Bilal Ahmad*
Muhammad Zakria**

Comparative Study of Box-Jenkins ARIMA and KNN Algorithm for Stock Price Prediction in Pakistan

ABSTRACT

Stock market is considered a vital part of modern economic systems in the world. The fluctuation in the stock prices is of complex nature because multiple causative factors control these movements. This study was carried out to forecast the stock prices by applying two different techniques of k-nearest neighbors algorithm and Box-Jenkins ARIMA to compare their effectiveness. Three major contributing companies in Pakistan Stock Exchange were selected and the daily stock price data during the period 2014-2018 were used. In the first phase, Box-Jenkins methodology was adopted to build parsimonious ARIMA model for each series separately. The k-nearest neighbors algorithm was also performed and forecasts were calculated. Lastly, Root Mean Square Error, Mean Absolute Error and Mean Absolute Percentage Error were used for comparison purpose of both techniques. It was observed that machine learning technique of k-nearest neighbors algorithm provided more accurate results as compared to ARIMA.

Keywords: Pakistan Stock Exchange, ARIMA, Forecasting, KNN algorithm

* Allama Iqbal Open University, Islamabad. bilalahmad.imcbh9@gmail.com
** Associate Professor, Allama Iqbal Open University, Islamabad. zakria@aiou.edu.pk

Introduction

Stock markets are a trading platform where shares of publicly-held companies are sold and bought. These markets provide backbone to modern economic infrastructure. There is an inherent complexity in the movements and behaviour of the market systems (Palmer et al., 1994). This has been a dream for traders to resolve such complexities. They desire to have some forecasting method that can minimize investment risk and guarantee easy profiting. They have been looking for some strong assistance to their financial management. Human capability has not been satisfactory in analysing such data and traditional economic methods did not perform well. Stock markets' dramatic movements, chaotic behaviour and non-linearity dull the available traditional techniques. This is a big motivation for a researcher to introduce and evolve new forecasting techniques (Hafezi et al., 2015).

The prediction of stock prices is an interesting and challenging endeavour that has been considered by economists, financial analysts, statisticians and computer scientists alike. The stock market has intrigued researchers due to the uncertainty of the market as well as the potential financial gain that can come from accurate predictions (Selvin et al., 2017).

Since the 90's, a significant transition has taken place. With the advancement in the field of machine learning, the initial theories and linear strategies are making way for advanced non-linear pattern analysis, which was not possible in the past. The results of this development have been very promising and, in some cases rather astonishing, most probably redefining the way modern financial markets will trade in the days to come (Zhang et al., 2017).

Pakistan, like many other developing countries, is suffering from economic crisis due to worldwide financial uncertainty and rising inflation. Political instability is one of the major factors which cause the market to behave abnormally every now and then. Statistical methods, especially Arima models, have been the choice of most researchers for forecasting stock price movements in Pakistan but only a very few have opted machine learning algorithms. There is a need to modify the techniques that will consequently help the investors in making wise decisions and gaining maximum output. This study not only provides a simple alternative to this conventional method but also gives the comparison for better understanding.

Approaches for Stock Price Prediction

Broadly speaking, the forecasting may be performed from two perspectives: traditional statistical techniques and methods based on machine learning algorithms. Various research studies have been conducted over the years on stock price prediction with several solution techniques proposed.

Statistical methods

Renowned statistical methods for stock price prediction include exponential smoothing, ARIMA and GARCH. The ARIMA models are most commonly used in analysing and forecasting time series data. It is considered as the most efficient technique for forecasting in social sciences and is extensively used for financial problems as well. ARIMA relies on previous error terms as well as past values of the series for forecasting. These models are relatively more efficient in short-run forecasting than other models with more structural complexity. This does not imply, however, that ARIMA models are always necessarily superior to alternatives, especially if the data do not conform to the necessary assumptions – and business data often do not.

Machine learning techniques

Machine Learning techniques include unsupervised learning and supervised learning algorithms such as artificial neural networks, decision tree induction, k-nearest neighbors and genetic algorithm. KNN algorithm has been used successfully since last couple of decades for predicting financial variables (Imandoust et al., 2013). This algorithm simply states that the objects near to each other will have similar target values as well. So if you know the prediction values of certain objects, you can use them for their nearest neighbors.

Literature Review

Many of the scholars have analyzed stock markets' trends and stock price movements by statistical tools as well as soft computing algorithms. In the following paragraphs we discuss some of the key studies and their findings.
Lin et al. (2012) applied KNN and ARIMA on the stock price data. Detailed experiment and analysis were performed using both techniques. The robustness of KNN method was also studied and it was found that the KNN outperformed ARIMA and produced better results.

Paul et al. (2013) investigated stock prices of major pharmaceutical firms in Bangladesh to determine the best ARIMA model. Initially, the stationarity of data was perceived by visual inspection of historigram and correlogram, and then by Dickey-Fuller test statistic. It was found that the original series was non-stationary. Even the log transformation did not work and transformed series was still non-stationary. Then first difference of the log transformed series was taken and tests of stationarity were applied. The resultant series was found to be stationary. Several criteria such as AIC, MAPE and RMSE were used to select the best model. According to these criteria, ARIMA (2, 1, 2) was found to be the most appropriate and parsimonious model for forecasting stock prices of the pharmaceutical companies.

Alkhatib et al. (2013) carried out a study on application of KNN algorithm to stock prices for Jordanian stock exchange. The results were judged on the basis of Average Estimated Error and Root Mean Square Error. It was found that KNN algorithm was robust and stable with small error ratio, making the results rational and reasonable.

Imandoust and Bolandraftar (2013) discussed the potential of KNN for both the classification and regression problems. The weighting scheme and pros and cons of the algorithm were given. Along with many other fields, it was suggested that it can be applied to financial and business fields such as forecasting stock market, predicting the prices of a stock on the basis of performance measures and economic data, bank bankruptcies, currency exchange rates, managing and understanding financial risk etc.

Ariyo et al. (2014) presented process of building predictive model for stock price using ARIMA methodology. The published stock price data obtained from Nigeria Stock Exchange (NSE) and New York Stock Exchange (NYSE) were used for this purpose. Results of the study showed that ARIMA model has relatively a strong potential for short-term prediction and can favorably compete with other methods of stock price forecasting.

Mondal et al. (2014) used stock prices of fifty-six stocks from different sectors to study the usefulness of ARIMA for forecasting. ARIMA was selected because of its wide acceptability and simplicity. The effect on prediction accuracy was also observed based on several possible past period data taken. Akaike information criterion (AIC) was used for parameterization and comparison of the ARIMA models.

Pathirana (2015) conducted empirical study on forecasting the foreign exchange rate with KNN algorithm. The results were compared to ARIMA using mean square error, U-statistic and normalized root mean square error and it was found that KNN has an overall advantage over ARIMA.

