Forecasting Stock Prices using Hidden Markov Models and Support Vector Regression with Firefly Algorithm

Joshua Reno S. Cantuba[1,*], Patrick Emilio U. Nicolas[1,*], and Frumencio F. Co[1]
[1]*Mathematics Department, De La Salle University*
**Corresponding Authors: joshua_cantuba@dlsu.edu.ph, patrick_nicolas@dlsu.edu.ph*

Abstract: The autoregressive integrated moving average (ARIMA) model is extensively used in the fields of economics and finance for forecasting stock prices. Using vast amounts of historical data, it is used in the Philippine Stock Exchange (PSE) to forecast future stock price movements. In this paper, two new forecasting techniques are introduced: hidden markov models (HMM) and support vector regression with firefly algorithm (SVR-FA). Both methods are compared to ARIMA in analyzing closing stock prices of five selected Philippine companies: SM, Ayala Corporations (AC), Philippine Long Distance Telephone Company (TEL), JG Summit (JGS), and Manila Electric Railroad and Light Company (MER). All five companies present closing stock price movements that present a challenge to the proposed models as well as for ARIMA. The HMMs are trained using observable states that can emit possible movement of hidden states. In stock price forecasting, we may assume an underlying hidden movement that governs the actual increases (decreases) in stock prices for model estimation. The SVR-FA model uses the ε-sensitive loss function and a kernel function for analyzing the best regression hyperplane that relates the current day's closing stock prices with the future closing stock prices. Results show that SVRFA and HMM performed better than ARIMA in forecasting the closing stock prices ofthe selected companies with SVR-FA yielding forecasts with the lowest mean absolutepercentage errors (MAPEs) and mean absolute deviations (MADs).

Keywords: Stocks, ARIMA, HMM, MAD, MAPE, SVR-FA

1. INTRODUCTION

Stock price forecasting has been a challenge due to its random patterns and unpredictable movements. The primary motivation for studying such price movement is to determine the times or schedules for which stocks can be bought and sold at profitable positions. Several studies have been made for forecasting the trend of stock prices; although there are a lot of methods that can be used for the process, an intelligent prediction model for stock price forecasting is highly desirable (Nath& Hassan, 2005).

One of the most famous models used in predicting stock prices is the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA has been extensively used in the fields of economics and finance. In the Philippines, the Philippine Stock Exchange (PSE) forecasts future stock movement by applying the ARIMA model on vast amounts of historical data (Trading Economics, n. d.). Although the use of the said model has been proven trustworthy over for so many years, is it the best model available for forecasting Philippine stock prices? In this paper, we would like to introduce the use of two new forecasting methods namely: HMM and Support Vector Regression (SVR). It is also the objective of the paper to use ARIMA in Philippine stock prices as the benchmark for the two proposed methods.

2. METHODOLOGY

2.1 Data

Stock prices of SM, AC, TEL, JGS, and MER from January 1, 2012 to February 23, 2016 (daily data) are used for the study. The opening price, closing price, highest price, and lowest price of the day were selected as variables of interest. The last ten closing stock prices from February 10, 2016 to February 23, 2016 of the five selected companies were set aside to test the forecasting capabilities of the constructed models and how good it will be in forecasting stock prices for at most two weeks. The other observations were then split into two parts: the training set and the testing set with the ratio of 90-10%, respectively. The ARIMA models are constructed using the closing stock prices of the entire training and testing set of each company.

Various HMM and SVR – FA models are constructed using the training set of each company and the best model is selected through the testing set. HMM incorporates all four variables of interest. On the other hand, similar to ARIMA, SVR – FA incorporates closing stock prices only. The following software are used: Murphy's Hidden Markov Model MATLAB Toolbox for HMM, Chang and Lin's LIBSVM MATLAB Toolbox for SVR and the forecast and lmtest toolbox in R for the ARIMA.

2.2 Model Fitting and Forecasting

The ARIMA model fitting and the validation of the presented model through series of statistical tests in our study were facilitated by the automatic selection option of R. The primary concern is to construct a model that satisfies the assumptions of an ARIMA model and to select the most accurate one based on forecast errors.

