding the problem.

On the other hand, the Quantitative forecasting models make use of the data available to make predictions into future. The model basically sums up the interesting patterns in the data and presents a statistical association between the past and current values of the variable. Likewise, we can say, that quantitative forecasting models are used to extrapolate the past and present behavior into future. Some examples of the Quantitative models include the regression analysis models, smoothing models and the time series models.

III. FINANCIAL FORECASTING

A Financial forecast can be elaborated as a forecast regarding the future business circumstances that are expected to affect a company, organization, or a country. A financial forecast visualizes the movements in relevant historical data and then projects these movements in order to help the decision-makers by providing information regarding the forthcoming financial status of the company. Simply we can say that, a financial forecast is a business plan or budget for a business. It is basically considered as an estimate of two vital forthcoming financial outcomes of a business – the projected revenue and the costs. Prediction of the financial state of a business is never an easy task; with most of the forecasts go wrong. But still it is a better idea to have an educated guess about the future than to not forecast at all, since ''Best'' educated guesses about future are more valuable for purpose of planning and budgeting. There are many advantages of an efficient financial forecast as below:

a) Controls the financial practicability of a new business project. Thus aides in designing of models for how the business would perform economically, if certain approaches, procedures and tactics are undertaken.
b) Helps in comparing the real financial operation with the forecasted financial plan and make modifications where needed.
c) Drives the business in an accurate direction and controls the flow of cash.
d) Specifies a point of reference against which the future performance can be supervised.
e) Identifies the probable threats and the cash deficits in order to keep the business away from the financial catastrophe.
f) Helps in knowing the future cash needs and whether any additional borrowing is needed.

IV. STOCK MARKET PREDICTION

A ''share market or equity market or a stock market'' is a public market that exists for issuing, buying and selling of stocks or shares [11], [42]. A Stock denotes a partial ownership in a company or an industry, with rights to share in its profits. A person, who invests in a stock or buys a stock of a company, is termed as stockholder of that company. A stock market is a dynamic ingredient of a free- market economy where organizations get access to capital in exchange by providing the sponsors a share in proprietorship of the organization. A stock market plays a significant role for organizations to raise revenue along with the debit markets. The stock market allows the business units to be openly traded and raise surplus financial capital for growth and development by selling shares of the proprietorship of a company in a public market. History has proved that stock prices and prices of other assets have a dynamic impact on the economic activity and is also an indicator of social mood state. An economy is considered as a rising economy if its stock market is on rise.

India, being one amongst the fast developing economies of the world provides investors, both domestic as well as foreign investors, an opportunity to make right investments in Indian stock market. India has two main stock exchanges as:

- ''National Stock Exchange (NSE)''
- ''Bombay Stock Exchange (BSE)''

The BSE is the oldest stock exchange in India that came into existence in 1875 while the NSE came much later and started working in 1994. All of the major companies or business firms in India are listed under these two stock exchanges. BSE has around 5000 firms listed to its name while the rival NSE has listed about 2000 firms to its name [11]. In spite of having a lesser number of firms listed to its name, NSE still enjoys the dominant share in share trading with about 70% of market share [11], [12]. The performance of overall stock market is calculated by Index. It is an indicator of the overall stock market movement. Indian stock market has two main indicators or indexes which are as:

- Nifty.
- Sensex.

There are many other indexes that reflect the performance of a particular sector example bank index, IT index, automobile index and many others, but the above two are the prime ones. The Sensex (or **sens**itive Ind**ex**) comprising of 30 (thirty) stocks, reflects the whole market sentiments of major firms listed under Bombay Stock Exchange (BSE). The UP or DOWN movement of Sensex indicates that the majority of stocks prices under BSE have gone up or down respectively. Nifty, on the other hand comprises of 50 (Fifty) stocks, is an indicator of the firms listed under National Stock Exchange (NSE) [13]. Stock market prediction can be concluded as an approach to estimate the upcoming worth of a company's stock traded on a stock exchange. A constructive prediction of a company's stock price can bring fortunes to the company.

Calculation of the Index

There are two main components that are required while calculating the value of index for the next day. These components include ''the index value'' and ''the total market capitalization of the previous day''. The index is calculated as follows:

Index Value = (Todays Market Capitalization / Yesterdays Market Capitalization) x Yesterdays Index point

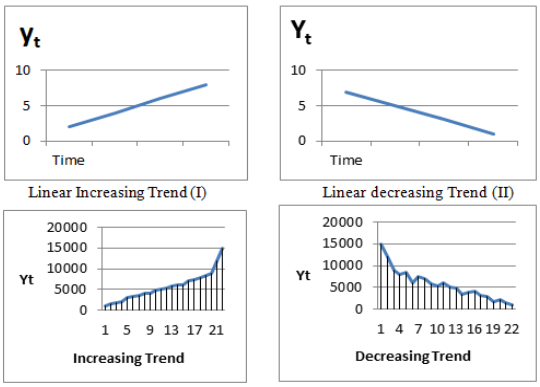

FIGURE 3. Trend component of time series.

V. TIME SERIES ANALYSIS

Time-series data is a well-ordered arrangement of data or a set of data points that a variable takes at equally placed time intervals. The objective of time series analysis is to develop models which are able to describe the given time series with a reasonable amount of accuracy. The recent years have witnessed a manifold increase in the research regarding the time-series modeling. The systematic research of Time series is of great significance so as to forecast regarding the changeability of the data in future, based on the past observations of the data. So, basically the Time series forecasting can be labeled as an approach of making an estimate about the future by understanding the past [14]. Time series analysis may be regarded as a decision making factor, for the future plan and estimate. It is of vital significance in numerous real-world areas which may include business, economics, science and engineering, banking and finance, governance, and many others [15], [16]. As such, proper attention need to be taken while selection of a time series model, so as to achieve better results of forecasts [17].

A ''Time Series'' is defined as a chronological set of data points, which are monitored usually over consecutive intervals of time. Mathematically it can be represented as a set of vectors $X(t)$, $t = 0, 1, 2, 3, \ldots$ where t signifies the time intervened [18].

$\boldsymbol{X_t: t \epsilon T}$ where ''T is an ordered set of time''.

