A Machine Learning Approach to Predicting Market Indices: A Case Study of Colombo Stock Exchange, Sri Lanka

Rathnayaka, R.M.K.T.[a], and Seneviratna, D.M.K.N.[b]

[a] *Department of Physical Sciences & Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.*
kapiar@appsc.sab.ac.lk

[b] *Department of Interdisciplinary Studies, Faculty of Engineering, University of Ruhuna, Galle, Sri Lanka.*
seneviratna@is.ruh.ac.lk

Abstract

Generally, stock prices are chaotic and show both linear and nonlinear behaviors. As a result, the ability of forecasting is notoriously problematic, and represents a major challenge with traditional time series mechanisms; most of the traditional approaches are especially weak in forecasting the future in the highly volatile and unbalanced frameworks under global and local financial depressions. This study is an attempt to develop a new hybrid forecasting approach based on back propagation neural network (BPN) to handle random walk data patterns under high volatility. The proposed methodology was successfully implemented to fulfil the daily demands of the All Share Price Index (ASPI) in Colombo Stock Exchange (CSE) Sri Lanka, from April 2009 to March 2017. The Autoregressive Integrated Moving Average (ARIMA) approach is used as a comparison mode.

Keywords: Autoregressive integrated moving average model, Geometric Brownian motion, All Share Price Index and Colombo stock exchange

1. Introduction

As a result of financial crisis in the current economy, the global growth of market prices has been changing with highly volatile fluctuations in an irregular manner (Sayanthan, 2005; Abraham, 1986). It is a common phenomenon that, when the company has obtained their capital needed, the shareholders will benefit through dividends paid by companies (Rathnayaka, 2014). As a result, stock price trend prediction is essential to make profitable investments.

Over last several decades, much effort has been devoted to investigate and develop several approaches to estimate the future phenomena. The most trustworthy approaches have been combing several methodologies to increase their strengths and mitigate the weaknesses. Simply, these techniques have created a channel to extrapolate past behaviors into the future (Asteriou,2011). Numerous types of methodologies are available in the literature to estimate the market predictions during times of high volatility, such as high order fuzzy algorithms, Markov- Fourier Grey models, Auto Regressive Moving Average methods (ARMA) and clustering, Genetic Fuzzy systems etc (Tilakaratne, 2010); (Dutta, 2007).

Among the various models, stationary, non-stationary time series models and neural networks (NNs) have become significant. Because of the poor forecasting ability with non-stationary behavioral, miscellaneous types of new methodologies have been generalized based on ARMA model under the different conditions (Cortez, 2004). Mitra et.al (2009) carried out a study to fit the optimal combination of trading rules using technical analysis with Neural Network. In 2010, Abraham et.al (1986)have conducted a similar study and forecasted Kuwait stock exchange(KSE) closing price movements, using two neural network architectures: multi-layer perceptron (MLP) neural networks, and generalized regression neural networks. Wang (2002)has also predicted stock prices using artificial neural networks. They have considered Taiwanese stock market data with fundamental indexes, technical indexes, and macroeconomic indexes. In 2014, Adebiyi et.al (2014) used ARIMA to predict short-term stock prices in New York. The data has been obtained from New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE)

According to literature, Artificial intelligence methods such as Artificial Neural Networks (ANNs), and Support Vector Machines (SVM) have been widely used in forecasting, pattern recognition and classification since 1990. A large number of studies can be seen to be have been conducted applying different frames. The ANNs have been widely used in stock market prediction during the past few decades. By using linear and nonlinear artificial neural network models, Kanas & Yannopoulos (2001) conducted an empirical study based on out-of-sample return forecast, and have successfully applied it to forecast Dow Jones (DJ), and the Financial Times (FT) indices, which have been conducted in USA and UK stock markets respectively. The results indicated that regarding the sample ANN forecast has given more accurate forecasting, than traditional linear approaches.

Because of the complications regarding financial volatility analysis with traditional time series approaches, the main purpose of this study is to develop a new hybrid forecasting approach to handle incomplete, noise, and uncertain data estimating in multidisciplinary systems. For this purpose, Artificial Neural Network (ANN) and Geometric Brownian Motion (GBM) are mainly used. Furthermore, the ARIMA model is used as a comparison mode. The

proposed methodology has been successfully implemented to forecast price indices in Colombo Stock Exchange (CSE), Sri Lanka.