Pradesh et al. (2018) applied ARIMA to predict Sensex Index of Bombay Stock Exchange. The training set consisted of data on closing price of 30 contributing companies of Sensex for April 2007 to March 2017. The ADF test was performed to check stationarity of each series separately as well as collectively. After fulfilment of stationarity assumption, models were specified on the basis of AIC criterion. The estimation was done and forecasts were calculated. For the measurement of accuracy, RMSE and MAE were used. The forecasted values were compared to actual values and very small residuals were observed. So it was concluded that ARIMA gives satisfactory results for stock prices.

Some researchers; (Pai and Lin, 2005), (Merh et al., 2010), (Musa and Joshua, 2020); also used combination of statistical methods and machine learning algorithms to overcome the limitations and to gain higher accuracy.

Methodology

Time series analysis is available uniquely to data that occurs sequentially over intervals of time. Data can be recorded at any time interval including daily, weekly, monthly, and yearly. The prices of stocks are time series data because the price can be observed over various time intervals. Stock prices are recorded and updated constantly during open market hours so that stock brokers and stock traders can make quick buying and selling decisions. Therefore, data for daily stock prices are easily obtainable for any stock in the market and this study prefers to consider daily prices.

Data

There are total 546 companies listed in PSX. These companies have been categorized into 35 sectors. The stock prices of 3 major companies (each from different sector) were included in the study. The details are as under:

Table 1
List of Selected Companies

Symbol	Company Name	Sector
BOP	Bank of Punjab Limited	Commercial Banks
PSO	Pakistan State Oil Company Limited	Oil and Gas Marketing
DGKC	D.G. Khan Cement Company Limited	Cement

Five years data (from 01-01-2014 to 31-12-2018) on daily stock prices of these companies were obtained from the official sources of Pakistan Stock Exchange. The Target/response variable is closing price (stock price at the end of trading day).

Box-Jenkins Methodology

The most flexible and versatile method for time series forecasting was introduced by George E. P. Box and G. M. Jenkins in 1970, known as the Box-Jenkins method. This method necessitates sufficient knowledge of identification of the model to fit the data and subsequent refinement of the model to make it satisfactory and parsimonious. This method has fewer assumptions as compared to other statistical methods which makes it widely applicable to almost any kind of time series data. Most of the series can easily be modified and transformed to fulfil the underlying assumptions. It can not only be used for univariate series but it is also capable of modelling causal relationships in which a number of independent variables are involved. This method consists of following steps.

Inspection of series for stationarity

Whenever we are dealing with time series data, one of the most important preliminary steps is to determine the stationarity. It is desired that the series under consideration should have constant mean, variance and autocovariance function for any of its sampled segments at any point in time. The three most commonly known and extensively used methods for inspection of stationarity are graphical analysis, the correlogram and the unit root test.

Model specification

In this phase the aim is to decide the reasonable but only tentative values of parameters p and q and select a suitable representative model. Initially we may possibly identify a no. of models which seem to explain and represent the behaviour and patterns of the series under consideration. We may take all these potential models to the estimation phase. The identification of a potential model involves use of the sample ACF and PACF of observed series. Besides correlogram, several other approaches are available to select a model. One of the most important is Akaike's Information Criterion (AIC). According to this criterion, the optimal model is one which has a minimum value of

$$AIC = -2\log(L) + 2k \qquad 3.1$$

where $\boldsymbol{k}$ = p + q + 1 (for a model with a constant term) and $\boldsymbol{k}$ = p + q otherwise. The term $\boldsymbol{k}$ here penalizes when we select model with too many parameters and it guarantees a parsimonious model (Akaike, 1973).

Another method to identify p and q is Schwarz Bayesian Information Criterion (BIC). It is defined as

$$BIC = -2\log(L) + k\log(n) \qquad 3.2$$

Several other criteria for model selection can also be utilized; such as root mean squared error, mean absolute error and mean absolute percentage error etc. That model is preferred which gives minimum value for these criteria (Schwarz, 1978).

Parameter estimation

After identification of the tentative model, we move forward to parameter estimation. The estimation of parameters is complicated and computer is used to carry out the necessary computation.

Diagnostic checking

Generally, more than one model fit the data sufficiently but our objective is to select a parsimonious model which represents the data adequately with minimum possible parameters. Diagnosing the suitability, adequacy and

parsimony of the estimated model is the final step of model building procedure. It can be done by plotting the residuals, correlogram of residuals and the Ljung-Box Q statistic defined by

$$Q = N(N+2)\sum_{k=1}^{k}\frac{r_k^2}{N-k} \qquad 3.3$$

If the correct model is specified and estimated, this statistic follows chi-square distribution with d.f = k-p-q. The calculated value of Q exceeding the critical values shows that residuals do not conform to white noise. In this case we need to respecify the model (Ljung and Box, 1978).

Forecasting

Forecasting the future values is an important objective of time series analysis. Once the optimal model is selected, it is used to forecast the series. The accuracy of these forecasts is also important and it indicates how much reliable are the results.

K-Nearest Neighbors Rule

The nearest neighbors rule, proposed by Fix and Hodges in 1951, has attracted many researchers. They numerically evaluated the performance of KNN rule for small sample under the assumption of normality. But Cover and Hart formally introduced the nearest neighbor algorithm in 1967. KNN is considered one of the most popular data mining techniques. It is very simple to understand but surprisingly versatile and performs incredibly well in practice. KNN can be applied to different areas such as finance, physics, computational geometry and engineering etc. It is a lazy learning technique. Being non-parametric is very useful and makes it possible to apply this algorithm to various different problems. NN and its derivatives are instance-based supervised learning algorithms. In supervised learning, the new example having unknown label/category is classified by observing those examples already observed and stored in memory. Objective is to allocate a suitable target to the new instance by learning from already labelled examples. An important aspect of this algorithm is the ability to learn quickly as compared to other methods. This technique works well even in the case where available data is limited.

KNN algorithm for time series forecasting

In KNN algorithm, ***k*** segments are selected from the training set which are closest to the most recent one. It detects histories with a similar pattern and uses them to predict the future behaviour. The predetermined parameters are ***k*** (number of nearest neighbors), **m** (embedding dimension) and ***h*** (forecasting horizon). When these parameters are determined, we can obtain the future

forecasts by utilizing the historical information from the time series as follows.