A time series can be uni-variate or multi-variate depending upon the set of observations of the variable. A time series comprising of observations related to only one variable is referred to as uni-variate time series while the one with observations related to more than one variable is referred as a multi-variate time series. Furthermore, a time series can be of two types [19], [20]:

- Continuous
- Discrete

A time series is continuous if the observations are measured at each point of time. The continuous time series are relatively very rare in economic data. Some examples of continuous time series include temperature readings of a city, flow of a river, and others. In a ''Discrete time series'' observations are calculated at discrete instants of time. The consecutive observations are noted at equal spaced time intervals in case of a discrete time series. These discrete time intervals may be hourly, daily, weekly, monthly or yearly. Some examples of discrete time series include the population of a country, manufacturing in a firm, exchange tariffs among two different currencies and others. While considering the discrete time series, the variable under consideration is expected to be measured as a continuous variable by implementing a real numerical scale. Moreover, it is easier for a continuous time series to be changed into discrete time series by combining data together over a definite time interval.

A. TIME SERIES COMPONENTS

Considering the patterns in the time series is an important aspect in time series analysis as it helps in selection of models. However the patterns in data can be best analyzed when the underlying different components of time series are examined [21]. Any time series is composed of the following components [22].

- *''Trend component (T)''*
- *''Cyclic component (C)''*
- *''Seasonal component (S)''*
- *''Irregular component (I)''*

All these components can be combined in many ways. However, it is commonly supposed that they are multiplied or added, likewise as below:

$$''Y(t) = T(t)xC(t)xS(t)xI(t)''$$
$$''Y(t) = T(t) + C(t) + S(t) + I(t)''$$

where:

Y(t) = Time series Observation.

In the above multiplicative model the assumption is that all these constituents of a time series are not essentially independent, they are interrelated and as such are supposed to impact each other. However the additive model consents to assume that the four components are independent of one other.

The trend component, an outcome of the long term movements of various factors forms the chief component in a time series. A time series may exhibit upsurge or downward movement or may show steady movements over an extensively long duration of time. This overall movement is referred to as a trend component of a time series [23]. Based on the pattern exhibited by a time series, a trend may be positive or negative in nature. However, there are cases when a time series does not show either an upward or a downward pattern, such series are stationary and have a constant mean. Example – Population growth time series displays an ascending trend, while descending trend is displayed in series relating to epidemics.

Trend Component can be represented graphically as below in following figures:

Seasonality occurs when the time series is influenced by seasonal factors and repeat at regular periodic intervals like weekly, fortnightly, monthly, or through the identical quarter

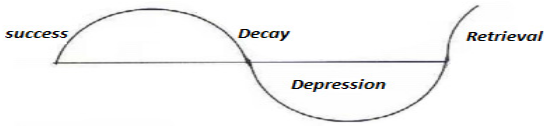

FIGURE 4. Phases of a business.

for each year. There are many factors like climate and weather conditions, traditional customs and habits and many others responsible for causing seasonal variations. For instance - sale of woolen cloths increase in winter, sale of ice-cream and cold drinks increase in summer, festival season etc. Businessmen, shopkeeper and producers monitor the seasonal variation very closely in order to make proper future plans and conquer the best returns over investments.

Cyclic component is present in a time series when the time series displays a rise and fall of irregular time period. The interval of a cycle is expressed by the type of business or industry being examined. The cyclic component usually stretches over longer intervals, which may range from two or more years. Cyclical variation is displayed by almost all categories of economic and financial time series.

Illustration: Four segments of a business cycle [22] with figure.

i) Success
ii) Decay
iii) Depression
iv) Retrieval.

Irregular or random variation component of a time series is unpredictable. They are initiated by unpredictable influence that are neither regular nor repeat in some pattern. The random variations may be caused by various external factors like earthquake, war, flood, etc. There is no statistical procedure to measure the random fluctuations in a time series. While making predictions, the motive is to model all the components of a time series to a point that we end up with only one component that remains unexplained called as the random component.

B. EXAMPLES OF TIME SERIES DATA

1). Average monthly performance of the Stock Exchange during the period of 2015 to 2017 as measured by the General Index.
2). Numbers of births per year in India from 1947 to 2018.
3). Monthly rainfall in Delhi over last 15 years
4). Daily stock prices for the last 5 years.
5). Monthly car sales of an automaker for the last 3 years

VI. FORECASTING MODELS

The selection of prediction model is of remarkable significance as it reveals the fundamental structure of the time series. Time series models can be ''linear or non-linear'' based on whether the present value of the series is a ''linear or non-linear'' function of earlier observations. Uni-variate time series models try to interpret a number of

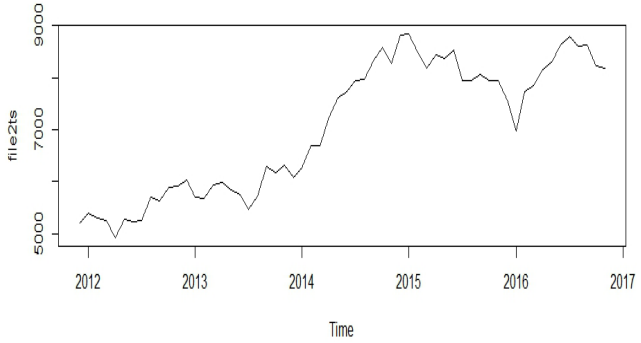

FIGURE 5. Monthly –Nifty (2012-2016).

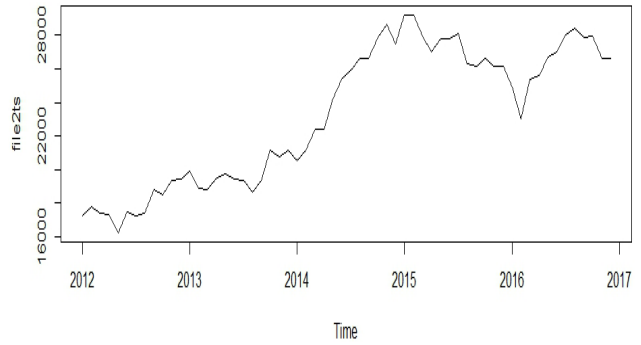

FIGURE 6. Monthly – Sensex (2012-2016).

economical phenomena through the historical behavior of a dependent variable. There are mainly two widely used linear models [24]–[27], [44], [46], ''Auto-Regressive models (AR)'' and ''Moving Average (MA) models''. A grouping of these two models result in the formation of another model referred as Auto-regressive Moving Average (ARMA). We have one similar type of model as ''Auto-Regressive Integrated Moving Average model'' (ARIMA) [28]–[30], [33], [45]. All these models have been elaborated below:

A. AR (p) MODEL

An ''Auto regressive (AR) model'' is used to calculate the future behavior of a variable under consideration, using a linear combination of historical values of the variable [34], [35]. The word ''*auto*-regression'' designates that there is a regression of the variable against itself. It could be considered as a function of past values. i.e.

$$\text{“}\mathbf{y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, \ldots\ldots \epsilon_t)}\text{”}$$

Auto regressive model that is influenced by 'p' of its earlier values, referred as AR(p) is represented as:

$$\text{“}\mathbf{y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + + \phi_p y_{t-p} + \epsilon_t}\text{”}$$
$$\mathbf{y_t = c + {}_{i=1}{}^{p}\phi_1 y_{t-I} + \epsilon_t} \quad (1)$$

where ϵ_t is the error term.

