Stock Price Forecast Using ARIMA and Support Vector Regression

Abstract

The prediction of the stock price has been challenging work for many people. The correct estimation of stock price gives enormous profits and return with low risks to investors. The advent of machine learning and time-series models make it possible. This essay focus on the forecast of Hang Seng Index (HSI) close price in Hong Kong stock market using support vector regression (SVR) and Autoregressive Integrated Moving Average model (ARIMA), then make the evaluation by calculating RMSE and MAPE as well as the directional accuracy of stock price changes. The results show that the SVR algorithm outperforms ARIMA. The impact of different time horizon is also considered.

Department of Economics

Student ID: 1836386

Word count: 4924

Key words: Machine learning, Support vector machine for regression, ARIMA, Time series analysis, Hang Seng index, Closing price, MAPE, RMSE.

Table of Contents

1 Introduction

Predicting the stock market has been challenging work for many investors, financial analysts and investment managers. This problem has attracted lots of experts and elites from different fields such as economics, mathematics, data science and computer science. There are numerous uncertainties in the stock market, so it is difficult to forecast future movement using only simple regression methods to observe past stock data trends.

Traditionally, there are two well-known methods to make stock predictions. The first method is fundamental analysis. It usually tries to find the stocks' intrinsic value by analysing a company's financial statements and some common financial ratios such as P/E ratio and P/B ratio. The second is technical analysis. It predicts the stock price by observing the past pattern of the stock. In most of the investment banks, the analysts commonly think that fundamental analysis should be a significant method to predict the stocks' movement. With the boom of Big Data and AI, however, it also shows an astonishing predicting result using machine learning techniques to predict the stock movement by the past price information. It is an extremely efficient approach to hedge financial risks for investors (Chen, Leung and Daouk, 2003). Hence, the multitude of commercial companies started to apply this methodology. For example, Castle Ridge Asset Management leveraged Geno-Synthetic Algorithms to detect behavioural patterns in large volumes of market data and claimed 32 percent average annual return. According to Bloomberg, also, BlackRock proposed a new set of 12 ETFs managed by machine-learning (Evans and Ponczek, 2017).

In this paper, we discuss the prediction of stock market indexes using Machine Learning and time-series techniques. The essay uses machine learning model, known as the Support Vector Machine for Regression (SVR) and time series model, called Autoregressive Integrated Moving Average (ARIMA), to forecast the closing price in Hang Seng Index (HSI) in the short, medium and long term. The goal is to provide a comparison between two models, studying which gives better performance.

This essay includes five sections. Section 2 introduces the background and hypothesis of financial markets as well as some prior research for the stock market using similar machine learning algorithms and time series models. In section 3, the collected data and the principle of SVR and ARIMA in the essay are introduced. Section 4 explains the steps and results of solving stock prediction problem using SVR and ARIMA. The last section discusses some comments on the two methods and recommendations for future work. In the essay, some crucial graphs and tables are shown in the context, while other figures and tables can be seen in the appendix.

2 Background and Literature Reviews

Efficient Market Hypothesis (EMH) is an essential component in financial markets. The idea behind EMH is random walk theory, claiming that stock prices follow a random walk, or that the price today is independent of price yesterday (Pinches, 1970). Also, EMH claims that share prices reflect all information (Fama, 1970). It is impossible to obtain an abnormal return, and the market cannot be beaten. There are three kinds of efficient markets: weak, semi-strong and strong forms (Brealey, Myers and Allen, 2014). Weak form market efficiency states that today's price reflects the information on historical price. Investors can use fundamental analysis to obtain abnormal gains instead of technical analysis. The semi-strong form of market efficiency suggests that today's share price reflects all publicly available information, which explains that either fundamental analysis or technical analysis fails to be leveraged for earning abnormal returns. Strong market efficiency shows that today's stock price reflects all private and public information, which means that investors cannot obtain abnormal returns using either inside information or technical analysis. Therefore, the prices of all securities should reflect their intrinsic values at any time based on EMH (Fama, 1970).

However, Quiggin (2020) argued that, in real life, Bitcoin could be the best case that EMH cannot hold. The reason is that unlike gold, tobacco or dollars, Bitcoin does not have any source of value. It cannot generate any benefits. However, investors hope Bitcoin's value, similar to that of other financial assets, appreciates and does not depreciate. This endless

appreciation is the kind of bubble that EMH believes cannot occur. Besides, the technical analysis should not give a precise forecast if EMH holds. However, many papers, to be introduced in the next part, show that many statistical models and algorithms achieve highly accurate predictions on stock prices. Hence, the development of these statistical models is deemed to be the biggest threat to EMH (Patel et al., 2015).

Machine learning is a mature technique that uses these statistical models to make forecasts with big data. Machine learning is an application of some models and algorithms, making data more useful (Raschka, 2015). It can predict future data by learning past data trends automatically, without explicitly programming. There are mainly two types of machine learning methods: supervised and unsupervised. Supervised learning use models, including numbers of features, to train the labelled input data, called training data, to predict the future output data, known as testing data (Raschka, 2015). Supervised means the actual results of testing data are known so that the predicted results can be compared with the actual results to see the forecasted performance. Unsupervised learning means the testing data is not labelled, and the prediction, therefore, needs to be done without knowing the actual results. In other words, unsupervised learning aims to find the structure of data without any guidance of known outcomes. The forecast in the stock market is supervised learning in this research. Supervised learning can solve two kinds of problem, which are classification and regression. Classification is a problem that identifies the categories of the testing data in which the result is already known using some features. In contrast, the regression problem gives continuous outcomes or solutions for prediction. In this study, the forecast of the stock market is considered as a regression problem since we estimate the actual closing price of HSI.

Several papers and research studies have shown that neural networks (NN), an algorithm in machine learning used for classification, performed very efficiently for stock predictions. Saad, Prokhorov and Wunsch (1996) found that it is possible to reduce false alarm rate using time-delay neural networks (TDNN) and recurrent neural networks (RNN) with two training-data methods, including conjugate gradient and extended Kalman filter. Moreover,

the probabilistic neural network (PNN) was found to offer more accurate predicted results when forecasting the direction of Taiwan Stock Exchange (TSE) Index, compared with generalized methods of moments (GMM) Kalman filter model and random walk model (Chen, Leung and Daouk, 2003).

However, NN includes some limitations to learn the data trend due to a large amount of noise and complex dimensionality. It is challenging for the artificial neural network (ANN) algorithms to predict noisy data. Thus, in Kim and Han's research paper, they proposed a method combining genetic algorithms with ANNs to predict the stock market index, which successfully optimised the dimension of the feature space and boosted the forecasting performance. There are other methods in addition to ANN algorithms that are used to predict the stock. Alkhatib, Najadatthat, Hmeidi and Shatnawi (2013) used k-nearest neighbour (k-NN), a supervised algorithm in machine learning for solving classification problems, to predict the stock market direction of five major companies in Jordan stock market exchange, and they mainly use root mean square deviation (RMSD) to evaluate the model. The best-predicted performance of k-NN with strong robustness and lowest RMSD was founded when k = 5. Additionally, Zaidi and Ofori-Abebrese (2016) leveraged logistic regression with four features—i.e., lower price, higher price, open price, and oil price—to obtain 80 percent accuracy as classifying KSA stock market direction.