HMMs shall be constructed on the assumption that a particular hidden movement influences the increase or decrease of closing stock prices. This involves computation of log-likelihood values of observable movement of stocks which would be handy in finding historic patterns for forecasting. The goal is to create multiple HMM models with different structures (in states) to extract log-likelihood values from the testing sets and locate a particular historic pattern with the closest log-likelihood value. The increases/decreases in the said pattern will be used to compute for the forecasts.

For the SVR, the firefly algorithm was used to carefully select the optimal input parameters C, γ, and ε. SVR-FA will refer to the firefly optimized SVR. An initial population of 25 fireflies was chosen because it makes the process faster. The parameter δ was set to 0.97 so that the optimization process can capture more of the random movements of fireflies. The task is to run the algorithm and search for SVR-FA with optimal parameters that can forecast future closing stock prices with the lowest Mean Absolute Percent Error (MAPE). All other parameters to be used will follow that of Table 1.

Table 1. Validated Parameters for SVR-FA

Parameter	Estimates	Validation	Source
α	0.1	Considered best through simulation	Arora and Singh, 2013
γ	0.01		
β	1.0	Based on usage experience	Yang, 2008
δ	[0.95, 0.97]		
Firefly population	[25, 40]		

Validation of the constructed models was done by computing the MAPE values of the forecasts on the testing set. After selecting the final models, measures of accuracy are obtained from the forecasts of the last ten observations that were set aside through the MAPE and Mean Absolute Deviation (MAD).

3. RESULTS AND DISCUSSION

3.1 ARIMA Model Fitting

The closing stock prices after the removal of the last ten observations was first subjected to the Goldfeld-Quandt Test as shown in Table 2. Of the five companies, only TEL and JGS closing stock prices showed significance with p-values less than 0.0001 indicating that their variances are significantly different at the beginning and the latter part of the time series. They were then log-transformed to eliminate non-stationarity in variance. Augmented Dickey-Fuller was then used as shown in Table 2. Results show that after first differencing, all closing stock prices are significantly mean stationary.

Table 2. P-values of the Goldfeld-Quandt Test and Augmented Dickey-Fuller Test

Company	Goldfeld-Quandt Test	Augmented Dickey-Fuller Test	
		Before Differencing	After First Differencing
SM	1.0000	0.5376	0.0100
AC	0.1680	0.2558	0.0100
TEL	<0.0001	0.7164	0.0100
MER	1.000	0.3884	0.0100
JGS	<0.0001	0.1219	0.0100

In accordance with the mean-stationarity tests, max ordinary differencing for the automatic selection was set at 1 since all five closing stock prices are stationary after first differencing and needless differencing produce less satisfactory models (Pankratz, 1983).The ACF and PACF plots after log transformation and ordinary differencing show no signs of seasonal components. As such, seasonal components and differencing are not included in the automatic selection criteria. Starting AR and MA operators for the automatic stepwise selection process are set at 0. The final constructed models for each of the five selected companies are shown in Table 3 while the validation of the constructed models is done using the Ljung-Box test as shown in Table 4. Through the said test with a maximum consideration of 250 lags, each constructed ARIMA model has shown p-values greater than 0.05 which is an indication that the said models are not significantly different from a white noise process.

Table 3. Final ARIMA Models

Company	Final Model	With Mean	Log Transformed
SM	ARIMA(5,1,0)	No	No
AC	ARIMA(1,1,1)	No	No
TEL	ARIMA(4,1,3)	No	Yes
MER	ARIMA(0,1,0)	Yes	No
JGS	ARIMA(2,1,2)	Yes	Yes

Table 4.Ljung-Box Test

Company	p-value
SM	0.3930
AC	0.0638
TEL	0.9944
MER	0.3267
JGS	0.4754

3.2 HMM Model Fitting

HMMs were used to get the log-likelihood value of the occurrence of the observation sequences in the testing set. The observation sequence in the training set was then partitioned into sizes similar to the testing set and the partition with the closest log-likelihood value was chosen as the best historic pattern that can replicate the movement of the testing set. This procedure was done in different HMM structures. The HMMs that located the best pattern and has the least parameters (states) for each company are selected as the final models. Table 5 shows the selected HMM for each of the 5 companies.Additionally, the log-likelihood values of the selected models are shown in Table 6.