In AR(p) model, P is the parameter, which can change its values as:

When $p = 0$, AR(0) becomes $y_t = c$

When p = 1, AR(1) becomes $y_t = c + \phi_1 y_{t-1}$ and so on for the values of p as 2, 3...

Normally we restrict autoregressive models to stationary data, but When p≥3, the restrictions are much more complicated. We need to find the best value of p for forecasting.

B. MA (q) MODEL

Instead of using the earlier values of a variable for forecasting as in case of AR(p) model, the "Moving Average (MA) model" uses earlier errors terms for prediction. MA model can be considered as a function of error terms as:

$$y_t = f(\epsilon_t, \epsilon_{t-1}, \quad \epsilon_{t-2}, \epsilon_{t-3}, \ldots\ldots \epsilon_{t-q})$$

In regression, we get an error term when we regress a series with its past values as:

Regression of y_t over y_{t-1}:

$y_t = \mu + \theta_1 \epsilon_{t-1} + \epsilon_1$ where μ = constant, and ϵ_1 = error

Similarly we get other error terms as ϵ_2, ϵ_3 and so on when we regress the series with different values. So instead of using the past values, we use the error terms in this model. As such the "moving average model" can be put forward as:

$$\text{"}y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots\ldots + \theta_q \epsilon_{t-q} + \epsilon_t\text{"}$$
$$y_t = \mu + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \quad (2)$$

The error terms ϵ_t are supposed to be white noise processes, i.e having "zero mean" and "variance constant" σ^2.

μ represents the "mean of the series", and "θ_j (j = 1, 2, 3, 4..q)" represent the parameters of the model and "q is the order of the model". MA model is more difficult than AR model to fit to a time series as the random error terms are not foreseeable [37].

C. ARMA (p,q) MODEL

"Autoregressive" (AR) along with "moving average" (MA) models are used collectively to get a new unit of time series models referred as "ARMA models" (i.e. AR + MA = ARMA model) [35]–[37], [43]. The general notation for the "ARMA (p,q) model" is as:

$$\text{"}y_t = (c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t) + (\mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + .. + \theta_q \epsilon_{t-q} + \epsilon_t)\text{"}$$
$$y_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_1 y_{t-1} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \quad (3)$$

In the above given equation, the parameters 'p' and 'q' correspondingly refer to the "auto-regressive" and "moving average model".

D. ARIMA (p,d,q) MODEL

"ARIMA (*AR+I+MA*)" stands for "Auto-regressive Integrated Moving Average". This model is also often called by the famous "Box-Jenkins model". The "ARMA model" is best suited for the stationary time series data but the thing is that most of the time series data from real world shows non-stationary behavior. This model claims that a non-stationary series could be changed to stationary by means of differencing it [37], [46]. The common form of an "ARIMA model" for y_t is as:

$$\text{"}y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + .. + \theta_q \epsilon_{t-q} + \epsilon_t\text{"} \quad (4)$$

where y_t is "differenced time series" that might have been differenced once or more.

This model is called as "ARIMA (p,d,q)" model, in which the parameters p, d, q represent the following:

'p' is the "order of auto-regressive part",

'd' is the "degree of differencing",

'q' is the "order of the moving average part".

The "ARIMA model" combines 3 basic methods:

1) AUTO-REGRESSIVE (AR)

This specifies that the values of the given time series is to be regressed with its own lagged value. This is specified through the "p value of an ARIMA model".

2) DIFFERENCING PART (I FOR INTEGRATED)

Integration is implied as the inverse of differencing. It is the degree on differencing that needs to be done on data. In order to transform a "non-stationary time series into a stationary one", the series needs to be differenced. Differencing can be implied as:

D(i) = Data (i) – Data(i-1).

This differencing of the time series is represented by the 'd' value of the "ARIMA model". There are some situations that may arise with the value of 'd', likewise below:

When 'd' = 0, it signifies that the series under consideration is stationary, so we don't require to take the difference of it.

If 'd' =1, it signifies that the current series is not stationary, we need to take the first difference of the series.

If 'd' = 2, it signifies that the series under consideration has been differenced two times.

3) MOVING AVERAGE (MA)

The "moving average" component of an ARIMA model is denoted by 'q'. This simply refers to the total number of lagged values of the error term. For instance, when 'q' = 1 it means that there exists an error term and there is auto-correlation with one lag.

VII. DATA COLLECTION

The work in this research paper is focused on the data regarding the stock market. Data considered as raw oil, is being generated with every passing second [31]. This pragmatic study began with the analysis of Indian stock market data related to Sensex and Nifty. The publically available Stock market data sets contain historical data about all the stocks [32] has been collected from yahoo. The dataset specifies the "opening price, lowest price, closing price, highest price, adjusted closing price and volume" against

each date. The historical data of the Indian stock market collected through a span of five years beginning from ''January 2012 to December 2016'' has been taken into consideration for this work. The data has been divided into two parts – ''the training part and the testing part''. The ''training part'' from the time series data is used for formulation of the model while the ''testing part'' is used for the validation of the proposed model.

A. METHODOLOGY

Auto-regressive processes have a certain degree of unpredictability or randomness built in, that occasionally makes it capable to predict future trends pretty well. However, this needs to be assumed that they are never 100% accurate [38], [39]. After analyzing the time series data collected regarding the stock market, the 1st thing to do, is to ensure that the series is stationary or not. If the series is non-stationary, then the series has to be differenced so as to make it stationary. As such, we need to find the auto-co-relation and partial auto co-relation of the series. ARIMA model relies on the stationarity of the series [39]. So we will start with a fleeting portion about the stationary time series.