First of all the time series $(x_1, x_2, \ldots x_t, \ldots, x_h)$ is converted into vectors. For this, the equal length segments are considered as m-dimensional vectors (also known as m-histories) as under:

$$d\left(x_T^{m,h}, x_t^{m,h}\right) \quad for\ t = m, m+1, \ldots, T \tag{3.4}$$

Where ***h*** is referred to the delay parameter and **m** is the embedding dimension. These segments help in detecting behavioural configurations of the data. In this manner, the delay vector (the most recent available vector) can be written as

$$x_{t_j+1}^{m,h} = \left(x_{t_j}, x_{t_j-h}, \ldots, x_{t_j-(m-1)h}\right)\ for\ j = 1,2, \ldots, k \tag{3.5}$$

The distances, $d(x_T^{m,h}, x_t^{m,h})$, between all the m-history vectors and the delay vector are calculated for selecting the most similar vectors. Then most relevant ***k*** vectors are selected which minimize $d(x_T^{m,h}, x_t^{m,h})$. The matrix of selected k-history vectors may be given as:

$$A_t = \begin{pmatrix} x_{t_1} & x_{t_1-h} & \cdots & x_{t_1-(m-1)h} \\ x_{t_2} & x_{t_2-h} & \cdots & x_{t_2-(m-1)h} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t_k} & x_{t_k-h} & \cdots & x_{t_k-(m-1)h} \end{pmatrix} \tag{3.6}$$

The next step is to forecast the value at point $t = T + 1$. For this, many authors proposed and applied several methods such as local linear regression, simple average and weighted average (Martinez et al., 2019).

Recursive strategy

In many cases, a sequence of values is predicted instead of a single value. There are several multi-step-ahead forecasting strategies available in the literature but we have used recursive strategy. It is also known as multi-stage or iterative strategy. It is very intuitive and uses one-step ahead model repeatedly. Let we have the series $(x_1, x_2, \ldots, x_t)$ and we want to predict next ***h*** observations $(x_{t+1}, x_{t+2}, \ldots, x_{t+h})$, then

$$\hat{x}_{t+1} = \hat{f}\left(x_t, x_{t-1}, \ldots, x_{t-p+1}\right) \tag{3.7}$$

Here $x_t, x_{t-1}, \ldots, x_{t-p+1}$ are past p observations and $\hat{f}(.)$ is the model/function. Once the forecast $\hat{x}_{t+1}$ is calculated, it is then used an input in next step. That means the forecast $\hat{x}_{t+2}$ may be expressed as under:

$$\hat{x}_{t+2} = \hat{f}\left(\hat{x}_{t+1}, x_t, x_{t-1}, \ldots, x_{t-p+2}\right) \tag{3.8}$$

This procedure continues until ***h***-steps forecasts are obtained.

Measures of Forecast Accuracy

There are several measures available for comparing the forecasting performance of different techniques. Some of them are listed below

$$Mean\ Absolute\ Error = MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\hat{Y}_t - Y_t\right| \quad 3.9$$

$$Mean\ Absolute\ Percentage\ Error\ =\ MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{\hat{Y}_t - Y_t}{Y_t}\right| \quad 3.10$$

$$Root\ Mean\ Square\ Error = RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{Y}_t - Y_t\right)^2} \quad 3.11$$

Where Y_t is the original value and $\hat{Y}_t$ denotes the corresponding estimated/ forecasted value.

Results and Discussion

This section reveals the analysis of daily closing stock price data of three companies i.e. Bank of Punjab (BOP), Pakistan State Oil (PSO) and DG Khan Cement (DGKC). The data consists of 1238 observations of each company. The stock prices from January 1, 2014 to December 31, 2018 were collected from the official source of Pakistan Stock Exchange. The stock prices from January 1, 2014 to November 30, 2018 were considered as training set and the remaining stock prices from December 1, 2018 to December 31, 2018 are considered as test set. The two most popular time series techniques i.e. KNN algorithm and ARIMA were applied on the training set, parameters of each parsimonious model were estimated and tested. Goodness of fit of the models was examined and projection was also made using parsimonious model. The necessary illustration and analysis are given below:

Table 2

Descriptive Statistics of Bank of Punjab, Pakistan State Oil and DG Khan Cement Stock Prices

	BOP	PSO	DGKC
Number of observations	1238	1238	1238
Mean	10.69	365.27	144.76
Standard Error of Mean	0.08	1.46	1.31
Standard Deviation	2.71	51.58	45.97
Variance	7.33	2660.07	2113.13
Minimum	7.05	213.74	71.71
Maximum	20.37	486.05	245.37

ARIMA Models

In the first phase of analysis, ARIMA methodology was adopted to obtain the forecasts. The method was performed iteratively to select the parsimonious model. Firstly, the time plots of original series were drawn to observe the general pattern of stock prices during the five years' period. Then stationarity was checked by applying ADF test to the series. Necessary transformation was made to make data stationary. In the identification stage, ACF and PACF were used and parsimonious models were selected by using AIC and BIC criteria. The parameters were tested for their statistical significance. The diagnostic checks were performed on residuals of fitted models to determine the goodness of fit. Forecasts were obtained and accuracy measures were computed.

Visual Inspection

The historigrams of daily stock prices from 2014 to 2018 are given in Fig. 1a, 1b and 1c.