Most of the methods mentioned above follow the empirical risk minimisation (ERM) principle, aiming to minimise the empirical errors. The models that apply the ERM principle have a precise leaning result as the sample size is large, while it causes the overfitting problem easily as the sample size is small (Kim, 2003). Therefore, recently, a distinct algorithm called support vector machine (SVM) is used by Kim (2003). SVM is a supervised-learning algorithm with the property of dimensionality control of the decision function and the usage of kernel functions (Vapnik, 1999; Cristianini and Taylor, 2000). It seeks hyperplane that is used to determine the support vector to operate a classification of the data (Vapnik, 1999). Figure 1 shows the linearly separable case. The solid straight line is a hyperplane in 2-dimensional space. The round points and stars represent classified

samples. The support vectors lie along the dash lines. The distance between L_1 and L_2 is called margin. SVM seeks maximal margin hyperplane to make a separation between classes. In other words, SVM wants to maximise the distance of the closest points from the hyperplane (Vapnik, 1999; Cristianini and Taylor, 2000). In a non-linear separable case, a

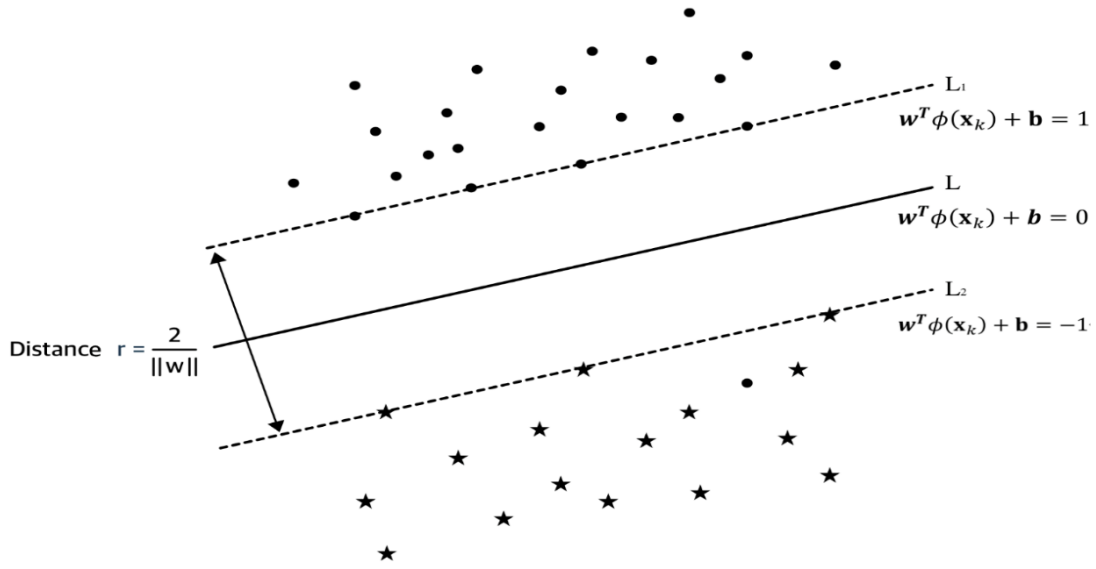

Figure 1. Linear plane separable hyperplane of SVM case

kernel function is needed to be introduced, and it maps input variables from Euclidean space into a space with higher dimensions, called Hilbert space (Tay and Cao, 2001). In the new space, SVM can be obtained by calculating its new maximal margin hyperplane.

Kim (2003) stated that SVM applied the structural risk minimisation (SRM) principle, which tries to solve the overfitting problem by a trade-off between hypothesis space complexity and the success of fitting training data. In Kim's paper, he used the polynomial and Gaussian radial kernel function to solve the problem of non-linear separable financial data. Kim made a comparison with SVM, back-propagation neural network (BPN) and case-based reasoning (CBR), the outcome revealed that SVM excessed BPN and CBR. Thus, he claimed that SVM could be an alternative method for time-series analysis. In addition, Zhang (2009) successfully used the time-series model, ARIMA (1,1,0), to fit the SSE composite index, and obtained forecast of the short-term trend with relatively low error's average value, at 0.04. Hence, Zhang (2009) believed that ARIMA is also an efficient method for prediction in the stock market. Therefore, based on the conclusions above, since this essay treats stock prediction problem as a regression problem, support vector machine

for regression (SVR) and ARIMA are applied.

Several papers focus on the prediction of the stock price using sentiment analysis by machine learning. According to Mbadi (2018), she believes that sentiment analysis is significant in the prediction of the stock market. Sentiment analysis can be extracted by natural language process, statistical methods and machine learning techniques. Mbadi (2018) used enhanced Naïve Bayes classifier, which is a classification algorithm in machine learning, to extract information from Facebook and Twitter and then classify them as positive, negative and neutral information. After that, she used a hybrid model to predict the stock prices of Apple and Amazon and finally concluded that using sentiment of investors as a feature to create a model can obtain a better performance than the predicted model without a sentiment indicator. In addition to the sentiment analysis using machine learning, Schumaker and Chen (2009) used machine learning approaches, such as Bag of Words and Named Entities, to do textual mining in the financial news articles. Then they determined the SVM derivative to forecast a discrete stock price in the S&P 500 index twenty minutes after the report was released. Finally, a very low MSE and 57.1% directional accuracy proved that their prediction was quite effective. However, due to the complexity of data mining and information scoring, and because the essay is mainly to make an attempt to the machine learning technique in the stock forecast, the essay excludes the sentimental analysis.

3 Experimental Data and Forecasting Methodologies

3.1 Experimental Data

The Hang Seng Index in Hong Kong Stock Exchange will be selected as a primary research object for the stock prediction in this essay. The collection of 20 years (01/12/1999-3/20/2020) historical time series dataset from HSI was gathered from Yahoo finance website.

Before using machine learning techniques, many financial economists prefer to use the econometric model with technical indicator variables to analyse the stock market, such as EMH model (Hsu et al., 2016). Technical indicators leverage the past price and volume to determine the future price trends of the stock. In addition, recent researches usually

combinate many indicators for the improvement of profitability, such as Simple Moving Average (SMA) and Relative Strength Index (RSI) (Henrique, Sobreiro and Kimura, 2018), Commodity Channel Index (CCI) (Kim, 2003), as well as Momentum (Cortes et al., 1995). Thus, the data of these indicators are collected as the independent variables in the essay. The formulas of these indicators are given as equation (1) and (2):

$$SMA = \frac{1}{n}\sum_{t=1}^{n} P_t \tag{1}$$

$$RSI = 100 - \frac{100}{1+EMA(n)_{up}/EMA(n)_{down}} \tag{2}$$

$$CCI = \frac{M_t - SMA_t}{0.015 \times D_t} \; where \; M_t = \frac{H_t + L_t + P_t}{3} \; and \; D_t = \frac{1}{n}\sum_{i=1}^{n} |M_{t-i+1} - SMA_t| \tag{3}$$

$$Momentum = P_t - P_{t-x} \tag{4}$$

where P_t is the closing price at time t; L_t is the low price at time t; H_t is the high price at time t; SMA_t is the simple moving average of t days; P_{t-x} is the closing price $t-x$ number of days ago; $EMA(n)_{up}$ is an upward change calculating by n-period Exponential Moving Average (EMA); $EMA(n)_{down}$ is a downward change calculating by n-period EMA.

The aim is to forecast the closing price of HSI in the short term (52 weeks), medium term (104 weeks) and long term (319 weeks) by SVR and ARIMA, and then make the comparison between them using RMSE and MAPE (Henrique, Sobreiro and Kimura, 2018). Also, comparing the directional accuracy of the change in forecasted stock price with each period. The formulas of RMSE and MAPE are given as equation (5) and (6) below:

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{t=1}^{n} (\hat{c}_t - c_t)^2} \tag{5}$$

$$MAPE = \frac{1}{n} \times \sum_{t=1}^{n} \left| \frac{c_t - \hat{c}_t}{c_t} \right| \times 100\% \tag{6}$$

where $\hat{c}_t$ is predicted closing prices at time t, c_t is actual closing prices at time t.