1) STATIONARITY OF A TIME SERIES

A time series needs to be lacking trend and seasonality, in order to be stationary. Such type of time series are characterized by having a constant variance and constant mean over a given period of time. The ''trend and seasonality'' component may affect a time series at different instants [38]. As ARIMA model takes into account, the earlier values of the series to model its prediction, so modeling a steady series with regular properties involves little insecurity. In order to design a model that is efficient in predicting future values of series, the primary time series has to be Stationary one. There are certain tests that assist in checking whether the series is stationary or not [33]. Some of these include ''W-D test'', ''Auto-correlation function (ACF)'', ''Partial auto-correlation function (PACF)'', ''L-jung-Box test'', ''t-statistic test'', the ''Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test'' and ''Augmented Dickey-Fuller unit root test (ADF test)''. Example- The ''augmented Dickey-Fuller (ADF)'' test is a standard statistical test for stationarity. In the ADF test, if the value of 'p' is less than 0.05 or 5% level for a time series, then the series is supposed to be stationary. However, there are cases when the series is non- stationary, at such times the ''p value is greater than 0.05 or 5% level''.

A 'non-stationary time-series' needs to be corrected by means of differencing [40]. An easy way to alter a non-stationary time series into stationary one needs to compute the differences between consecutive observations. This is referred as **differencing** i.e. ($\mathbf{y_t – y_{t-1}}$).

This differencing is the 'I' (integration) part of ARIMA model and denoted by 'd'. If we put on the differencing method twice or more, this gives birth to 2nd order difference and so on. These differenced values are then noted and thus it gives rise to a new dataset of the time series nature that can be used to test and discover new remarkable statistical properties and correlations.

2) DETERMINATION OF ACF AND PCF

An important step while selecting the model is the determination of ideal parameters for the model. Plotting the ACF and PACF against the consecutive time lags for the series is one simple approach to choose the parameters of the model. ''ACF and PACF'' are statistical methods that signify the relation between the observations in a time series with one another [40], [41]. They aid in defining the parameters of ''AR and MA'' terms.

The general form of ACF is as:

$$\frac{Covariance\ (Xt, Xt-h)}{Std.dev(Xt).Std.dev(Xt-h)} = \frac{Covariance\ (Xt, Xt-h)}{Variance(Xt)}$$

The above equation gives the ACF between x_t and x_{t-h}, Where x_t 'denoted the value of time series at time t for h values $= 1, 2, 3 \ldots$' etc.

In case of 'AR models', ACF for time series will shrink exponentially, so, PACF is implemented to identify the order of 'p'. ''Partial autocorrelation plots (PACF)'' signify the correlation between a variable and its lags and are very beneficial for describing the order of the 'AR(p) model'. The common form is of PACF is as:

$$\frac{Covariance\ (y, X3|X1, X2)}{variance(y|X1|X2)variance(X3|X1, X2)}$$

where: ''y = response variable'', ''X_1, X_2, and X_3 are the predictor variables''.

In case of 'MA models', the 'PACF' for time-series will shrink exponentially, so the order of 'q' in a MA method is identified through an ACF plot.

3) FRAMEWORK

See Figure 7.

4) METHOD

The first thing to do for analysis of the time series data requires plotting the data as below.

The next important thing that needs to be done involves the decomposition of the time series data into its essential constituents. Since the data under consideration is the time series data from indian stock market, the figure below shows the components of time series data related to stock market.

Decomposing the time series data helps in revealing a lot of hidden patterns inside the time series. A time series in general consists of four components as given in the figure above. We can also see the effect of season on the time series data by using the Boxplot function in R. The Boxplot for both the Nifty and Sensex time series data is as:

The Box plots provide a pictorial representation of the effect of season on time series data. The Boxplot helps in analyzing for each month, how the data is varying over the various years of time series [41]. Box plots

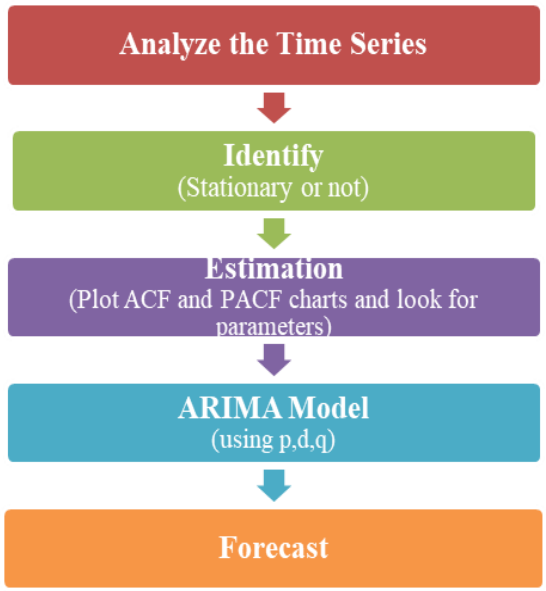

FIGURE 7. Framework for prediction.

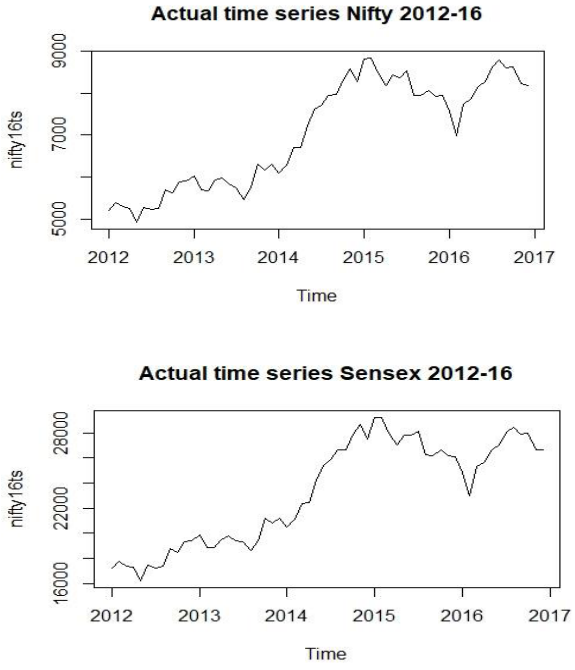

FIGURE 8. Time series data of Nifty (2012-2016) and Sensex (2012-2016).