1.1. Problem Definition

The study was carried out on the basis of secondary data, which were obtained from Colombo Stock Exchange official database, annual statistical reports from Central Bank of Sri Lanka, different types of background readings, and other relevant sources. Colombo Stock Exchange (CSE) is one of the modernized stock exchange markets in the South Asian region, with a fully automated trading platform. It maintains market capitalization over US$23 billion with average daily turnover of above US$18 billion.

This study mainly makes an attempt to predict future patterns in the CSE under the new proposed Geometric Brownian Motion (GBM) frame work. The paper is organised as follows. The estimated methodology is described in section II. Section III discusses the experimental findings, and concluding remarks are presented in Section IV.

2. Methods and Materials

Highly volatile and instable patterns are a common phenomenon in finance today; especially in developing stock markets in the South Asian region. Usually, the innumerable micro and macro-economic conditions with market conditions directly affect high volatility. The proposed methodology consists of two major parts; namely, ARIMA based on traditional stationary time series methods in the first part, and artificial neural network and GBM based new hybrid approach in the second. The model comparison study is conducted in the final.

2.1. Stationary time series

A stationary time series is one whose properties do not depend on the time at which the series was observed. In general, a stationary time series will have no predictable patterns long-term. Time plots will show the series to be roughly horizontal (although some cyclic behavior is possible) with constant variance, if the series is non- stationary The level data can be differentiating to calculate the constant mean, and for a non- constant variance any transformation can be applied to calculate the variance constant.

2.2. AR process

AR(p) model define as

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + Z_t, \quad (1)$$

Where {Zt} is white noise, i.e., {Zt}~WN (0, σ2), and Zt is uncorrelated with Xs for each s<t.

2.3. MA process

MA(q) Model defined as

$$X^t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad (2)$$

Where Zt~WN (0, σ2) and θ1..., θq are constants.

By using equation (01) and (02), the ARIMA model can be summarized as;

$$\emptyset(B)\nabla^{\mathrm{d}} X_t = \theta(B) a_t \quad (3)$$

The famous ARIMA sample models are summarized in Table 01.

Table 1: Arima models

ARIMA(0,0,0)	White noise
ARIMA(1,0,0)	First order Auto regressive Model
ARIMA(0,1,0)	Random Walk
ARIMA(0,0,1)	First order Moving Average Model

2.4. Geometric Brownian Motion with Ito's Lemma Approach

The Geometric Brownian motion approach is one of the popular data mining tasks, which can be used to take proper decisions in finance today; especially regarding financial data with high volatility (Ali, 2011).

Due to high volatility and unstable patterns, the traditional time series forecasting approaches cannot achieve successful predictions in both linear and non-linear domains.

Therefore, the proposed new hybrid methodology, composed of two main phases based on their linear and non-linear domains, is as follows [4] (Wang, 2002).

$$Y_t = L_t + N_t$$

L_t and N_t denote the linear autocorrelation and non-linear component of the time series pattern Y_t respectively (Rathnayaka, 2014). In the initial step, the GBM with Ito' lemma approach is used to forecast the stock market indices under the stationary and non-stationary conditions (Ho, 1998).

As the next step, the residual of the linear component is evaluated using equation (5) (Zhang, 1998).

$$e_t = Y_t - \widehat{L_t}$$

e_t denotes the residual of the GBM and $\widehat{L_t}$ presents the forecasted value of the estimated time series at time t. However, if we can see any non-linear behavioral patterns in residuals, as a next step, the ANN modeling approach is used to discover the non-linear behavioural patterns (Rathnayaka, 2014).

$$e_t = f(e_{t-1}, e_{t-2}, e_{t-3} \ldots, e_{t-n}) + \varepsilon_t$$

$$\widehat{y_t} = \widehat{L_t} + \widehat{N_t}$$

n represents the input nodes and f is the non-linear function which determined based on ANN approach.

The proposed hybrid model exploits the unique features of ARIMA, GBM and ANN in determining different patterns. Tues, it creates an additional advantage to model linear and nonlinear patterns separately by using separate models and then to combine the forecasts to improve the overall modelling performances.