This research splits the dataset into two groups, 70% of training data and 30% of testing data. Figure 2 below shows the actual closing price of HSI from 1999 to 2013 (706 weeks). The entire analysis is performed using R statistical software.

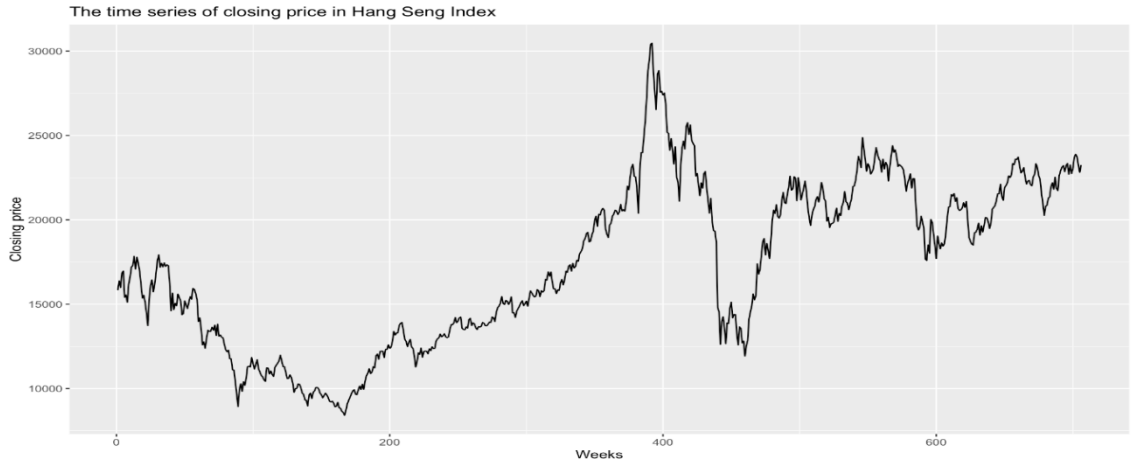

Figure 2. The time series of closing price in Hang Seng Index

3.2 Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a popular technique for prediction of time series data, which have two properties: linear regression and the univariate regression. ARIMA combines autoregressive (AR) model and Moving Average (MA) model together with taking the difference of time series data for stationarity. Therefore, for creating the ARIMA model, we need three parameters, which are 'p', 'd' and 'q'. 'P' and 'q' values in ARIMA represents the order of AR and MA terms. 'd' value specifies the number of times of differencing the series for making it stationary. The ARIMA model is given as equation (7):

$$\phi(L)(1-L)^d X_t = \mu + \Theta(L)\varepsilon_t \qquad (7)$$

where $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, is p-th order Autoregressive (AR) polynomial coefficient; $\Theta(L) = 1 - \Theta_1 L - \cdots - \Theta_q L^q$, is q-th order Moving Average (MA) polynomial coefficient.

There are six steps of prediction using ARIMA. First, cleaning and pre-processing the data. Second, testing the stationarity of time series. Third, differencing if the time series is not stationary. Next, confirming the order of AR and MA. Then, fitting the series and operating

a significance test for regression coefficients. Finally, comparing and selecting the model.

3.3 Support Vector Machine for Regression (SVR)

SVR is an extended application of SVM for solving regression problem with the property of structural risk minimisation. SVR follows a similar principle with SVM that was mentioned in Section 2. However, differently with SVM, SVR wants to minimise the distance of the farthest support vectors from hyperplane as figure 3 shows (Vapnik, 1999). It can achieve high accuracy of prediction under the condition of finite sample size and low complexity of the model (Vapnik, 1999). SVR obtains the optimal decision function by convex quadratic programming (QP) so that it can gain a globally optimal solution.

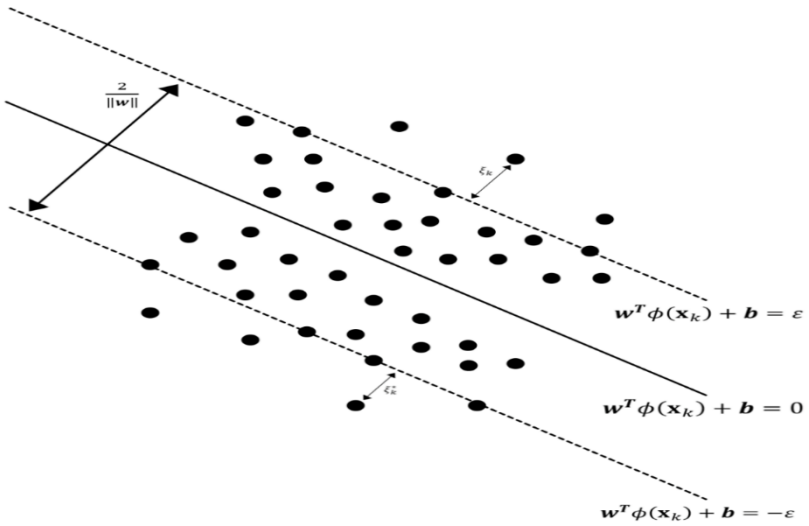

Figure 3. The diagram of SVR case

The non-linear support vector regression problem can be described as follows:

Given training set $t = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (\mathbb{R}^n \times \mathbb{R})^l$, where $x_k \in \mathbb{R}^n, y_k \in \mathbb{R}, k = 1, \dots, l$. An optimal decision function $g(x) = (\boldsymbol{w^T}\phi(\mathbf{x}_k) + \mathbf{b})$ can be found in $\mathbb{R}^n$ based on training set, so that $\widehat{y_s} = g(x_s)$ is predicted value y_s corresponding input variables x_s, where $\phi(\mathbf{x}_k)$ represents mapping the variable $\mathbf{x}_k$ from Euclidean space into Hilbert space. This problem can be solved by creating the convex QP as equation (8):

$$\min_{\boldsymbol{w},\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{k=1}^{l}(\xi_k + \xi_k^*), \tag{8}$$

$$\text{subject to } \begin{cases} (\boldsymbol{w^T}\phi(\mathbf{x}_k) + \mathbf{b}) - y_k \leq \varepsilon + \xi_k, k = 1, \dots, l \\ y_k - (\boldsymbol{w^T}\phi(\mathbf{x}_k) + \mathbf{b}) \leq \varepsilon + \xi_k^*, k = 1, \dots, l \\ \xi_k^{(*)} \geq 0, k = 1, \dots, l \end{cases}$$

where C is the tuning parameter that controls the trade-off between bias and variance, if C is a small value, the model has a low bias but high variance and vice versa; $\xi_k \; and \; \xi_k^*$ are slack variables showed in figure 3, meaning the upper training error and the lower training error, respectively (James, Witten, Hastie and Tibshirani, 2013); ε is margin of tolerance; $\xi_k^{(*)} = (\xi_1, \xi_1{}^*, \dots, \xi_l, \xi_l{}^*)^T$. The problem above can be solved by creating a Lagrange dual function (Hastie, Friedman and Tibshirani, 2017):

$$L(\boldsymbol{w}, \boldsymbol{b}, \xi^{(*)}, \eta^{(*)}, \alpha^{(*)}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{k=1}^{l}(\xi_k - \xi_k^*) - \sum_{k=1}^{l}(\eta_k\xi_k - \eta_k^*\xi_k^*) - \sum_{k=1}^{l}\alpha_k(\varepsilon + \xi_k + y_k -$$

$$(\boldsymbol{w^T}\phi(\mathbf{x}_k) + \mathbf{b}) - \sum_{k=1}^{l}\alpha_k{}^*(\varepsilon + \xi_k - y_k + (\boldsymbol{w^T}\phi(\mathbf{x}_k) + \mathbf{b}) \quad (9)$$

where $\alpha^{(*)} = (\alpha_1, \alpha_1{}^*, \dots, \alpha_l, \alpha_l{}^*)^T$, $\eta^{(*)} = (\eta_1, \eta_1{}^*, \dots, \eta_l, \eta_l{}^*)^T$ are Lagrange multiplier vectors.