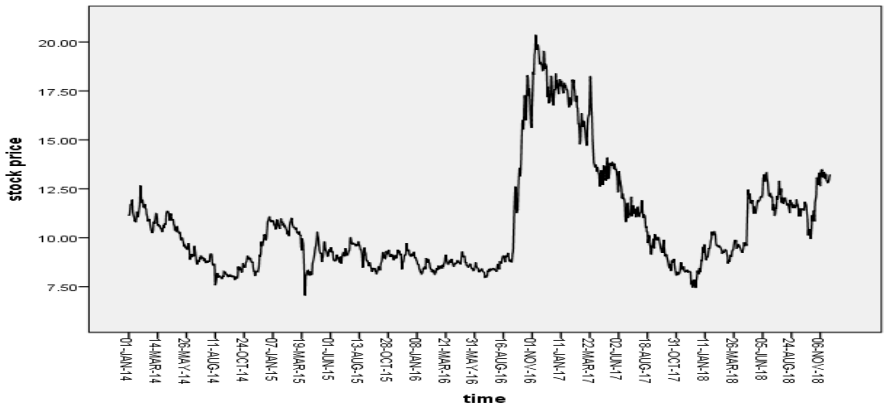

Figure 1a: Historigram of Bank of Punjab Stock Prices during 2014-2018

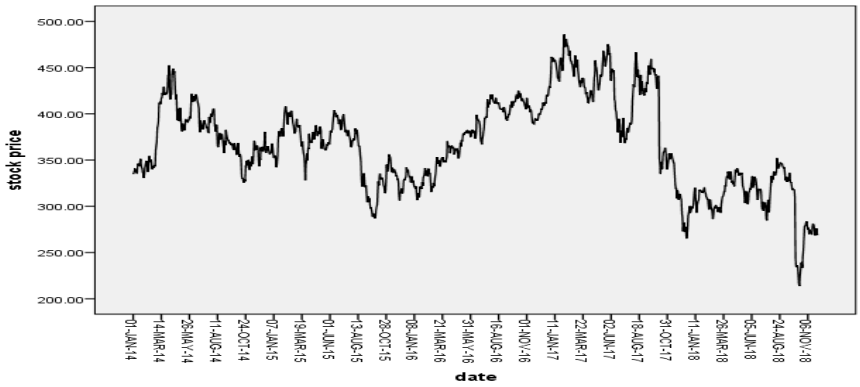

Figure 1b: Historigram of Pakistan State Oil Stock Prices During 2014-2018

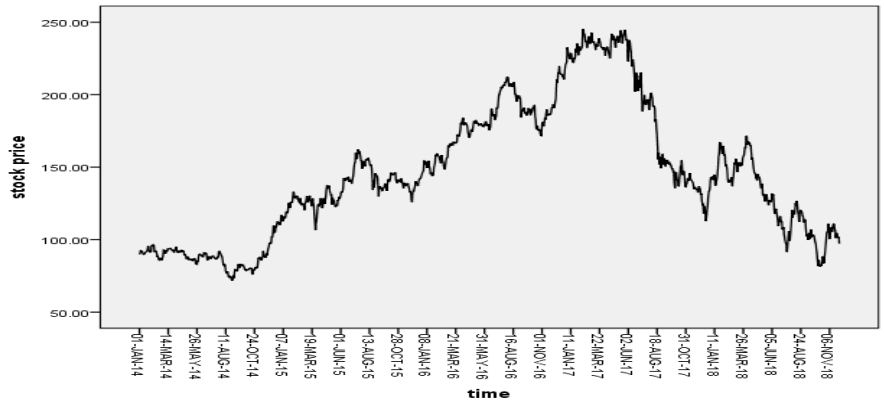

Figure 1c: Historigram of DG Khan Cement Stock Prices During 2014-2018

There are significantly visible fluctuations in all the three series. The period from November 2016 to September 2017 has the highest values while second half of 2018 showed the fast decline in the prices. The figures also indicate the mean is time-variant and data is non stationary.

Transformation and Differencing for Stationarity

Since all the three series were non-stationary, necessary transformation was performed to each series. The first series was differenced after log-transformation while first difference of the original values was sufficient for the other two series. The sequence charts of resultant transformed series are given in Fig. 2a, 2b and 2c.

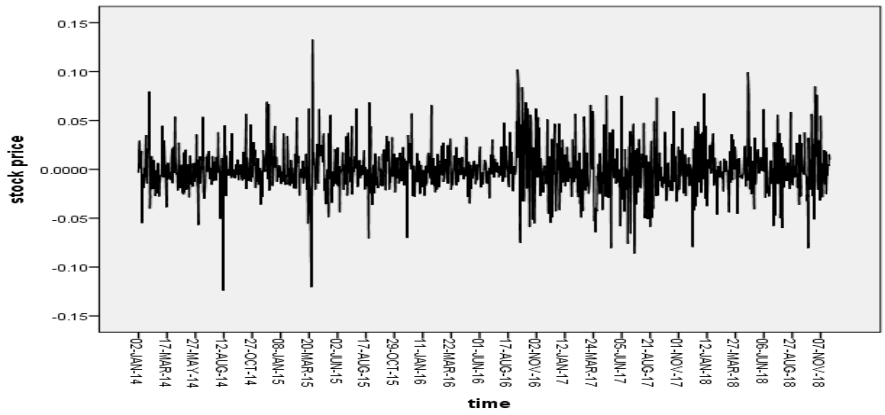

Figure 2a: Sequence Plot of Differenced Log-Series(BOP)

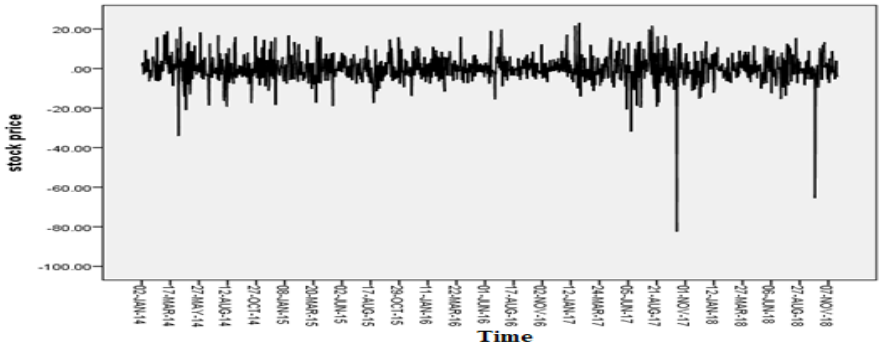

Figure 2b: Sequence Plot of Differenced Series (PSO)

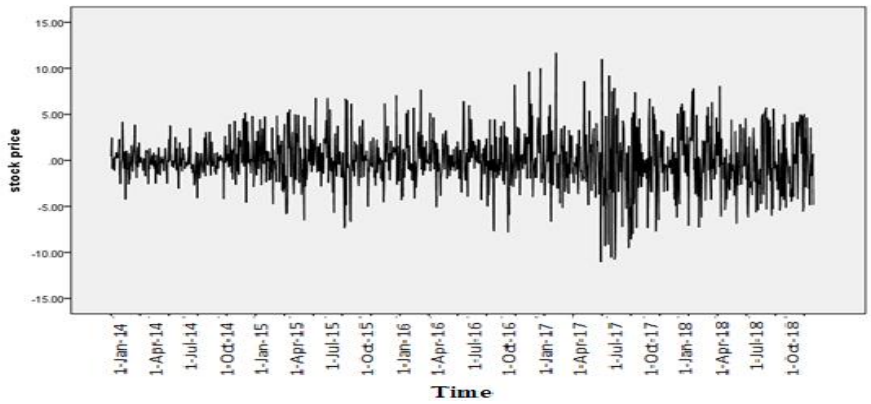

Figure 2c: Sequence Plot of Differenced Series (DGKC)

The sequence plots of transformed series, indicating that the series are centred at zero mean with no visible trend. In addition, the results of ADF test are given in Table 2 which are results were consistent with graphical evidence.

Table 3
ADF Test of Stationarity

Series	ADF Statistic	p-value
BOP	-10.731	0.009
PSO	-10.189	0.009
DGKC	-10.663	0.010

Identification by ACF and PACF

Once the series became stationary, the next step was to identify the parameters p and q of ARIMA model. ACF and PACF of the transformed series were used for that as shown in Fig. 3a, 3b and 3c.