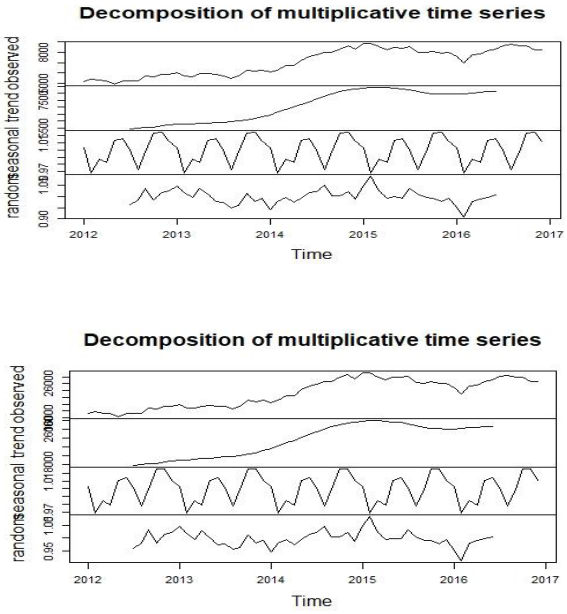

FIGURE 9. Decomposition of Nifty and Sensex time-series data.

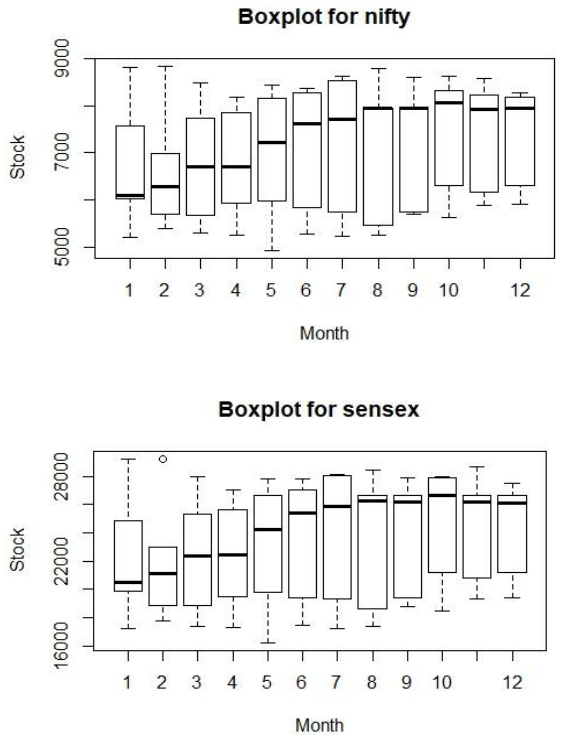

FIGURE 10. Boxplot for Nifty and Sensex time-series data.

are of great use in explanatory data analysis to demonstrate the shape of the distribution, its central value and variability.

Since the stock prices are based on returns, and the return value is based on percentage, so we need to change our time series data into logarithmic format. The next thing we need to do is to plot the ''ACF and PACF'' plots meant for time series data. These plots are as:

While observing the ACF and PACF plots we conclude that both of these series are stationary. As such we need to calculate the difference of lags. The plot showing the difference among the lags for both Nifty and Sensex is plotted below:

Next we use the 'dickey fuller test' to screen for stationarity and discard 'the null hypothesis of non-stationarity' across both the original time-series and the differenced time-series of both Nifty and Sensex time series data. We can also plot the ''ACF and PACF'' of the differenced series

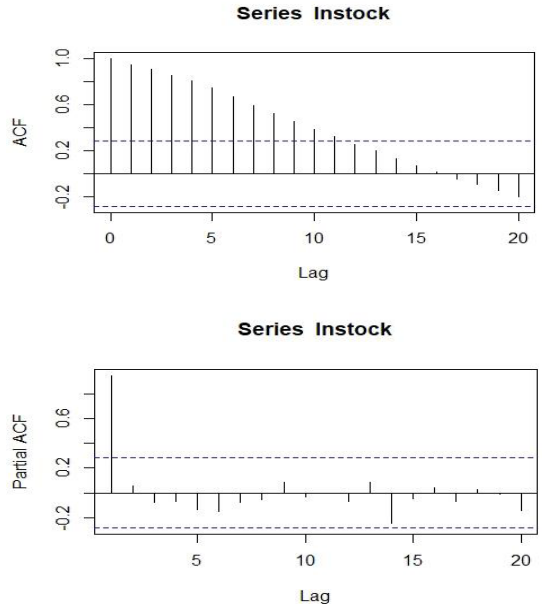

FIGURE 11. "ACF and PACF" plot for Nifty time series data.

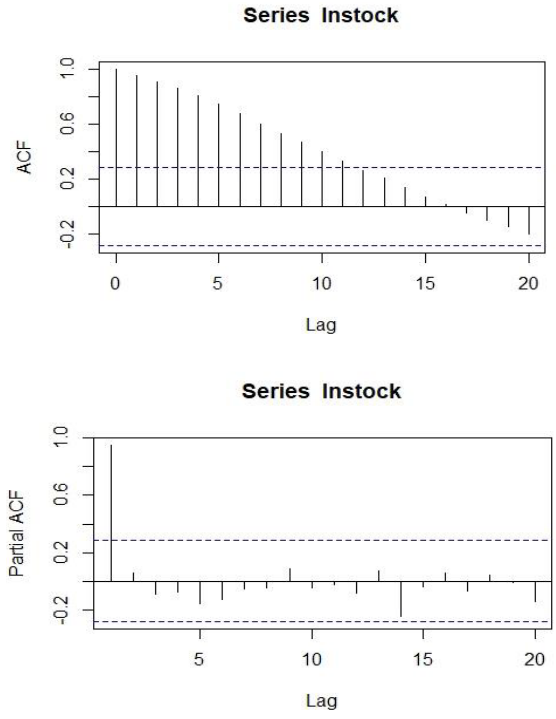

FIGURE 12. "ACF and PACF" for Sensex time series data.

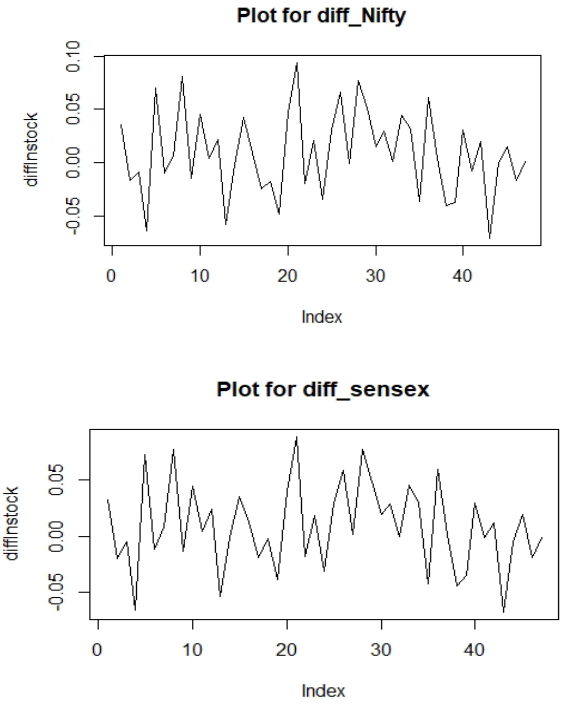

FIGURE 13. Difference plots of Nifty and Sensex.