Therefore, the dual problem of the primal problem is:

$$\min_{\alpha^{(*)} \in \mathbb{R}^l} \frac{1}{2}\sum_{k,j=1}^{l}(\alpha_k^* - \alpha_k)(\alpha_j^* - \alpha_j)K(\mathbf{x}_k, \mathbf{x}_j) + \varepsilon\sum_{k=1}^{l}(\alpha_k^* + \alpha_k) - \sum_{k=1}^{l} y_k(\alpha_k^* - \alpha_k) \quad (10)$$

$$s.t. \sum_{k=1}^{l}(\alpha_k^* - \alpha_k) = 0, 0 \leq \alpha_k^{(*)} \leq C, k = 1, \dots, l$$

where $K(\mathbf{x}_k, \mathbf{x}_j) = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}^T)$ is kernel function. The application of kernel function makes the inner product of vectors in Hilbert space can be calculated efficiently and hence avoids "Dimensional Disaster".

Therefore, by solving the minimisation problem, the coefficient of support vector $\bar{\alpha}^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \dots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$ can be obtained, then we can create an optimal decision function for regression as equation (11):

$$g(x) = \sum_{k=1}^{l}(\bar{\alpha}_k^* - \bar{\alpha}_k)K(\mathbf{x}, \mathbf{x}_k) + \boldsymbol{b} \quad (11)$$

The kernel functions of three regression models are shown below:

$$Linear\ SVR: K(x_k, x_j) = x_k^T x_j \tag{12}$$

$$Polynomial\ SVR: K(x_k, x_j) = (1 + x_k^T x_j)^d, where\ d\ is\ in\ the\ set\ of\ \{2,3, \dots\}. \tag{13}$$

$$Gaussian\ radial\ SVR: K(x_k, x_j) = \exp\left(-\gamma \left\| x_k - x_j \right\|^2\right)\ \text{if}\ \gamma > 0 \tag{14}$$

For polynomial SVR, d is the degree, indicating that it fits the SVR into a higher-dimensional space with the polynomials of degree d. For Gaussian radial SVR, γ is a positive constant. In this study, we use SVR to estimate the real values instead of classification.

4 The Prediction of HSI Closing Price

4.1 Fitting data using ARIMA

As mentioned before, there are some steps using the ARIMA model for prediction. The first step is data pre-processing. The weekly data of HSI is imported from the local repository into R. We transform closing price into logarithmic format since the price is based on the return that is depended by percentage, and then transform back to the original closing price after forecasting. Table of data of the first 30 weeks can be found in the appendix. Then, the data are processed by dropping the missing value. Also, the data can be divided into 70% training data (706 weeks) and 30% testing data (319 weeks) for prediction of the closing price of HSI since it is how the most of research paper divide data.

The differencing data	P values of Augmented Dickey-Fuller Test
The original training data	0.253
The second-order differencing training data	<0.01
Alternative Hypothesis: Stationary	

Table 1. The results of ADF test

Secondly, determining the order of differencing and testing the stationarity. For the data that are not stationary, differencing is needed. The unit root test, such as DF and ADF test, can be leveraged to test the stationarity of the series. According to figure 1 shown above, it can

be seen that the closing price series in HSI has an upward trend but no seasonal effect. In addition, as table 1 shows, large p-value, at 0.253, means that the null hypothesis where the

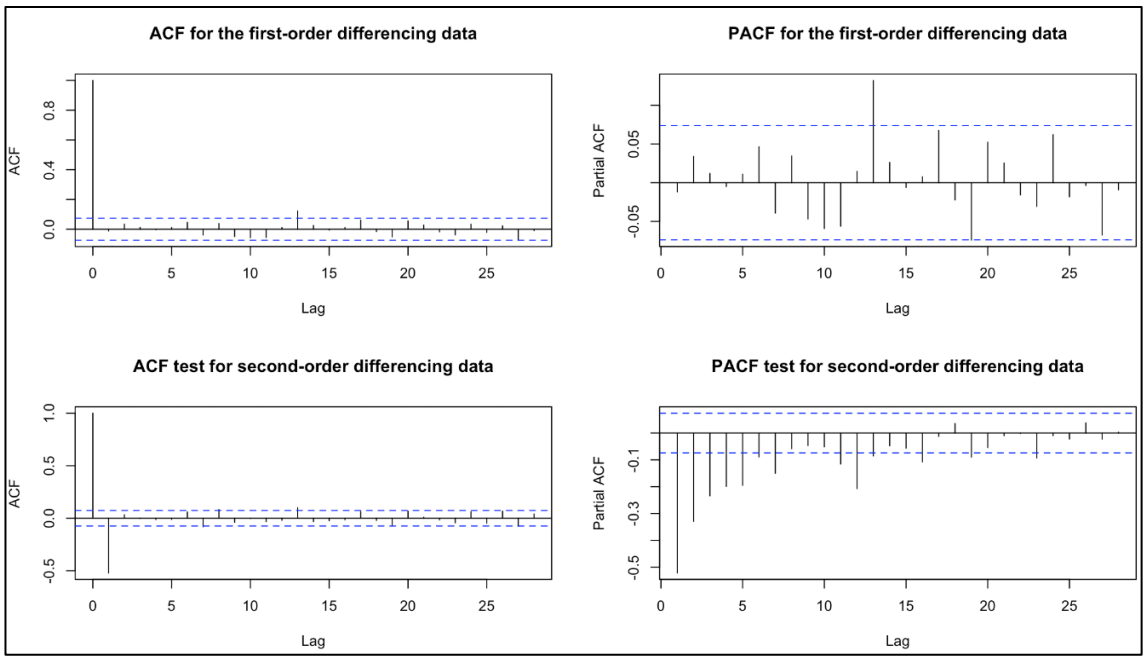

Figure 4. ACF and PACF diagrams for the first and second order differencing data

series is not stationary fails to be rejected at any three of significance levels. For the data with second-order differencing, the p-values are far less than 0.01, indicating that the series is stationary.

Then, determination of ARIMA (p, d, q). Figure 4 shows the ACF and PACF diagrams of the series after taking the first and second-order differences. By the figure, it is better to determine the AR and MA terms using the series after taking the second-order difference. From the figure, the first-order MA model can be chosen since there is a first-order truncation in the ACF diagram. Also, we choose the fifth-order or seventh-order AR model based on the PACF diagram. Hence, ARIMA (5,2,1) and ARIMA (7,2,1) are chosen. However, by Akaike's information criterion (AIC) and Bayesian information criterion (BIC), only ARIMA (5,2,1) should be determined as the final predicted model. AIC and BIC provide a trade-off between the complexity of the model and the goodness-of-fit (James, Witten, Hastie and Tibshirani, 2013). The smaller the two values, the better fit on the training data. In this experiment, ARIMA (5,2,1) gives smaller values of both AIC and BIC than ARIMA (7,2,1), so ARIMA (5,2,1) should be determined for the prediction on the

logarithmic closing price of HSI.

Therefore, we use ARIMA (5,2,1) to fit the data. Figure 5 below is a time series diagram after fitting the training data using ARIMA (5,2,1). By that figure, the red fitting curve almost covers the black logarithmic closing price curve, indicating that the model fits the data successfully. In addition, table 2 above illustrates the p-value of the Ljung-Box test. The p-value of the LBQ test is 0.089, indicating that the null hypothesis that the residuals are random cannot be rejected at the 18 degrees of freedom and 5% significance level. Hence the residuals can be treated as white noise series.