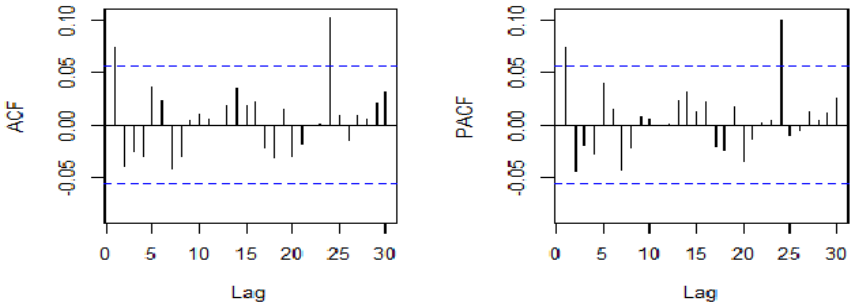

Figure 3a: ACF and PACF of Differenced Log-Stock Prices (BOP)

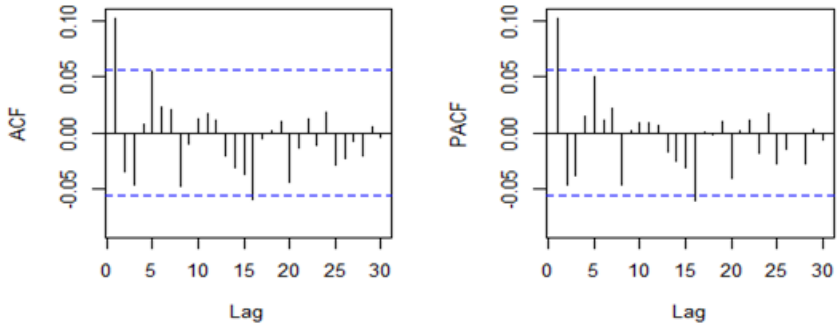

Figure 3b: ACF and PACF of Differenced- Stock Prices (PSO)

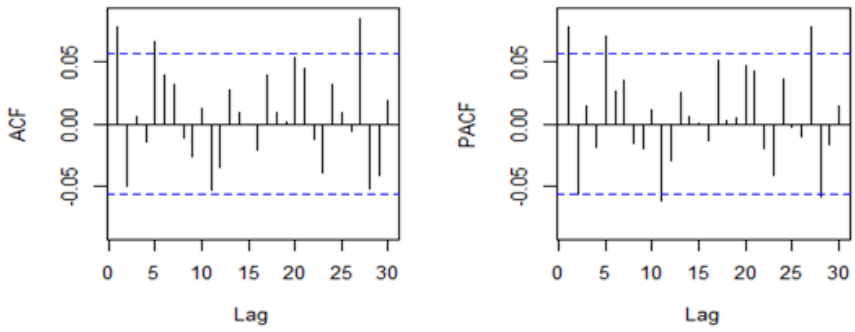

Figure 3c: ACF and PACF of Differenced Stock Prices (DGKC)

There is somewhat same type of decay in both ACF and PACF at different lags. Looking at the correlogram, the values of p and q were selected by alternate change in the AR and MA order for each series. Two information criteria i.e. AIC and BIC were used for selection of the tentative parsimonious models.

Estimation of Parameters

After identification and selection of models, the next step was the estimation and testing the parameters of these models given in Table 3.

Table 3

Estimates of Selected Models for Bank of Punjab, Pakistan State Oil and DG Khan Cement Stock Prices

Model	Term	Lag	Estimate	S.E.	t-statistic	p-value
ARIMA (0, 1, 1) for BOP	MA	Lag 1	-0.08	0.029	-2.812	0.005
ARIMA (0, 1, 1) for PSO	MA	Lag 1	-0.11	0.029	-3.851	0.000
ARIMA (1, 1, 1) for DGKC	AR	Lag 1	-0.607	0.159	-3.827	0.000
	MA	Lag 1	-0.691	0.144	-4.786	0.000

Diagnostic Checks for Estimated ARIMA Models

After estimation of the models, the next step was to perform certain diagnostic checks to see the adequacy of these models. This was done by inspection of ACF and PACF of residuals as shown in Fig. 4a, 4b and 4c.