2.5. Error Implementation

In the current study, MAE and MAPE are utilized to evaluate the accuracy of one-step ahead forecast. The error measures are as follows (Rathnayaka, 2014).

$$MAE = \left|P_i - \widehat{P}_i\right|$$

$$MAE = \left|\frac{P_i - \widehat{P}_i}{P_i}\right|$$

P_i and $\widehat{P}_i$ are the actual value of the original series and predicted value from the proposed hybrid model respectively. The smaller values of these error measures are considered to find the most accurate forecast result among the focused models.

3. Results and Discussion

The study was completed based on secondary data, which were obtained from Colombo Stock Exchange official database, different types of background readings, and other relevant sources. In the current study, stock prices from 1994 to 2017 are used. According to Figure 1, non-seasonal upward trend can be seen during the period from 1994 to 2017; especially, after the end of civil war in 2009, the stock prices have fluctuated with an upward (2012-2017) trend.

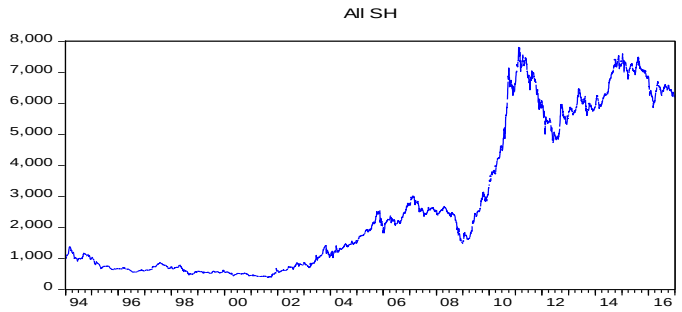

Figure 1: Stock prices during 1994-2016

In the ASPI estimation stage, first 85% of 1453 daily observations were used during the training, and the remaining 257 (about 15% of the sample) were considered as the out of sample. The visual inspection of the daily ASPI pattern in Figure 2 indicates that the data observations contain considerable noise with a significant non-linear trend with considerable volatility during the sample period of time.

To compare the unit root performances based on three different unit methods, ADF, PP and KPSS, test methodologies are applied under the 0.05 level of significance. According to Table 1, intercept (0.3335>0.05) and trend component (0.5525>0.05) of the model is not

significant under the 0.05 levels. Furthermore, the result clearly indicates that the ASPI can be categorized under the non- stationary random walk.

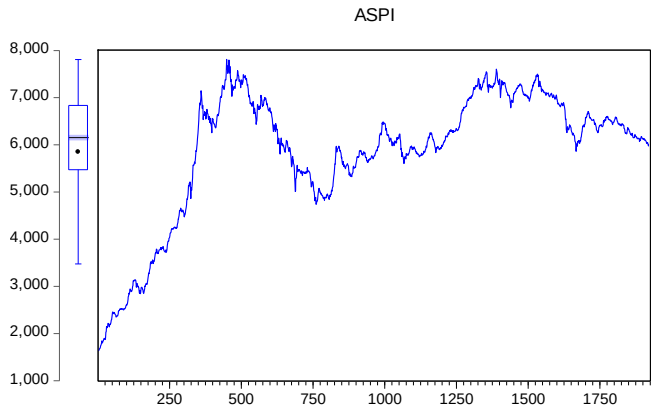

Figure 2: Time Series Plot of ASPI

According to the ACF plot in Figure 2, there are no seasonality patterns in their levels. Furthermore, ADF result (P-value=0.4189 >0.05) suggests that level data are non-stationary.

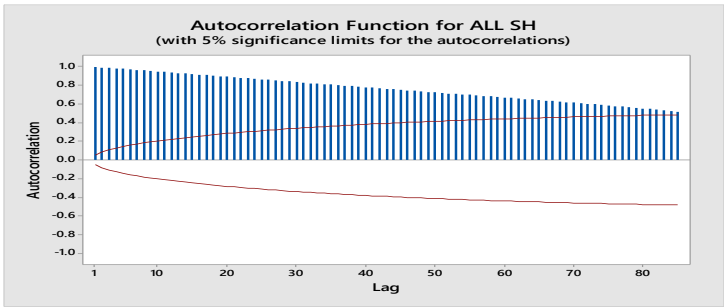

Figure 3: ACF Plot for stock price during 2010-2016

Furthermore, ADF results in Table 2 suggest that a data stationary (P-value=0.00<0.05) in their first difference conditions at 5% level of significance. It makes the analysis proceed to fit an ARIMA model.