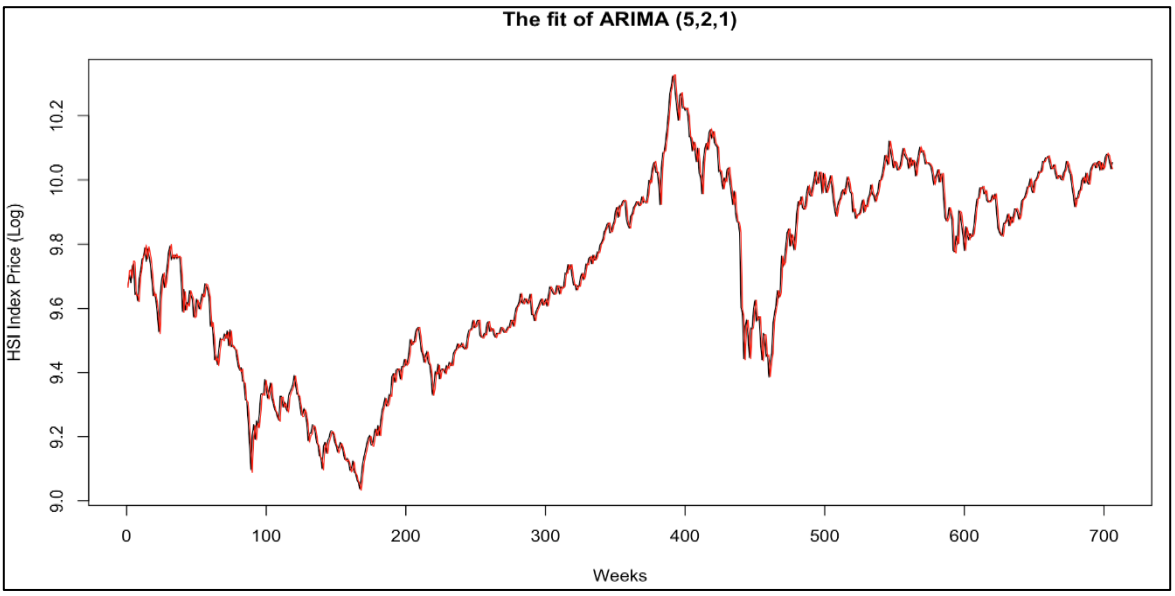

Figure 5. The fit of ARIMA (5,2,1)

Model Fit		Ljung-Box Q Test		
AIC	BIC	χ-squared statistics	DF	Sig.
-2783.41	-2751.51	26.48	18	0.089

Table 2. LBQ Test for fitting residuals

4.2 Fitting data using SVR

Afterwards, the essay fits the data using SVR. For SVR model, the closing price of HSI is the output variable, four technical indicators, including SMA, RSI, CCI and Momentum, are used as input variables. SMA indicator can eliminate the effects that stock prices fluctuate; RSI indicator incorporates some principles of judgment, such as overbought,

oversold and divergence, in which the theory and practice are extremely suitable for short-term investment in the stock market; Momentum indicator reveals the change in stock prices by observing the speed of stock price fluctuations; CCI indicator measures the change in stock prices from the statistical mean. Hence, by these indicators, the SVR can recognise the trends more precisely. In this essay, similar to most papers, a calculation period of P = 6 is fixed for all indicators. In addition, all input and output variables are scaled with zero mean and one standard deviation, and the standardised actual and forecasted closing price are transformed back to the original value after prediction.

The linear SVR, polynomial SVR and Gaussian radial SVR are applied. Due to the use of kernel function, the most challenging part for SVR is to choose appropriate parameters, such as C, σ and the number of degrees (d). This research applies the 10-fold Cross-Validation (CV) and repeats the CV for five times to obtain the optimal tuning parameters. The principle of 10-fold CV is to let training set be further divided into ten subsets, including nine subsets for modelling and 1 for predicting, and iterate the process for ten times. Each subset can be the predicting set only once during the iteration, then the average RMSE of the iterated results can be computed. The 10-fold CV can be operated automatically in R. The results can be seen in table 3.

Linear SVR	**Polynomial SVR**		**Gaussian Radial SVR**	
C	C	d	C	ε
0.25	1	3	45	0.01

Table 3. The optimal parameters chosen by 10-fold CV with 5 times repeat

After obtaining the optimal parameters for three kinds of SVR, we use them to fit the

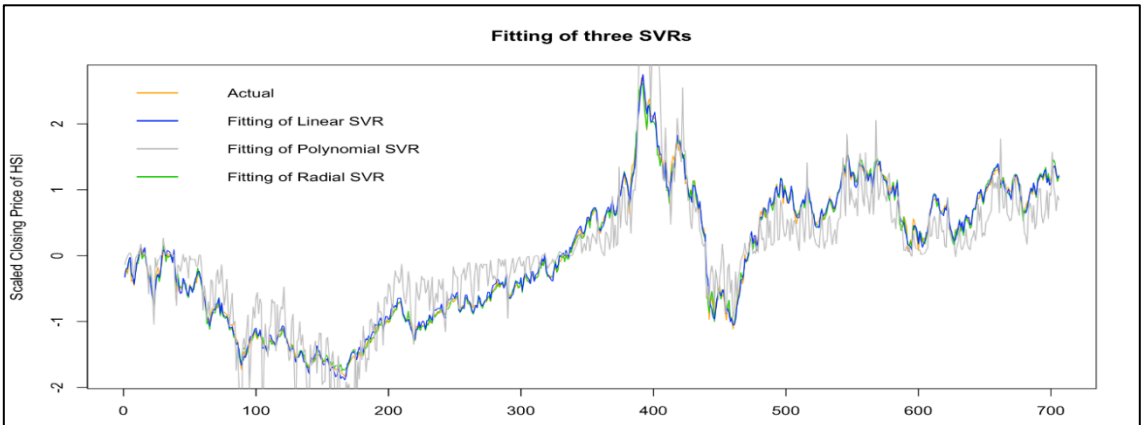

Figure 6. Fitting scaled data by three SVRs

training data. By figure 6, it can be seen that the linear and radial SVRs fit the training data precisely. Hence, these two SVRs will be used to do prediction and comparison with the ARIMA model.

4.3 Forecast using ARIMA (5,2,1) and SVRs

According to the models introduced above, we predict the closing price of HSI in the short, medium and long term using ARIMA (5,2,1), linear SVR and Gaussian radial SVR, obtaining the tables between corresponding forecasted prices and actual prices as table 4 and 5 show. The predicted diagrams of each model can be seen in the appendix.

Periods	**RMSE**	**MAPE**
52 weeks	869.26	3.16%
104 weeks	1706.19	5.40%
319 weeks	2479.40	7.90%

Table 4. The RMSE and MAPE of Prediction of ARIMA (5,2,1)

Periods	**RMSE**		**MAPE**	
	Linear SVR (C = 0.25)	**Radial SVR (C = 45, σ = 0.1)**	**Linear SVR (C = 0.25)**	**Radial SVR (C = 45, σ = 0.01)**
52 weeks	437.35	445.36	1.54%	1.59%
104 weeks	550.40	537.14	1.69%	1.67%
319 weeks	408.40	421.86	1.14%	1.24%

Table 5. The RMSE and MAPE of Prediction of Linear SVR and Radial SVR

Table 4 shows the RMSE and MAPE of the forecasted HSI closing price using ARIMA (5,2,1). MAPE represents the mean absolute percentage error, which tells the average deviation from the actual value. Thus, the smaller the values of MAPE are, the more precise fitting with the actual value. RMSE represents the root mean square error, the smaller the values of RMSE are, the higher the accuracy of data description. The RMSE and MAPE of prediction for 52 weeks are dramatically lower than the ones for 104 weeks and 319 weeks.

It concludes that the performance of ARIMA (5,2,1) is better in the short-term forecast.