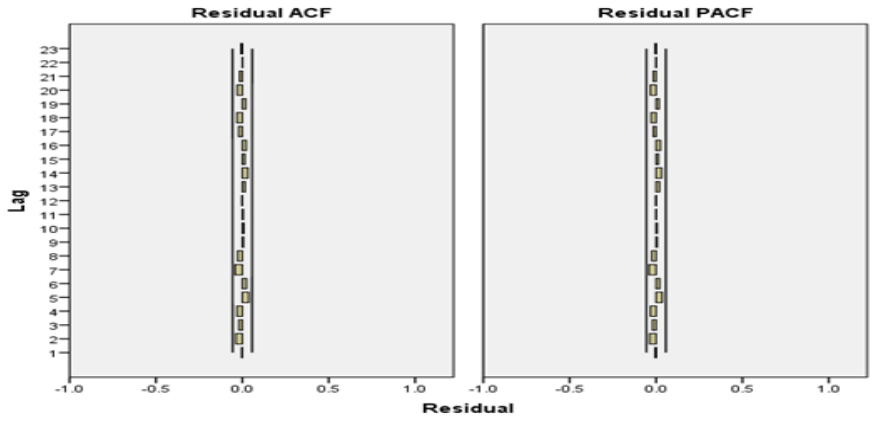

Figure 4a: ACF and PACF of Residuals for ARIMA (0, 1, 1) for BOP

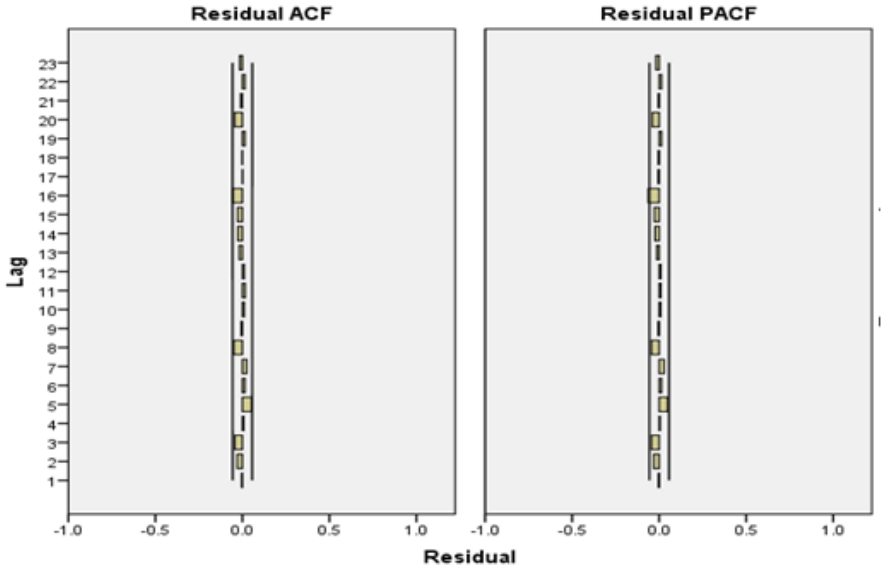

Figure 4b: ACF and PACF of Residuals for ARIMA (0, 1, 1) for PSO

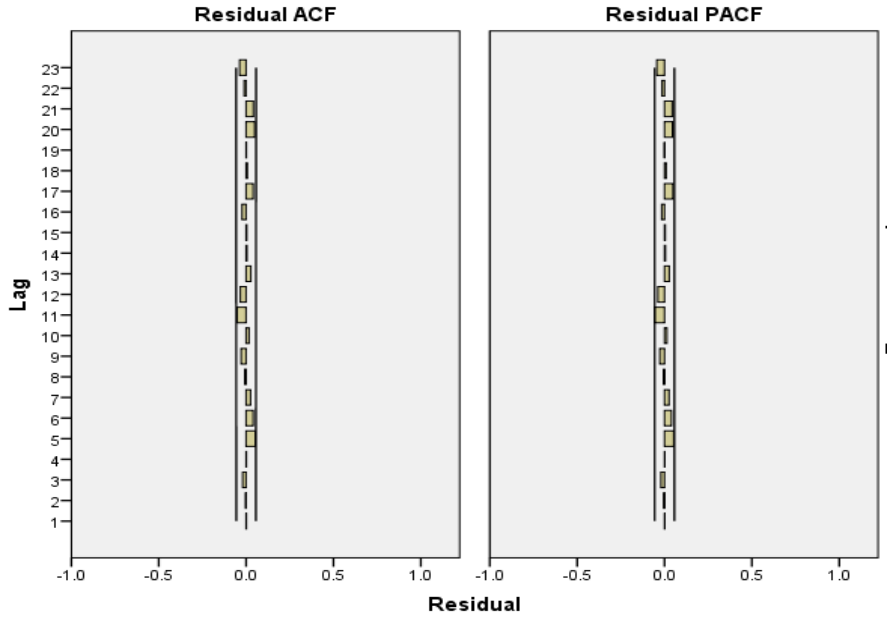

Figure 4c: ACF and PACF of Residuals for ARIMA (1, 1, 1) for DGKC

All the ACF and PACF are within confidence bounds which is consistent with the theoretical property of noise term. In addition to that, the value of Ljung-Box Q statistics also indicated that ACF and PACF of the residuals are not significant, so the proposed models are good fit.

Forecasting

The final step was to use the parsimonious ARIMA models for forecasting the stock prices of December 2018. Figure 5a, 5b and 5c present forecasts with their lower and upper confidence limits.

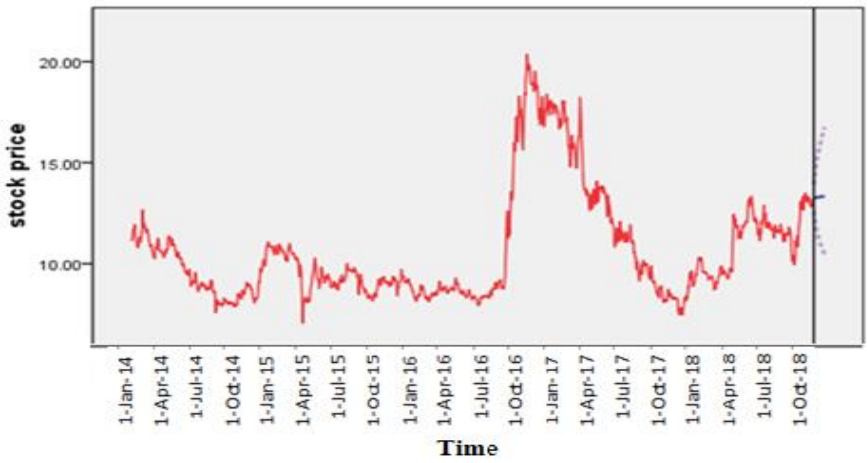

Figure 5a: Forecasts for Bank of Punjab Stock Prices by ARIMA (0, 1, 1)

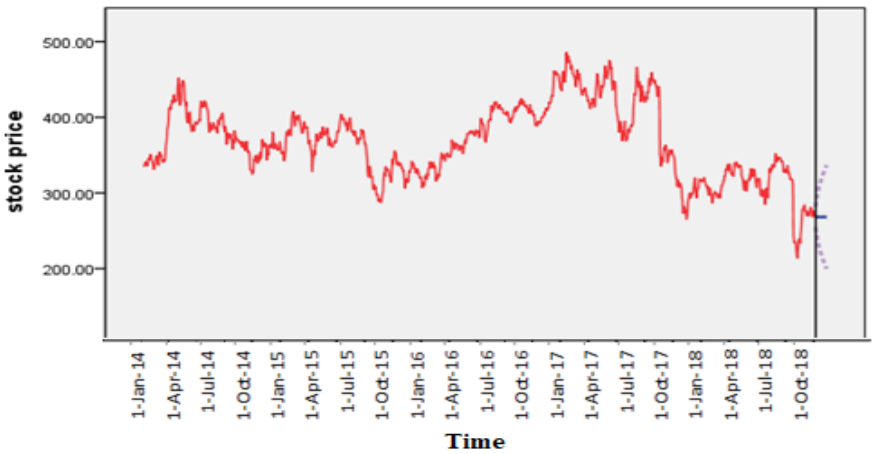

Figure 5b: Forecasts for Pakistan State Oil Stock Prices by ARIMA (0, 1, 1

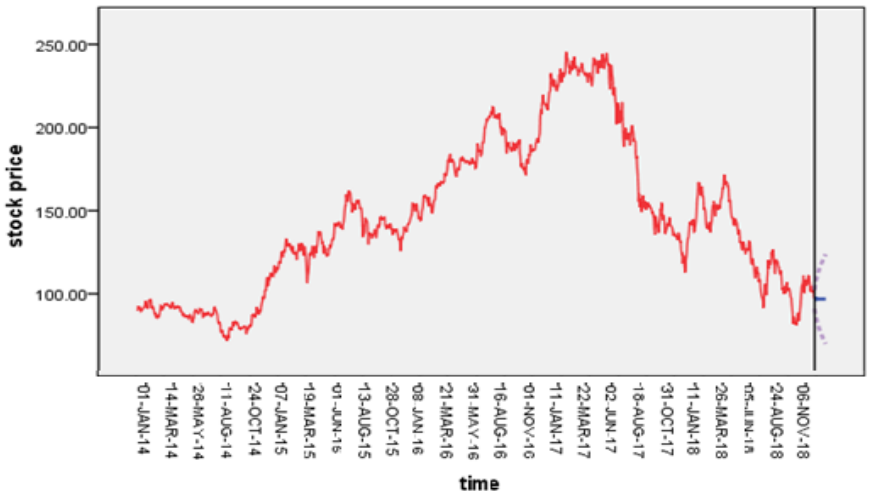

Figure 5c: Forecasts for DGKC Stock Prices by ARIMA (1, 1, 1)