Table 2: Unit root test result for testing the variable

Level Data		t-statistic	Prob*.
	Augmented Dickey-Fuller test statistic	-2.0347	0.5811
	Test critical values: 1% level	-3.9646	

		5% Level	-3.4130	
		10% Level	-3.1285	
	Phillips-Perron test statistic		-2.2123	0.4817
	Test critical values:	1% level	-3.9646	
		5% Level	-3.4130	
		10% Level	-3.1285	
	KPSS test statistic (LM Test)		0.6547	
	Asymptotic critical	1% level	0.2160	
	values*	5% Level	0.1460	
		10% Level	0.1190	
1st	Augmented Dickey-Fuller test statistic		-29.916	0.0000
Difference	Test critical values:	1% level	-3.9646	
		5% Level	-3.4130	
		10% Level	-3.1285	
	Phillips-Perron test statistic		-30.2394	0.0000
	Test critical values:	1% level	-3.9646	
		5% Level	-3.4130	
		10% Level	-3.1285	
	KPSS test statistic (LM Test)		0.13079	
	Asymptotic critical	1% level	0.21600	
	values*	5% Level	0.14600	
		10% Level	0.11900	

3.1. Data Modeling and Forecasting

Table 3: Arima model selection

		AR			
		0	1	2	3
MA	0		10.48213	10.48374	10.48284
			10.48537	10.49021	10.49255
			10.48333	10.48614	10.48644
	1	10.48411	10.48321	10.47883	10.47906
		10.48735	10.48968	10.48853	10.492
		10.48531	10.4856	10.48242	10.48386
	2	10.48387	10.47737	10.47891	10.48025
		10.49033	10.48706	10.49185	10.49643
		10.48626	10.48096	10.4837	10.48625

	3	10.48383	10.47847	10.48007	10.4811
		10.49352	10.4914	10.49625	10.50052
		10.48742	10.48326	10.48607	10.4883

The summary of the ARIMA models in Table 03 suggests that ARIMA (1, 1, 2) is the best model to forecast future predictions with 10.47737 AIC value, and 10.48706 BIC.

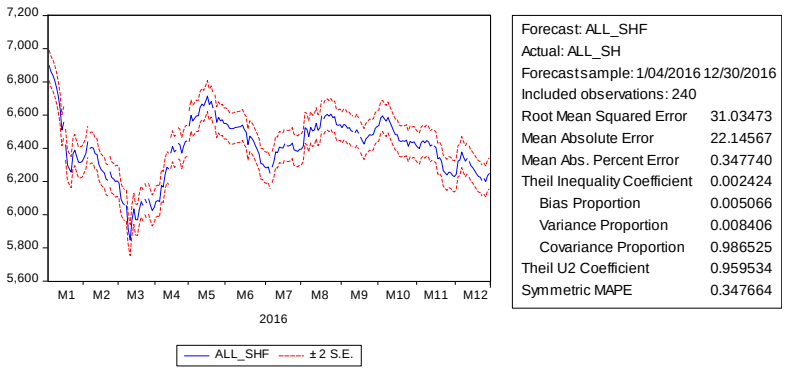

Figure 4: The graph of actual Vs. forecast

Since the models fits well, residual diagnostic has to be conducted. The Jarque-bera test results (P-value=0.00<0.05) indicate that model residuals are not normally distributed at 5% of significance level.

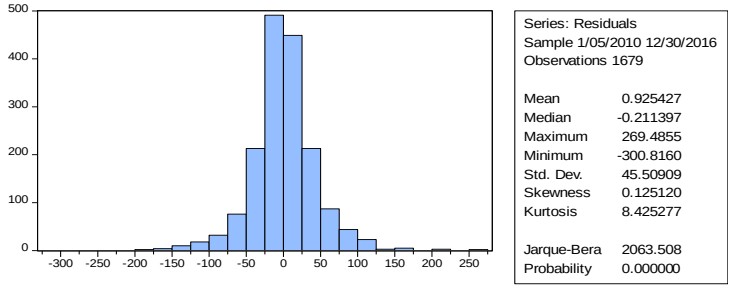

Figure 5: The graph of residuals

Furthermore, Ljung-Box test for residuals (Correlagram Q-stat) indicated (P-value=0.866>0.05) no autocorrelation between residuals. The ARCH LM test result (P-value=0.32>0.05) suggests no ARCH effect in residuals.