Table 5 shows the RMSE and MAPE of predicted HSI closing price using linear SVR and Gaussian radial SVR. Differing with ARIMA (5,2,1), the RMSE and MAPE of linear and radial SVR models are low in the long-term prediction, indicating that they are better at forecasting the long-term closing price in HSI. However, if comparing ARIMA with the SVRs, the predicted errors of ARIMA are larger than the SVRs in all periods. Thus, from the perspective of forecast errors, the SVRs always give a better performance than ARIMA in the short, medium and long-term prediction.

Periods	**Accuracy**		
	ARIMA (5,2,1)	**Linear SVR (C = 0.25)**	**Radial SVR (C = 45, σ = 0.01)**
52 weeks	52.94%	78.85%	82.69%
104 weeks	54.37%	79.81%	81.73%
319 weeks	54.72%	82.76%	81.19%

Table 6. The accuracy of the direction of closing price in HSI

In addition to measuring the errors, the directional accuracy of the stock prices is also vital. In order to measure the accuracy, we mark “Up” and “Down” for each actual and forecasted prices. Then creating a confusion matrix and calculating the accuracy of the directions that are predicted correctly with each model. Table 6 above shows the directional accuracy of the closing prices in HSI. From the table, the directional accuracy of the closing price in HSI forecasted by ARIMA (5,2,1) is in the range between 52.94% and 54.72% during the three periods, which is not very precise but predictable since it is somewhat more precise than guessing. Nevertheless, the predicted accuracy of linear and radial SVR reaches the range between 78% and 83%, indicating that SVR models capture at least 78% of trends of the closing price in HSI. It also can be concluded that SVR models outperform the ARIMA model dramatically.

In conclusion, from measuring either the errors or directional accuracy of the stock price, support vector regression produces better results than ARIMA model, which means that

machine learning techniques provide an alternative method to time-series analysis for the forecast in the stock market. This is because SVR applies structural risk minimisation, which is more advantageous for processing the non-linear and heteroscedastic data. Additionally, ARIMA is a univariate model, while SVR is a multivariate model so that the prediction of SVR does not only depend on the past data of the dependent variable.

5 Discussion and Future Works

The motivation of the essay is to make a comparison between ARIMA and SVR for stock prediction and to provide a model with a better result. However, there are some limitations to the essay. Initially, for prediction of ARIMA (5,2,1), the AIC and BIC are used for model selection. But there are other criteria, such as Cp-criterion and adjusted R-squared value, might give different results for model selection.

Secondly, the essay made a preliminary forecast on the closing price of HSI using SVR. However, there are still lots of work to be done in the optimisation of tuning parameters. For instance, instead of using the CV to determine tuning parameters, other methods, such as genetic algorithm (GA) and particle swarm optimisation (PSO), are worthy of consideration.

Another limitation is that the essay only focuses on the weekly stock market data. It might be the essential reason for the decrease in predicted accuracy. The daily, monthly data should be collected as well. In addition, this essay selects 20 years of closing price data in HSI for studying. Further research can focus on more data, such as 30 years or 50 years, which may result in more accurate forecasting results.

Feature or independent variable selection plays an essential role in either machine learning models or regression models. Another limitation of the essay is that the model is simple. In this essay, only four technical indicators are used as independent variables in SVR. Thus, the further work, in addition to the technical indicators, can be based on other factors, such as macroeconomic factors including GDP growth, change in inflation, working capital ratio and LIBOR or even oil prices.

Besides, the research mainly focuses on the closing price in the Hang Seng Index, which is the relatively mature stock market in the world. The model might not have the same results as forecasting from other stock markets since the factor that affects the stock market in each country may be different because of the various governments, policies and the rules of law. Therefore, the same techniques can be applied in distinct stock markets or companies.

6 References

Alkhatib, K., Najadat, H., Hmeidi, I. and Shatnawi, M. (2013). Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm. *International Journal of Business, Humanities and Technology*, 3(3), pp.32-44.

Brealey, R., Myers, S. and Allen, F. (2014). *Principles of corporate finance*. 11th ed. New York, NY: McGraw-Hill Education, pp.325-326.

Chen, A., Leung, M. and Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research*, 30(6), pp.901-923.

Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, 20, 273-297.

Evans, R. and Ponczek, S., 2017. *Blackrock Is Launching New Active Etfs That Are Built By Bots*. [online] Bloomberg. Available at: <https://www.bloomberg.com/news/articles/2017-11-20/blackrock-king-of-indexing-ditches-passive-for-bot-built-etfs> [Accessed 16 March 2020].

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2), p.383.

Hastie, T., Friedman, J. and Tibshirani, R., 2017. *The Elements Of Statistical Learning*. 2nd ed. New York: Springer, pp.423-431.

Hsu, M., Lessmann, S., Sung, M., Ma, T. and Johnson, J., 2016. Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, 61, pp.215-234.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. 8th ed. New York Springer, pp.337-365.

Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2), pp.307-319.

Kim, K. and Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, 19(2), pp.125-132.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), pp.259-268.

Pinches, G. (1970). The Random Walk Hypothesis and Technical Analysis. *Financial Analysts Journal*, 26(2), pp.104-110.

Quiggin, J. (2020). *The Bitcoin Bubble and a Bad Hypothesis*. [online] The National Interest. Available at: https://nationalinterest.org/commentary/the-bitcoin-bubble-bad-hypothesis-8353

Raschka, S. (2015). *Python machine learning*. Birmingham: Packt Publishing, pp.4-7.

Saad, E., Prokhorov, D. and Wunsch, D. (1996). Advanced Neural Network Training Methods for Low False Alarm Stock Trend Prediction. *IEEE*, 4, pp.2021-2026.

Schumaker, R. and Chen, H., 2006. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, 27.

Mbadi, S., 2018. Predicting Stock Market Movement Using Naïve Bayes Model for Sentiment Analysis. Research Gate.

Castle Ridge Asset Management. 2020. *Meet W.A.L.L.A.C.E. - Castle Ridge Asset Management*. [online] Available at: <https://castleridgemgt.com/meet-w-a-l-l-a-c-e/> [Accessed 16 March 2020].

Tay, F. and Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, 29(4), pp.309-317.

Wei, W. (2007). *Time series analysis*. 2nd ed. Boston: Pearson Addison Wesley, pp.33-86.

Zaidi, M. and Ofori-Abebrese, A. (2016). Forecasting Stock Market Trends by Logistic Regression and Neural Networks: Evidence from KSA Stock Market. *Euro-Asian Journal of Economics and Finance*, 4(2), pp.50-58.

Zhang, J., Shan, R. and Su, W. (2009). Applying Time Series Analysis Builds Stock Price Forecast Model. *Modern Applied Science*, 3(5).