KNN Algorithm

Recursive strategy was adopted for KNN algorithm to forecast the stock prices. The algorithm consisted of two major steps; determining the optimal value of ***k*** and forecasting. The optimal number of nearest numbers was decided by iterative scheme for each series using R package *tsfknn* and the distances were calculated for each instance. These distances were then indexed along with the labels of targets. Then for each forecasting horizon, the first ***k*** neighbors were selected and recorded. In the recursive strategy, forecasts were obtained one by one and the forecast of some specific step was utilized as a part of training examples for the next step. The results were combined by using simple mean of the targets of ***k*** nearest neighbors.

Comparison of ARIMA and KNN

In this section, the forecasts of stock prices along with the three forecast accuracy measures during the month of December 2018 are given in Table 4a, 4b and 4c.

Table 4a

Forecast and Accuracy Measures for Bank of Punjab Stock Prices by ARIMA and KNN Algorithm

h	Y_t	KNN Algorithm					ARIMA (0, 1, 1)				
		$\widehat{Y}_t$	Error	RMSE	MAE	MAPE	$\widehat{Y}_t$	Error	RMSE	MAE	MAPE
h=1	13.00	13.177	-0.177	0.177	0.177	0.014	13.253	-0.253	0.253	0.253	0.019
h=2	13.06	13.202	-0.142	0.161	0.16	1.225	13.258	-0.198	0.227	0.226	1.733
h=3	13.11	13.209	-0.099	0.143	0.139	1.069	13.263	-0.153	0.206	0.201	1.544
h=4	12.86	13.227	-0.367	0.222	0.196	1.516	13.268	-0.408	0.271	0.253	1.951
h=5	12.92	13.242	-0.322	0.245	0.221	1.710	13.273	-0.353	0.289	0.273	2.106
h=6	13.20	13.242	-0.042	0.224	0.191	1.478	13.277	-0.077	0.266	0.240	1.853
h=7	12.99	13.202	-0.212	0.222	0.194	1.500	13.282	-0.292	0.270	0.248	1.910
h=8	12.71	13.171	-0.461	0.264	0.228	1.765	13.287	-0.577	0.324	0.289	2.238
h=9	12.58	13.152	-0.572	0.314	0.266	2.075	13.292	-0.712	0.387	0.336	2.618
h=10	12.74	13.121	-0.381	0.321	0.277	2.166	13.297	-0.557	0.407	0.358	2.794
h=11	12.95	13.107	-0.157	0.310	0.266	1.955	13.302	-0.352	0.402	0.357	2.609
h=12	12.77	13.098	-0.328	0.311	0.272	2.120	13.306	-0.536	0.415	0.372	2.904
h=13	12.74	13.084	-0.344	0.314	0.277	2.164	13.311	-0.571	0.429	0.388	3.026
h=14	12.83	13.081	-0.251	0.310	0.275	2.149	13.316	-0.486	0.434	0.395	3.080
h=15	12.88	13.067	-0.187	0.303	0.269	2.103	13.321	-0.441	0.434	0.398	3.103
h=16	12.72	13.082	-0.362	0.307	0.264	2.150	13.326	-0.606	0.447	0.395	3.206
h=17	12.86	13.101	-0.241	0.304	0.273	2.134	13.331	-0.471	0.448	0.414	3.233
h=18	12.72	13.108	-0.388	0.309	0.28	2.185	13.335	-0.615	0.459	0.425	3.322
h=19	12.40	13.116	-0.716	0.343	0.303	2.374	13.340	-0.940	0.496	0.452	3.546
h=20	11.97	13.107	-1.137	0.420	0.344	2.730	13.345	-1.375	0.573	0.499	3.943
Overall Accuracy				***0.420***	***0.344***	***2.730***			***0.573***	***0.499***	***3.943***

h = Forecasting Horizon, Y_t = Actual Values, $\widehat{Y}_t$ = Forecasted Values

Table 4b

Forecasts and Accuracy Measures for Pakistan State Oil Stock Prices by ARIMA and KNN Algorithm

h	Y_t	KNN Algorithm					ARIMA (0, 1, 1)				
		$\widehat{Y}_t$	Error	RMSE	MAE	MAPE	$\widehat{Y}_t$	Error	RMSE	MAE	MAPE
h=1	255.27	276.314	-21.044	21.044	21.044	0.082	268.280	-13.010	13.010	13.010	0.051
h=2	248.15	277.387	-29.237	25.472	25.140	10.013	268.280	-20.130	16.948	16.570	6.604
h=3	241.78	277.844	-36.064	29.429	28.781	11.647	268.280	-26.500	20.629	19.880	8.056
h=4	230.30	278.247	-47.947	34.990	33.573	13.94	268.280	-37.980	26.073	24.405	10.165
h=5	235.19	278.947	-43.757	36.910	35.610	14.873	268.280	-33.090	27.619	26.142	10.946
h=6	239.11	279.678	-40.568	37.545	36.436	15.222	268.280	-29.170	27.884	26.647	11.155
h=7	234.28	280.453	-46.173	38.894	37.827	15.863	268.280	-34.000	28.837	27.697	11.635
h=8	229.98	281.104	-51.124	40.625	39.489	16.659	268.280	-38.300	30.183	29.023	12.262
h=9	229.76	281.586	-51.826	42.017	40.860	17.314	268.280	-38.520	31.219	30.078	12.762
h=10	236.58	281.995	-45.415	42.369	41.315	17.502	268.280	-31.700	31.268	30.240	12.826
h=11	234.69	282.163	-47.473	42.858	41.875	17.001	268.280	-33.590	31.486	30.545	12.498
h=12	234.24	282.405	-48.165	43.325	42.399	17.984	268.280	-34.040	31.706	30.836	13.092
h=13	234.00	282.559	-48.559	43.750	42.873	18.197	268.280	-34.280	31.912	31.101	13.212
h=14	232.69	282.626	-49.936	44.221	43.378	18.43	268.280	-35.590	32.188	31.421	13.361
h=15	239.75	282.662	-42.912	44.135	43.347	18.395	268.280	-28.530	31.958	31.229	13.263
h=16	244.67	282.748	-38.078	43.781	41.702	18.218	268.280	-23.610	31.501	29.939	13.037
h=17	243.29	282.889	-39.599	43.546	42.816	18.104	268.280	-24.990	31.155	30.414	12.875
h=18	237.91	283.002	-45.092	43.633	42.943	18.151	268.280	-30.370	31.112	30.411	12.869
h=19	228.15	283.087	-54.937	44.300	43.574	18.463	268.280	-40.130	31.651	30.923	13.117
h=20	225.43	283.051	-57.621	45.060	44.276	2.730	268.280	-42.850	32.303	31.519	13.412
Overall Accuracy				***45.060***	***44.276***	***18.818***			***32.303***	***31.519***	***13.412***

h = Forecasting Horizon, Y_t = Actual Values, $\widehat{Y}_t$ = Forecasted Values