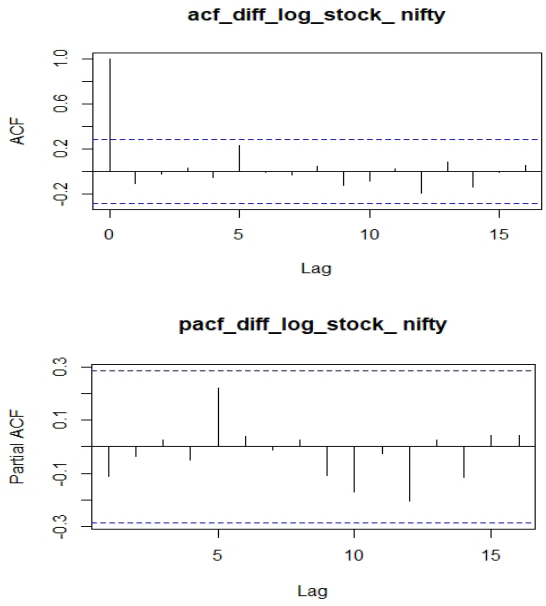

FIGURE 14. "ACF and PACF" (differenced Nifty series).

of both Nifty and Sensex for more conception regarding stationarity.

The ACF and PACF plots of differenced Sensex series are given as below in figure 15:

5) MODEL SELECTION

In the next step we go for forecasting the series. We use the "ARIMA (0, 1, 0) model" for predicting the next values in the time series. We use the auto.arima () function in R to get the results. Auto.arima() function chooses the best parameters of "ARIMA(p,d,q)" to get the forecasted series. The auto.arima() function uses a 'trace' that justifies why the parameters (p,d,q) chosen are best suited for the "ARIMA(p,d,q) model". In case of the current time-series regarding the Nifty and Sensex, the results of trace function are as:

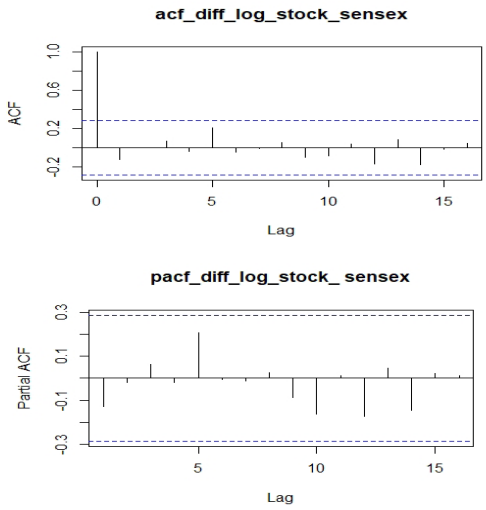

FIGURE 15. "ACF and PACF" (differenced Sensex series).

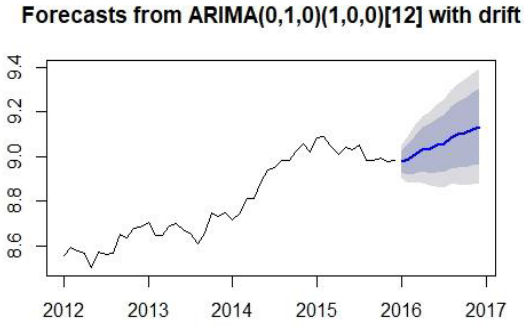

FIGURE 16. Plot for nifty 2016 (predicted series).

Case a: Test time series data from Nifty
auto.arima (lnstock_Nifty, ic="aic", trace = TRUE)
"ARIMA(2,1,2) with drift" : Inf
"**ARIMA(0,1,0) with drift" : −168.2363**
"ARIMA(1,1,0) with drift" : −166.8252
"ARIMA(0,1,1) with drift" : −166.8558
"ARIMA(0,1,0)" : −167.7504
"ARIMA(1,1,1) with drift" : −164.8622
"***Best model: ARIMA(0,1,0) with drift"***
Series: lnstock_Nifty
"ARIMA(0,1,0) with drift"
"Coefficients"
'drift'
0.0090
s.e. 0.0057
"sigma^2 estimated as 0.001534: log likelihood=86.12"
AIC = -168.24 AICc = −167.96 BIC = −164.54

Basically 'AIC' is what defines the accuracy of the model. The lowest value of "AIC" suggest the best model for prediction. In this case it is (0,1,0). For the Test time series data from the Sensex, the auto.arima() with trace depicts the following results:

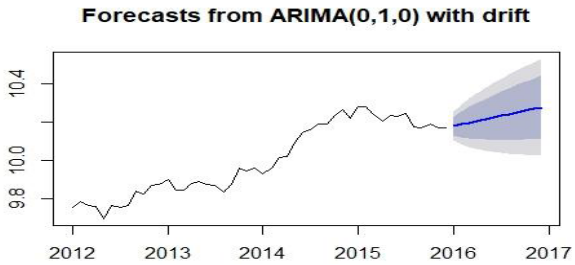

FIGURE 17. Plot for Sensex time series (Predicted 2016).

Case b: Test time series data from Sensex:
auto.arima(lnstock_Sensex, ic="aic", trace = TRUE)
"ARIMA(2,1,2) with drift" : Inf
"**ARIMA(0,1,0) with drift" : −172.4501**
"ARIMA(1,1,0) with drift" : −171.2168
"ARIMA(0,1,1) with drift" : −171.2201
"ARIMA(0,1,0)" : −171.8132
"ARIMA(1,1,1) with drift" : −169.2248
"***Best model: ARIMA(0,1,0) with drift"***
Series: lnstock_Sensex
"ARIMA(0,1,0) with drift"
"Coefficients":
drift
0.0089
s.e. 0.0054
"sigma^2 estimated as 0.001403: log likelihood=88.23"
AIC=−172.45 AICc=−172.18 BIC=−168.75

6) RESULTS

After we have got the best suited parameters (p,d,q) for the model, the next step is to predict the series using the test time series for Nifty and Sensex. The predicted series for both nifty and Sensex for the year 2016 is plotted in the figure 16 and figure 17 below:

Nifty (2010-2016) Time series

Since the predicted time series can never be 100% because of the irregular component, a comparison of the predicted series with the actual series shows roughly a deviation of 5% mean percentage error for both Nifty and Sensex on average. The monthly series values for both Nifty and Sensex are given as:

We can also use the L-Jung-Box test for the validation of the predicted series to check if the residuals are random. This test is used to decide "if the residuals of our time series follow a random pattern, or if there is a significant degree of non-randomness". The reason being that, if there are correlations between the residuals then we can say that the model is not good enough to handle the time series behavior and that is going to create problems in our time series. So we use the L-Jung Box test with our Null hypothesis that our residuals are random. The results of L-Jung Box test for both Nifty and Sensex predicted series is as:

L-Jung Box Test Results (Nifty)
Box.test(fitlnstock$residuals,lag = 10,type = "Ljung")
Box-Ljung test
data: fitlnstock$residuals

TABLE 1. Nifty predicted values (2016).