3.2. Geometric Brownian motion with Ito's Lemma Approach Based Hybrid Method

To find more accurate results, best two forecast horizons of 85% training sample sizes are used, and their error measures MAE and MAPE are summarized in Table 02. The corresponding forecast results, and their best error performance with minimum MAD, MSE and MAPE (%) performances with respect to the actuals are summarized in Table 4.

Table 4: Forecasting performances

Model	One-step-ahead Forecast	Actual Value	Error Accuracy Testing		
			MAD	MSE	MAPE(%)
ARIMA	5799.98	5754.31	3.75	14.06	0.053
GBM	5790.76		0.87	0.75	0.012
ANN-ARIMA	5770.89		2.29	5.26	0.038
ANN-GBM	5758.89		0.22*	0.05*	0.003*

*denotes the model with the minimum error values

Table 04 results suggest that, 85% testing sample gives the best performance with minimum MAD, MSE and MAPE (%) of 0.228, 0.051984 and 0.00324 respectively. Furthermore, results show that applying neural networks alone can improve the forecasting accuracy than single ARIMA and GBM.

4. Conclusion

Economical Time series models are widely used to develop economic relationships, especially for the nonlinear models under the stationary and non-stationary frameworks. Various research studies have been carried out to find the best forecasting methods to make long and short term predictions in the real-world context; especially in the areas of finance and investments.

The current study was carried out based on CSE daily trading data from January 2010 to May 2018, extracted and tabulated for calculations. Because of the nonlinear behavioural patterns in CSE, the proposed hybrid methodology is more suitable and appropriate to handle incomplete, noise and uncertain data in multidisciplinary systems.

References

Abraham, B. &Ledolter, J. (1986). Forecast functions implied by autoregressive integrated moving average models and other related forecast procedures. *International Statistical Review*, 54, 51-66.

Ali, S., Butt, B. Z. & Rehman, K. u. (2011). Movement Between Emerging and Developed Stock Markets: An Investigation Through Co integration Analysis. *World Applied Sciences Journal*, 12(4), 395-403.

Asteriou, D. & Hall, S. G. (2011). ARIMA Models and the Box–Jenkins Methodology. *Applied Econometrics*,2, 265–286.

Box, G. and Jenkins G., (1970).*Time series analysis*: *Forecasting and control*. San Francisco: Holden Day.

Cortez, P. R. M. &. J. (2004). Evolving Time Series Forecasting ARMA Models. *Journal of Heuristics*, 10, 415-429.

Dutta, G., Jha, P., Laha, A. K. & Mohan, N. (2007). Artificial Neural Network Models for Forecasting Stock Price Index in the Bombay Stock Exchange. *Journal of Emerging Market Finance*, 5(3), 283-295.

Rathnayaka, R. K., Jianguo, W. & Seneviratne, D. (2014).*Geometric Brownian Motion with Ito lemma Approach to evaluate market fluctuations: A case study on Colombo Stock Exchange*. Paper presented at International Conference on Behavioral, Economic, and Socio-Cultural Computing, Shanghai, China.

Sayanthan, B. (2005). *An Intelligent System to Predict the Stock Indices of the Colombo Stock Exchange*. Paper presented at Peradeniya University Research Sessions, Sri Lanka.

Tilakaratne, C. D. (2010).*A Neural Network Approach for the directional prediction of a Stock Market: An Application to the Australian All Ordinary Index*. Colombo: Department of Statistics, University of Colombo, Sri Lanka.

Wang, Y.F. (2002). "Predicting stock price using fuzzy Grey prediction system", *Expert Systems with Applications*, 22, 33-39.

Wang, Y. F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems with Applications*. 22, 33-39.

Zhang, G. P. B. E. & H. M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*. 14,35–62.