Table 4c

Forecasts and Accuracy Measures for DG Khan Cement Stock Prices by ARIMA and KNN Algorithm

h	Y_t	KNN Algorithm					ARIMA (1, 1, 1)				
		$\widehat{Y}_t$	Error	RMSE	MAE	MAPE	$\widehat{Y}_t$	Error	RMSE	MAE	MAPE
h=1	92.34	96.868	-4.528	4.528	4.528	0.049	96.643	-4.303	4.303	4.303	0.047
h=2	89.79	96.467	-6.677	5.704	5.602	6.170	96.975	-7.185	5.922	5.744	6.331
h=3	87.74	96.143	-8.403	6.725	6.536	7.306	96.774	-9.034	7.112	6.841	7.653
h=4	85.48	95.816	-10.336	7.787	7.486	8.502	96.896	-11.416	8.397	7.984	9.078
h=5	89.72	95.633	-5.913	7.450	7.171	8.120	96.822	-7.102	8.155	7.808	8.846
h=6	89.78	95.620	-5.840	7.206	6.949	7.851	96.867	-7.087	7.987	7.688	8.687
h=7	89.22	95.550	-6.330	7.088	6.861	7.743	96.839	-7.619	7.935	7.678	8.666
h=8	87.66	95.431	-7.771	7.177	6.975	7.883	96.856	-9.196	8.104	7.868	8.894
h=9	87.67	95.375	-7.705	7.237	7.056	7.984	96.846	-9.176	8.230	8.013	9.069
h=10	88.54	95.419	-6.879	7.202	7.038	7.962	96.852	-8.312	8.238	8.043	9.101
h=11	87.57	95.399	-7.829	7.262	7.110	7.605	96.848	-9.278	8.338	8.155	8.813
h=12	87.91	95.450	-7.540	7.285	7.146	8.095	96.851	-8.941	8.390	8.221	9.314
h=13	87.67	95.470	-7.800	7.326	7.196	8.157	96.849	-9.179	8.453	8.294	9.403
h=14	88.39	95.467	-7.077	7.309	7.188	8.146	96.850	-8.460	8.454	8.306	9.415
h=15	88.15	95.456	-7.306	7.308	7.196	8.155	96.850	-8.700	8.470	8.332	9.445
h=16	87.93	95.437	-7.507	7.321	6.932	8.179	96.850	-8.920	8.499	8.100	9.489
h=17	87.85	95.435	-7.585	7.337	7.237	8.206	96.850	-9.000	8.529	8.406	9.534
h=18	86.07	95.439	-9.369	7.464	7.355	8.355	96.850	-10.780	8.670	8.538	9.700
h=19	81.77	95.513	-13.743	7.920	7.692	8.800	96.850	-15.080	9.120	8.882	10.16
h=20	80.15	95.545	-15.395	8.452	8.077	9.320	96.850	-16.700	9.642	9.273	10.694
Overall Accuracy				***8.452***	***8.077***	***9.32***			***9.642***	***9.273***	***10.694***

h = Forecasting Horizon, Y_t = Actual Values, $\widehat{Y}_t$ = Forecasted Values

Table 4a, 4b and 4c reveal the forecasts for the month of December 2018. The forecasting horizon is 20. Forecasts at each value of ***h*** are given along with the errors and accuracy measures; Root Mean Square Error, Mean Absolute Error and Mean Absolute Percentage Error for each forecasting horizon.

Conclusion

Having a reliable statistical forecast model for predicting stock market is a very crucial issue among investors, portfolio managers and other stakeholders. This study was carried out with the aim of forecasting stock prices of three different companies listed in PSX by two different well-known techniques. Looking at the performance evaluation criteria, it was observed that, for BOP and DGKC stock prices, KNN algorithm provided better forecasts with higher level of accuracy as compared to ARIMA. Whereas ARIMA outperformed KNN algorithm in case of PSO stock price data. The final conclusion can be made that KNN algorithm has more potential for forecasting stock prices which is consistent with the previous studies by AlKhatib et al.(2013) and Pathirana (2015). This can be further investigated that what kind of characteristics and underlying patterns of the PSO stock price data which were captured more effectively by ARIMA model.

REFERENCES

Adebayo, F. A., Sivasamy, R., & Shangodoyin, D. K. (2014). Forecasting stock market series with ARIMA Model. *Journal of Statistical and Econometric Methods*, *3*(3), 65–77.

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In *Proc. 2nd Int. Symp. on Information Theory* (pp. 267–281).

Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (KNN) algorithm. *International Journal of Business, Humanities and Technology*, *3*(3), 32–44.

Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In *Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on* (pp. 106–112). IEEE.

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

Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications*, *80*(2), 340–355.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Fix, E., & Hodges, J. L. (1951). *Nonparametric discrimination: Consistency properties*. Randolph Field, Texas, USA.

Hafezi, R., Shahrabi, J., & Hadavandi, E. (2015). A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price. *Applied Soft Computing*, *29*(1), 196–210.

Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (KNN) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*, *3*(5), 605–610.

Jonathan, D. C., & Kung-Sik, C. (2008). *Time series analysis with applications in R*. SpringerLink, Springer eBooks.

Lin, A., Shang, P., Feng, G., & Zhong, B. (2012). Application of empirical mode decomposition combined with K-nearest neighbors approach in financial time series forecasting. *Fluctuation and Noise Letters*, *11*(02), 125–134.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Mandelbrot, B., & Taylor, H. M. (1967). On the distribution of stock price differences. *Operations research*, *15*(6), 1057–1062.

Martínez, F., Frías, M. P., Pérez, M. D., & Rivera, A. J. (2019). A methodology for applying k-nearest neighbor to time series forecasting. Artificial Intelligence Review, *52*(3), 2019–2037.

Merh, N., Saxena, V. P., & Pardasani, K. R. (2010). A comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting. *Business Intelligence Journal*, *3*(2), 23–43.

Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, *4*(2), 13–16.