Month	Actual Stock value of Nifty. close (2016)	Forecasted Nifty
1	7563.55	7919.173
2	6987.05	8006.043
3	7738.40	8184.790
4	7849.80	8359.883
5	8160.10	8396.378
6	8287.75	8513.470
7	8638.50	8573.895
8	8786.20	8831.254
9	8611.15	8937.142
10	8638.00	9011.454
11	8224.50	9156.635
12	8185.80	9263.184

TABLE 2. Sensex predicted values (2016).

Month	Actual stock value of Sensex. close (2016)	Forecasted Value
1	24870.69	26350.90
2	23002.00	26586.34
3	25341.86	26823.88
4	25606.62	27063.55
5	26667.96	27305.36
6	26999.72	27549.33
7	28051.86	27795.47
8	28452.17	28043.82
9	27865.96	28294.39
10	27941.51	28547.20
11	26652.81	28802.26
12	26626.46	29059.60

X-squared = 4.7061, df = 10, **p-value = 0.9099**

L-Jung Box Test Results (Sensex)

Box.test(fitlnstock$residuals,lag = 10,type = "Ljung")

Box-Ljung test

data: fitlnstock$residuals

X-squared = 5.3276, df = 10, **p-value = 0.8682**

Since the values of 'p' for both the predicted series of Nifty and Sensex are greater than **0.05**. This specifies that there is probable a high degree of uncertainty demonstrated by our residuals and so our "ARIMA model" is free of autocorrelation.

VIII. CONCLUSION

This paper introduces the concept of time series analysis and forecasting in the perspective of Indian economy. The major downfall of the Indian rupee in the recent times has led to the critical need for stock market prediction so as to safeguard the interest of the investors. This paper tries to build an efficient ARIMA model to predict the Indian stock market volatility. The publically available time series data of Indian stock market has been used for this study. The predicted time series has been compared with the actual time series, which shows roughly a deviation of 5% mean percentage error for both Nifty and Sensex on average. Various tests can be used for the validation of the predicted time series. However, in this study we have used the "ADF test and the L-jung box tests" for purpose of validation. We suggest that ARIMA approach is good enough for handling time series data, and as such can be very constructive in various real world problems like that of health sector, education, finance and other practical domains for prediction.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their useful suggestions and comments, which will help in improving this paper.

REFERENCES

[1] G. González'Rivera and T. H. Lee, "Nonlinear time series in financial forecasting," in *Encyclopedia of Complexity and Systems Science*. New York, NY, USA: Springer, 2009.

[2] P.-F. Pai and C.-S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," *Omega*, vol. 33, pp. 497–505, Dec. 2005.

[3] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, and S.-P. Guo, "Stock index forecasting based on a hybrid model," *Omega*, vol. 40, pp. 758–766, Dec. 2012.

[4] L.-Y. Wei, "A hybrid model based on ANFIS and adaptive expectation genetic algorithm to forecast TAIEX," *Econ. Model.*, vol. 33, pp. 893–899, Jul. 2013.

[5] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*. Hoboken, NJ, USA: Wiley, 2015.

[6] R. J. Hyndman. *Forecasting Overview*. Accessed: Nov. 8, 2009. [Online]. Available: https://robjhyndman.com/papers/forecastingoverview.pdf

[7] *Data Collection for Demand Forecasting*. Accessed: Jul. 2018. [Online]. Available: http://www.economicsdiscussion.net/demand-forecasting/data-collection-for-demand-forecasting/3583

[8] *Forecasting*. Accessed: Aug. 2018. [Online]. Available: https://en.wikipedia.org/wiki/Forecasting

[9] L. Kurzak, "Importance of forecasting in enterprise management," *Adv. Logistic Syst.*, vol. 6, no. 1, pp. 173–182, 2012.

[10] A. Khodabakhsh, I. Arí, M. Bakír, and A. O. Ercan, "Multivariate sensor data analysis for oil refineries and multi-mode identification of system behavior in real-time," *IEEE Access*, vol. 6, pp. 64389–64405, 2018. [Online]. Available: https://www.selfgrowth.com/articles/an-introduction-to-the-indian-stock-market date

[11] *An Introduction to Indian Stock Market*. Accessed: Jul. 2018. [Online]. Available: https://www.selfgrowth.com/articles/an-introduction-to-the-indian-stock-market

[12] *An Introduction to Indian Stock Market*. Accessed: Jul. 2018. [Online]. Available: https://www.investopedia.com/articles/stocks/09/indian-stock-market.asp

[13] A. M. Ashik and K. S. Kannan, "Time series model for stock price forecasting in India," in *Logistics, Supply Chain and Financial Predictive Analytics*. Singapore: Springer, 2019, pp. 221–231.

[14] T. Raicharoen, C. Lursinsap, and P. Sanguanbhokai, "Application of critical support vector machine to time series prediction," in *Proc. Int. Symp. ISCAS*, vol. 5, May 2003, p. 5.

[15] G. P. Zhang, "A neural network ensemble method with jittered training data for time series forecasting," *Inf. Sci.*, vol. 177, no. 23, pp. 5329–5346, 2007.

[16] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003.