Table 5. Patterns selected for the HMM procedure for the five companies

Number of Hidden States	Company			
	SM		AC	
	Pattern Selected from	MAPE	Pattern Selected from	MAPE
2	**06/04/14 to 10/28/14**	**1.4965%**	**01/03/12 to 05/24/12**	**1.7008%**
3	03/31/15 to 08/24/15	1.8191%	10/17/12 to 03/13/13	1.7202%
4	03/31/15 to 08/24/15	1.8191%	03/30/2015 to 08/20/15	1.9833%
5	10/29/14 to 03/30/15	1.7645%	03/14/13 to 08/06/13	2.0721%

Table 5 (Continuation)

Number of Hidden States	Company			
	TEL		MER	
	Pattern Selected From	MAPE	Pattern Selected From	MAPE
2	**01/03/12 to 05/24/12**	**1.9881%**	03/14/13 to 08/06/13	2.0128%
3	01/03/12 to 05/24/12	1.9881%	08/07/13 to 01/08/14	1.7255%
4	01/03/12 to 05/24/12	1.9881%	**10/28/14 to 03/27/15**	**1.2230%**
5	01/03/12 to 05/24/12	1.9881%	10/28/14 to 03/27/15	1.2230%

Table 5 (Continuation)

Number of Hidden States	Company	
	JGS	
	Pattern Selected From	MAPE
2	03/12/13 to 08/01/13	2.0164%
3	**01/03/14 to 05/26/14**	**1.8000%**
4	10/20/14 to 03/18/15	2.0121%
5	05/27/14 to 10/17/14	1.9658%

Table 6. Loglikehood of the patterns selected

Company	Log-likelihood Value			
	Pattern Selected	Date	Testing Set	Date
SM	-157.2148	06/04/14 to 10/28/14	-156.9371	09/08/2015 to 02/09/2016
AC	-155.8995	01/03/12 to 05/24/12	-156.2866	
TEL	-153.3345	01/03/12 to 05/24/12	-151.9672	
MER	-157.8273	10/28/14 to 03/27/15	-158.0489	
JGS	-155.7367	01/03/14 to 05/26/14	-155.5603	

It can be seen from Table 6 that the final HMMs has located historic patterns that have very close log-likelihood values to the corresponding testing set of each company.

3.3 SVR-FA Model Fitting

Table 7 shows the firefly-optimized parameters for SVR that are used to construct the final SVR model for each of the five selected companies. Note that all parameters are used for analyzing normalized closing stock price movement. Thus, only normalized data may be fed in the constructed SVR models for forecasting. Their corresponding MAPEs for the testing set are included as well. The parameters C^*, γ^*, and ε^* represent the optimal parameters for the SVR model.

Table 7. Firefly-optimized parameters for SVR

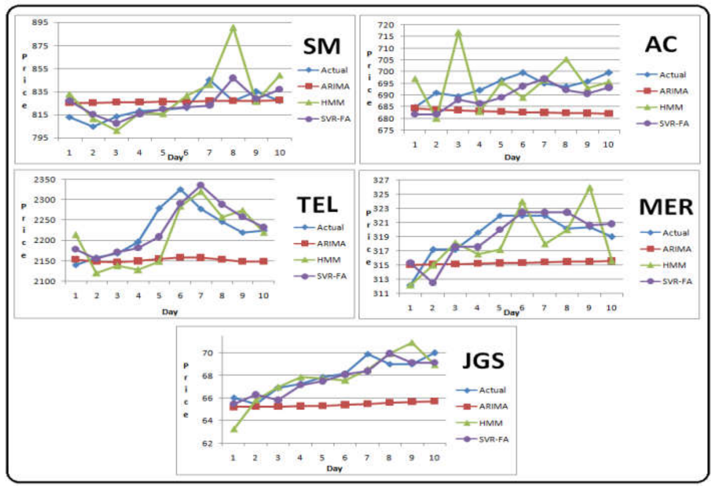
Company	C^*	γ^*	ε^*	MAPE
SM	0.6483	1.0500	0.0005	1.3896%
AC	3.3967	4.8345	0.0390	1.3733%
TEL	2.6189	0.0261	0.0002	1.7694%
JGS	94.4953	83.3669	0.0201	1.6218%
MER	23.4454	33.9205	0.0009	1.0834%

3.4 Forecasting

In this stage, the three constructed models are used in forecasting the last ten closing stock prices of the five companies that were set aside.