7 Appendix

Table a1: The first 30 weeks of original data from Yahoo finance

Date	Close	SMA	RSI	Weeks	MTM	CCI
1999/12/3	15840.41	14540.82	82.79	1	2977.33	97.93
1999/12/10	16380.21	15061.37	85.87	2	3123.26	112.02
1999/12/17	15986.35	15457.38	74.25	3	2376.08	74.69
1999/12/24	16833.28	15897.98	80.91	4	2643.61	98.8
1999/12/30	16962.1	16212.81	81.77	5	1889	116.72
2000/1/7	15405.63	16234.66	49.45	6	131.1	-62.57
2000/1/14	15542.23	16184.97	51.47	7	-298.18	-101.56

2000/1/21	15108.41	15973	44.67	8	-1271.8	-116.85
2000/1/28	16185.94	16006.27	60.31	9	199.59	-29.2
2000/2/18	16599.16	16130.69	58.1	10	1193.53	100.61
2000/2/25	17200.98	16407.15	64.11	11	1658.75	53.01
2000/3/3	17285.24	16769.96	64.95	12	2176.83	76.92
2000/3/10	17831.86	17044.28	70.37	13	1645.92	136.59
2000/3/17	17082.99	17230.09	56.1	14	1114.87	-23
2000/3/24	17784.57	17297.47	64.25	15	404.27	57.69
2000/3/31	17406.54	17432.03	57.37	16	807.38	67.76
2000/4/7	16941.68	17388.81	49.53	17	-259.3	-106.13
2000/4/14	16142.76	17198.4	38.65	18	-1142.48	-128.2
2000/4/21	15367.14	16787.61	30.77	19	-2464.72	-159.44
2000/4/28	15519.3	16527	33.94	20	-1563.69	-87.72
2000/5/12	15111.94	15725.24	29.3	21	-2294.6	-83.64
2000/5/19	14478.26	15314.67	22.81	22	-2463.42	-114.89
2000/5/26	13722.7	14911.33	17.32	23	-2420.06	-157.99
2000/6/2	15284.1	14897.49	48.22	24	-83.04	-15.86
2000/6/9	16120.26	14997.65	58.25	25	600.96	133.14
2000/6/16	16434.38	15191.94	61.6	26	1165.74	110.19
2000/6/23	15738.08	15296.3	50.76	27	626.14	61.2
2000/6/30	16155.78	15575.88	56.29	28	1677.52	54.4
2000/7/7	16829.96	16093.76	64.11	29	3107.26	105.27
2000/7/14	17586.16	16477.44	71.07	30	2302.06	97.93

Figure a1: The Augmented Dickey-Fuller test for the original data

```
        Augmented Dickey-Fuller Test

data:  train_hsiweek
Dickey-Fuller = -2.7684, Lag order = 8, p-value = 0.253
alternative hypothesis: stationary
```

Figure a2: The Augmented Dickey-Fuller test for the first order differencing data

```
        Augmented Dickey-Fuller Test

data:  difftrain_hsiweek
Dickey-Fuller = -8.7609, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(difftrain_hsiweek) : p-value smaller than printed p-value
```

Figure a3: The Dickey-Fuller test for the data with the second order of difference

```
        Augmented Dickey-Fuller Test

data:  second_difftrain_hsi
Dickey-Fuller = -13.794, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(second_difftrain_hsi) : p-value smaller than printed p-value
```

Figure a4: ACF and PACF diagram for the original data

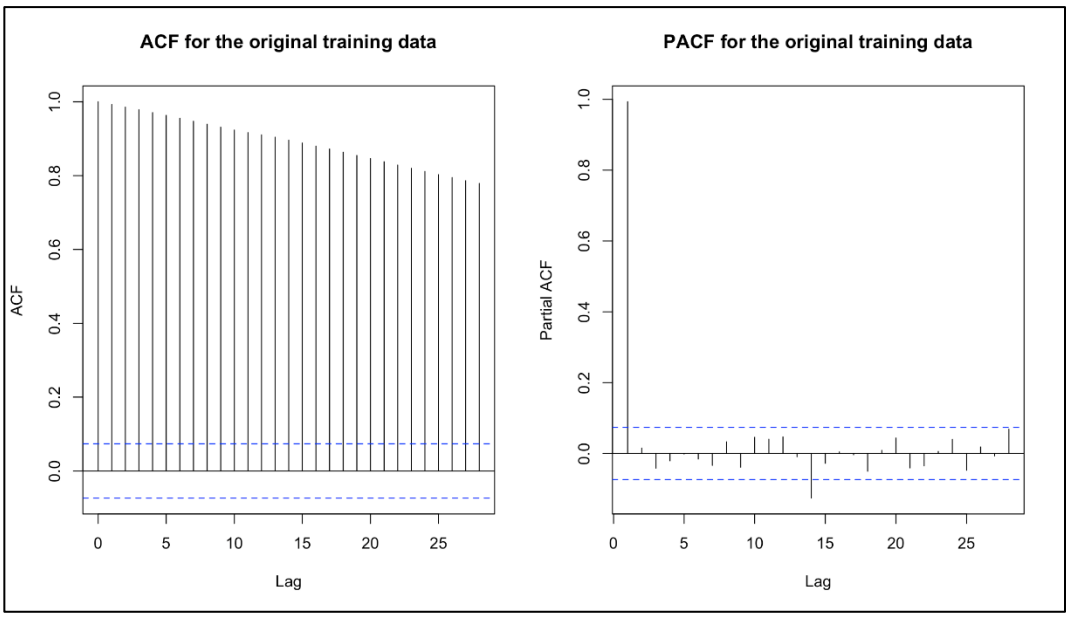

Figure a5: The fitting information of ARIMA (7,2,1)

```
Call:
arima(x = tsarima, order = c(7, 2, 1), optim.control = list(maxit = 1000))

Coefficients:
          ar1     ar2     ar3      ar4     ar5     ar6      ar7      ma1
      -0.0092  0.0361  0.0125  -0.0049  0.0145  0.0479  -0.0385  -1.0000
s.e.   0.0377  0.0376  0.0377   0.0377  0.0379  0.0379   0.0379   0.0051

sigma^2 estimated as 0.001087:  log likelihood = 1400.03,  aic = -2782.06

Training set error measures:
                      ME       RMSE        MAE          MPE      MAPE      MASE         ACF1
Training set 0.0007125806 0.03292545 0.02434887 0.007128048 0.2509808 0.9928801 0.001642439
```

Figure a6: The fitting information of ARIMA (5,2,1)

```
Call:
arima(x = tsarima, order = c(5, 2, 1), method = "ML", optim.control = list(maxit = 1000))

Coefficients:
          ar1     ar2     ar3      ar4     ar5      ma1
      -0.0106  0.0356  0.0129  -0.0034  0.0127  -1.0000
s.e.   0.0377  0.0377  0.0378   0.0377  0.0380   0.0051

sigma^2 estimated as 0.001091:  log likelihood = 1398.71,  aic = -2783.41

Training set error measures:
                      ME       RMSE        MAE          MPE      MAPE      MASE          ACF1
Training set 0.0007279065 0.03298753 0.02434248 0.007280766 0.2509671 0.9926193 -0.0002772262
```

Figure a7: The prediction of logarithmic closing price by ARIMA (5,2,1)