[17] H. Tong, *Threshold Models in Non-Linear Time Series Analysis*, vol. 21. New York, NY, USA: Springer, 2012.
[18] T. Raicharoen, C. Lursinsap, and P. Sanguanbhokai, "Application of critical support vector machine to time series prediction," in *Proc. Int. Symp. ISCAS*, vol. 5, May 2003, p. 5.
[19] J. M. Dufour, "Introduction to time series analysis," McGill Univ., Montreal, QC, Canada, Tech. Rep., 2008, pp. 1–16.
[20] F. Lu, K. Lin, and A. Chorin, "Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems," *Commun. Appl. Math. Comput. Sci.*, vol. 11, no. 2, pp. 187–216, 2016.
[21] D. Gerbing, "Time series components," Portland State Univ., Portland, OR, USA, Tech. Rep., 2016, p. 9.
[22] R. Adhikari and R. K. Agrawal. (2013). "An introductory study on time series modeling and forecasting." [Online]. Available: https://arxiv.org/abs/1302.6613
[23] C. W. J. Granger and P. Newbold, *Forecasting Economic Time Series*. New York, NY, USA: Academic, 2014.
[24] B. Petrevska, "Predicting tourism demand by A.R.I.M.A. models," *Econ. Res.-Ekonomska Istraživanja*, vol. 30, no. 1, pp. 939–950, 2017.
[25] M. Zhang and D. Pi, "A new time series representation model and corresponding similarity measure for fast and accurate similarity detection," *IEEE Access*, vol. 5, pp. 24503–24519, 2017.
[26] S. Green, "Time series analysis of stock prices using the box-Jenkins approach," Tech. Rep., 2011.
[27] J. Pati, B. Kumar, D. Manjhi, and K. K. Shukla, "A comparison among ARIMA, BP-NN, and MOGA-NN for software clone evolution prediction," *IEEE Access*, vol. 5, pp. 11841–11851, 2017.
[28] W. W. S. Wei, "Forecasting with ARIMA processes," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Germany: Springer, 2011.
[29] Q. Zhang, F. Li, F. Long, and Q. Ling, "Vehicle emission forecasting based on wavelet transform and long short-term memory network," *IEEE Access*, vol. 6, pp. 56984–56994, 2018.
[30] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of ARIMA and artificial neural networks models for stock price prediction," *J. Appl. Math.*, vol. 2014, Mar. 2014, Art. no. 614342.
[31] S. M. Idrees, M. A. Alam, and P. Agarwal, "A study of big data and its challenges," *Int. J. Inf. Technol.*, pp. 1–6, 2018.
[32] *NIFTY 50 (NSEI)/S&P BSE SENSEX (BSESN)*. Accessed: Jul. 15, 2018. [Online]. Available: https://in.finance.yahoo.com
[33] V. Tsioumas, S. Papadimitriou, Y. Smirlis, and S. Z. Zahran, "A novel approach to forecasting the bulk freight market," *Asian J. Shipping Logistics*, vol. 33, no. 1, pp. 33–41, 2017.
[34] C. Imai, B. Armstrong, Z. Chalabi, P. Mangtani, and M. Hashizume, "Time series regression model for infectious disease and weather," *Environ. Res.*, vol. 142, pp. 319–327, Oct. 2015.
[35] E. W. Frees, "Analytics of insurance markets," *Annu. Rev. Financial Econ.*, vol. 7, pp. 253–277, Dec. 2015.
[36] T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys, "Modeling and forecasting realized volatility," *Econometrica*, vol. 71, no. 2, pp. 579–625, 2003.
[37] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1035–1042, May 2005.
[38] K. V. N. Murthy, R. Saravana, and K. V. Kumar, "Modeling and forecasting rainfall patterns of southwest monsoons in North–East India as a SARIMA process," *Meteorol. Atmos. Phys.*, vol. 130, no. 1, pp. 99–106, 2018.
[39] F.-M. Tseng, G.-H. Tzeng, H.-C. Yu, and B. J. C. Yuan, "Fuzzy ARIMA model for forecasting the foreign exchange market," *Fuzzy Sets Syst.*, vol. 118, no. 1, pp. 9–19, 2001.
[40] E. G. Jain and B. Mallick, "A study of time series models ARIMA and ETS," *Int. J. Mod. Educ. Comput. Sci.*, vol. 9, no. 4, pp. 57–63, 2017. doi: 10.5815/ijmecs.2017.04.07.
[41] W.-C. Wang, K.-W. Chau, D.-M. Xu, and X.-Y. Chen, "Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition," *Water Resour. Manage.*, vol. 29, no. 8, pp. 2655–2675, 2015.
[42] P. Agarwal, "Introduction to the stock market," *Intell. Economist*, Dec. 2017.
[43] M. F. Anaghi and Y. Norouzi, "A model for stock price forecasting based on ARMA systems," in *Proc. 2nd Int. Conf. Adv. Comput. Tools Eng. Appl. (ACTEA)*, Dec. 2012, pp. 265–268.
[44] R. G. Kavasseri and K. Seetharaman, "Day-ahead wind speed forecasting using f-ARIMA models," *Renew. Energy*, vol. 34, no. 5, pp. 1388–1393, 2009.
[45] N. F. Rahim, M. Othman, R. Sokkalingam, and E. A. Kadir, "Forecasting crude palm oil prices using fuzzy rule-based time series method," *IEEE Access*, vol. 6, pp. 32216–32224, 2018.
[46] K. K. Suresh and S. R. K. Priya, "Forecasting sugarcane yield of Tamilnadu using ARIMA models," *Sugar Tech*, vol. 13, no. 1, pp. 23–26, 2011.

SHEIKH MOHAMMAD IDREES is currently a Research Scholar with the Department of Computer Science, Jamia Hamdard, New Delhi. He has published several research articles in reputed journals and international conferences. His research interests include data mining, time series analytics, and predictive analytics and modeling.

M. AFSHAR ALAM received the Ph.D. degree in computer science from Jamia Millia Islamia, New Delhi. He has served as the Dean of School, DSW, and a Foreign Student's Advisor with Jamia Hamdard, New Delhi, where he is currently a Professor and the Head of the Department of Computer Science and Engineering. He is well known internationally for his work. He is also an Invited Professor at various universities all over the world. He is also an Expert Member of various Government committees in India. He has more than 24 years of teaching experience and has supervised more than 25 Ph.D. students in all these years. He has published over 130 research articles in well-reputed journals, besides authoring several books. His research interests include the areas of data analytics, cloud computing, big data, applied machine learning and predictive modeling, sustainable development, cyber security, and time series analysis.

PARUL AGARWAL received the Ph.D. degree in computer science from Jamia Hamdard, New Delhi. She is currently an Assistant Professor with the Department of Computer Science and Engineering, Jamia Hamdard. She has got more than 15 years of teaching experience. She has got various research papers published in well-reputed international journals indexed in Scopus, Springer, and others. Her research interests include fuzzy data mining, algorithms and artificial intelligence, big data analytics, cloud computing, and predictive analytics and modeling.