Forecasting future closing stock prices in R using the constructed ARIMA models is done using the "forecast" command. As for HMM, the observation sequence of the last ten observations were considered as the new testing set and the remaining data were partitioned into sizes of ten. The partition with log-likelihood value closest to the sequence in the testing set is considered as the best historic pattern. The changes in the closing stock prices of the selected pattern were then extracted and used to make forecasts. For SVR-FA, the last ten closing stock prices were normalized and each normalized current day closing stock prices was fed to the final SVR models for each company to generate forecasts. The plots in Figure 1 show the comparison between the movements of the forecasts versus the actual prices.

Figure 1. Forecasts for the actual (blue) closing stock prices of the five selected companies from ARIMA (red), HMM (green), and SVR-FA(violet) models



As shown in the previous figure, the plots portray reasonable forecast values in each of the five company's actual prices.The MAPE of the forecasts yielded by the three models on the five companies are shown in Table 9.

Table 9. MAPE of the forecasts for the last 10 observations

Company	Mean Absolute Percent Error		
	ARIMA	HMM	SVR-FA
SM	1.1178%	1.8522%	1.1061%
AC	1.5506%	1.3226%	0.6867%
TEL	3.2931%	2.2285%	1.3879%
MER	1.3831%	0.8178%	0.5461%
JGS	3.6994%	1.4548%	0.9501%

It can be seen from Table 9 that among the three models, SVR-FA's forecasts have the lowest MAPEs. Table 10 shows the mean of the last 10 observations for each company that were set aside and the corresponding MADs for the three forecasting methods which portrays the same result.

Table 10. Mean of the last 10 observations and MAD of their corresponding forecasts (in Php)

Company	Mean	Mean Absolute Deviation		
		ARIMA	HMM	SVR-FA
SM	822.40	9.17	15.25	9.13
AC	693.75	10.79	9.15	4.77
TEL	2223.60	74.45	49.6	31.18
MER	319.18	4.43	2.62	1.74
JGS	67.97	2.54	0.99	0.65

5. CONCLUSION

It is clear from the results that the SVR-FA is the best pick in forecasting the closing stock prices of the five selected companies since it has the lowest MAPEs and MADs. ARIMA gives stationary forecasts which are not good representations of actual stock price movements. HMM still did well in the study, however, there are sudden increases/decreases in the forecasts which give relatively high or low forecasts compared to the actual values. The HMMs used in the study also incorporated all four variables of interest unlike the other two methods; even though SVR-FA only incorporated one variable it was able to yield better forecasts for the five selected companies.

6. RECOMMENDATIONS

ARIMA, HMM, and SVR – FA require complicated structures. For the HMM, continuous variations and new ways to determine the number of hidden states may be considered for more accurate forecasts. For researchers that will consider discrete HMMs, they can consider trying different ways on how to transform the data into a discrete variable. Researchers may try different optimization procedures for SVR and analyze their corresponding accuracy. All models and procedures specified in this paper may also be applied to other stock prices and with longer time periods to test whether the same reasonable results can be obtained.

7. REFERENCES

Arora, S., & Singh, S. (2013). The Firefly Optimization Algorithm: Convergence Analysis and Parameter Selection. International Journal of Computer Applications IJCA, 69(3), 48-52.doi:10.5120/11826-7528 kernel-based learning methods. Cambridge: Cambridge University Press.

Hassan, M., &Nath, B. (2005). Stock Market Forecasting using Hidden Markov model: A new Approach. 5th International Conference on Intelligent Systems Design and Applicationsc(ISDA'05). Retrieved March 30, 2016.

Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., &Hussain, O. K. (2013). Support Vector Rgression with Chaos-based Firefly Algorithm for Stock Market Price Forecasting. Applied Soft Computing, 13(2), 947-958. doi:10.1016/j.asoc.2012.09.024

Magadia, J. C. (2012, May 28-29). Risk Measures for the PSE Index Using Hidden Markov Models.

Philippines Stock Market (PSEi) Forecast 2016-2020. (n.d.). Retrieved June 12, 2016, from http://www.tradingeconomics.com /philippines/stock-market/forecast

Pankratz, A. (1983). Forecasting with Univariate Box-Jenkins Models (Concepts and Cases).New York: John Wiley & Sons.

Yang, X. (2008). Nature-inspired Metaheuristic AlgorithmsFrame, U.K.: LuniverPress.