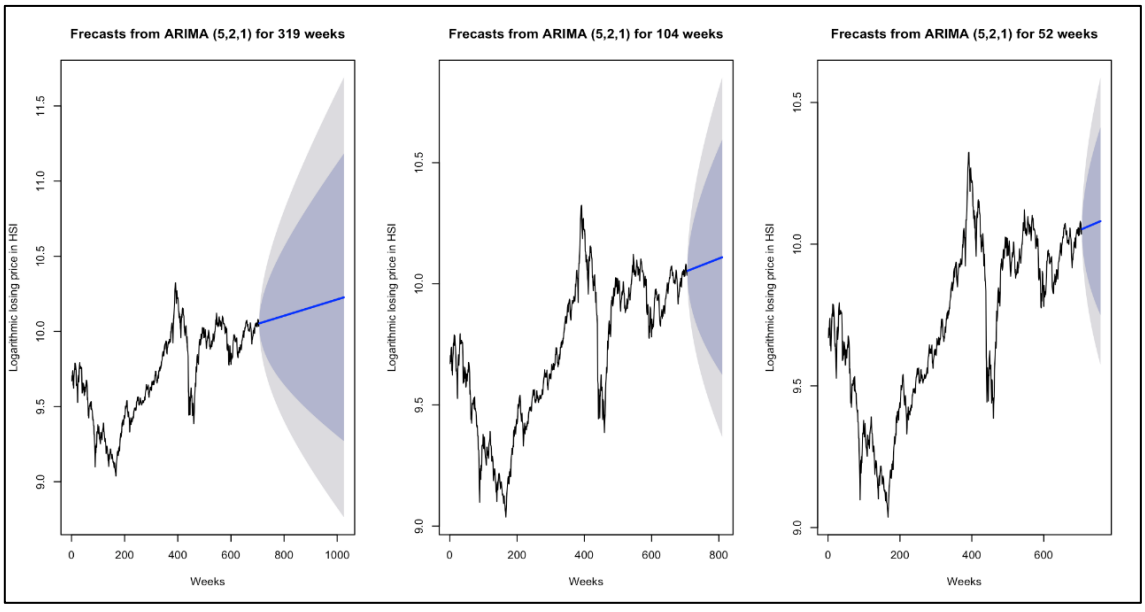

Figure a8: The prediction of closing price by ARIMA (5,2,1) for each period

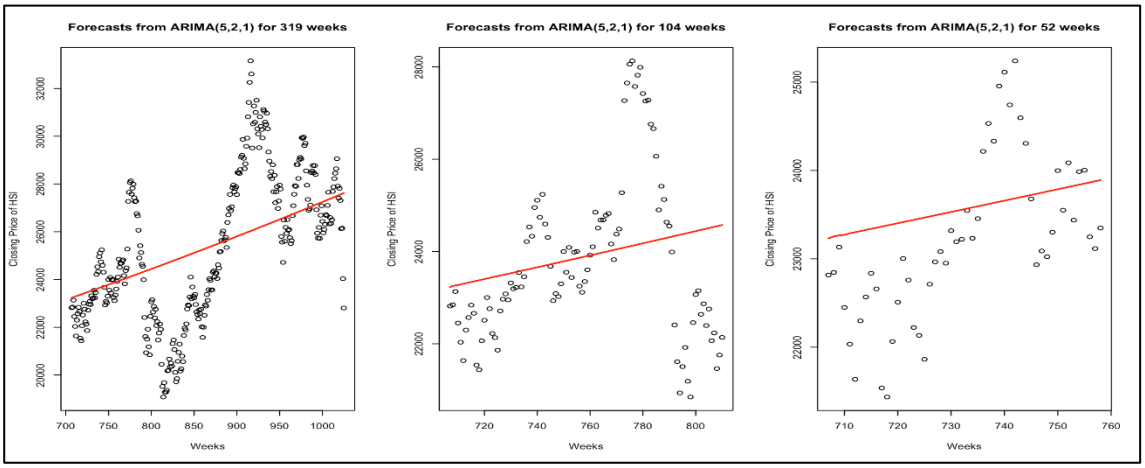

Figure a9: The LBQ test for residuals fitting the training data by ARIMA (5,2,1)

```
        Box-Ljung test

data:  fitB$residuals
X-squared = 26.475, df = 18, p-value = 0.08938
```

Figure a10: The predicted closing price of HSI by SVRs for 319 weeks

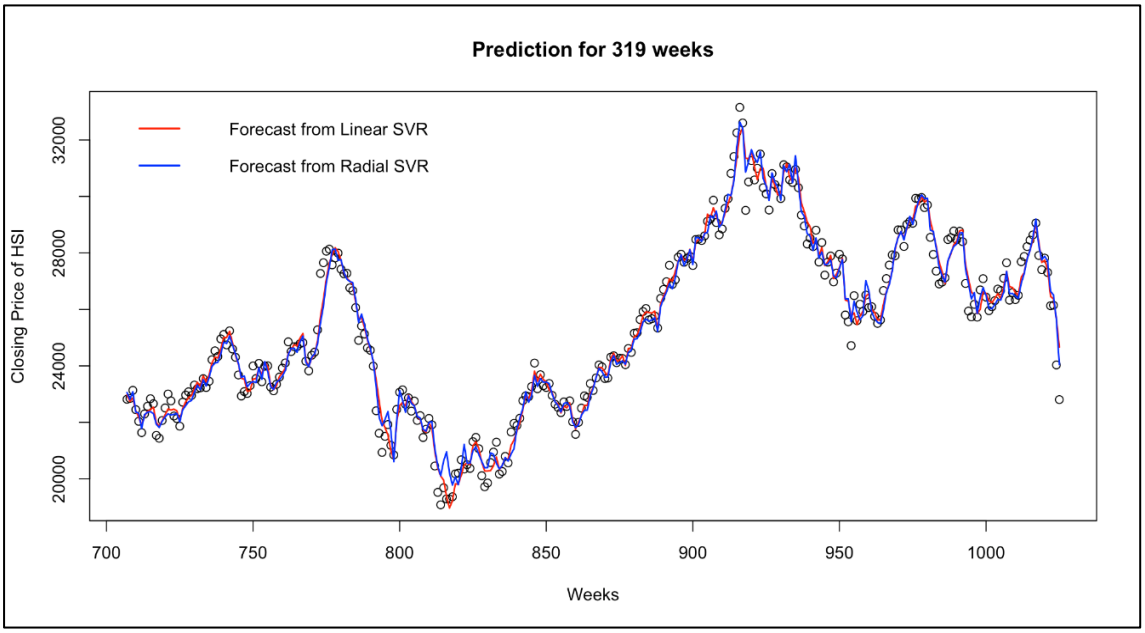

Figure a11: The predicted closing price of HSI by SVRs for 104 weeks

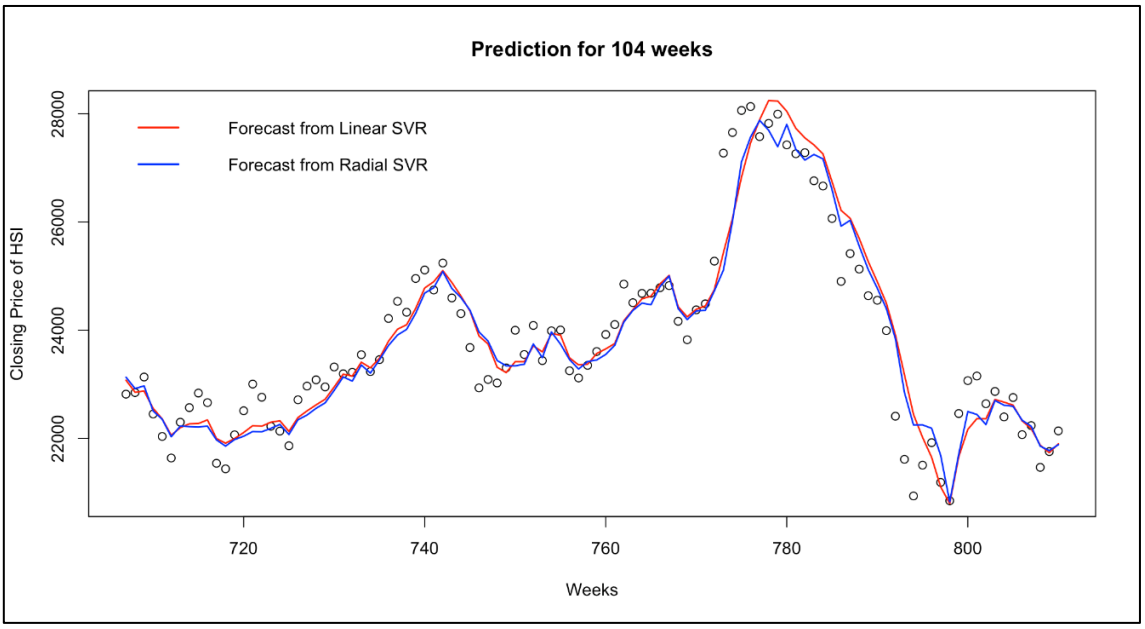

Figure a12: The predicted closing price of HSI by SVRs for 52 weeks