Available online at www.sciencedirect.com

ScienceDirect

The Journal of Finance and Data Science 4 (2018) 183–201
http://www.keaipublishing.com/en/journals/jfds/

Stock price prediction using support vector regression on daily and up to the minute prices☆

Bruno Miranda Henrique, Vinicius Amorim Sobreiro*, Herbert Kimura

University of Brasília, Department of Economics, Campus Darcy Ribeiro, Brasília, Federal District, 70910–900, Brazil

Received 22 January 2018; revised 14 March 2018; accepted 20 April 2018
Available online 27 April 2018

Abstract

The purpose of predictive stock price systems is to provide abnormal returns for financial market operators and serve as a basis for risk management tools. Although the Efficient Market Hypothesis (EMH) states that it is not possible to anticipate market movements consistently, the use of computationally intensive systems that employ machine learning algorithms is increasingly common in the development of stock trading mechanisms. Several studies, using daily stock prices, have presented predictive system applications trained on fixed periods without considering new model updates. In this context, this study uses a machine learning technique called Support Vector Regression (SVR) to predict stock prices for large and small capitalisations and in three different markets, employing prices with both daily and up-to-the-minute frequencies. Prediction errors are measured, and the model is compared to the random walk model proposed by the EMH. The results suggest that the SVR has predictive power, especially when using a strategy of updating the model periodically. There are also indicative results of increased predictions precision during lower volatility periods.

Keywords: Prediction; Stock market; Machine learning; Support vector regression; High frequency trading

1. Introduction

Stock price prediction mechanisms are fundamental to the formation of investment strategies and the development of risk management models[6]; p. 43). The Efficient Market Hypothesis (EMH), however, states that it is not possible to consistently obtain risk-adjusted returns above the profitability of the market as a whole.[20] Computational advances have led to several machine learning algorithms used to anticipate market movements consistently and thus estimate future asset values such as company stock prices[7]; pp. 193–194). Models based on the Support Vector Machine (SVM) are among the most widely used techniques.

☆ This document was a collaborative effort.

* *Corresponding author.*
E-mail addresses: brunomhenrique@hotmail.com (B.M. Henrique), sobreiro@unb.br (V.A. Sobreiro), herbert.kimura@gmail.com (H. Kimura).

Peer review under responsibility of China Science Publishing & Media Ltd.

https://doi.org/10.1016/j.jfds.2018.04.003

Information is a valuable resource when building predictive models in the pursuit of profitable financial market transaction systems. Given the peculiarities of financial time series, various challenges must be faced when developing price forecasting systems[1]; p. 4081). From a theoretical point of view, under the EMH, relevant information would be widely available to all market participants and immediately reflected in price, according to Malkiel and Fama (1970, p. 383)[20]'s EMH claims that it is impossible, consistently and over the long term, to achieve above-market returns adjusted to the level of risk assumed. As summarised by Malkiel (2003)[19]; the EMH has been questioned since its introduction, especially with the development of Malkiel (2003)[19]; the EMH has been questioned since its introduction, especially with the development of predictive systems, as shown in studies based on SVM and other algorithms (for example, Ballings et al. (2015); Nayak et al. (2015)[2,22]; and Qu and Zhang (2016))[25] that can generate profit in the long term. Malkiel and Fama (1970, pp. 386−387), however, argue that the market follows a random walk and that attempts to predict its movements in a consistent manner will be vain.

Computational advances have led to the introduction of machine learning techniques for predictive systems in financial markets. In a review of articles on predictive systems, Hsu et al. (2016, p. 215) observed that it is common to use financial series to measure the efficiency of predictive algorithms and classifiers in machine learning. Classifiers are systems that can learn, through training, to recognise patterns and thus assign a class to new data. As an example, machine learning algorithms can be used to predict insolvency, as observed by Zhou et al. (2012)[33] and Li et al. (2012).[16] In such cases, the aim is to classify companies with the highest probability of insolvency, according to an automatic classifier algorithm. Other examples are credit risk measurement, as in Li et al. (2006)[17]; and asset price forecasting, as proposed by Kao et al. (2013)[11] and Xiao et al. (2013).[30]

In addition to developing transactional strategies, progress in computational information systems has enabled rapid electronic transactions to take place in financial markets. Based on high-frequency trading algorithms, the submission and execution of purchase, sale or cancellation orders can be performed in seconds and microseconds, as Goldstein et al. (2014, pp. 182−183) note. The intensive use of rapid computational systems by some market participants may increase profitability, but the effects on normal market functioning are questionable, as not all participants have access to this type of technology[8]; pp. 182−183).

To analyse the EMH, as Hsu et al. (2016)[9] have sought to do, this study tests stock predictability in Brazil, the United States and China, based on prediction error analysis. The study does not seek to identify trading strategies that can lead to extraordinary gains but rather to evaluate prediction errors by comparing a machine learning model with a base model that follows a random walk. The choice of countries is due to the desire to evaluate results of machine learning techniques in both developed and developing markets. In regard to stock price frequency, daily data are traditional in prediction studies, such as those of Kumar et al. (2016), Zbikowski (2015)[13,32] and Patel et al. (2015b).[24] However, other studies, such as those of Qu and Zhang (2016)[25] and Manahov et al. (2014)[21]; use up-to-the-minute data. Our study shows results for both daily and up-to-the-minute prices.

Blue chip and small cap stocks are selected, as higher and lower capitalisations in each country, respectively. One hypothesis tested involves the argument that a more accurate prediction can be obtained using a frequency greater than daily. Moreover, aiming to capture changing market conditions more quickly, results are evaluated by comparing periodic updating of models with an absence of updating in terms of prediction performance in each case. The selected prediction method is a regression method based on SVM, as used by Qu and Zhang (2016), Patel et al. (2015b)[24,25] and Choudhury et al. (2014)[5]; called Support Vector Regression (SVR). Finally, it should be noted that three kernel functions are tested for SVR to identify the most suitable kernel function for this type of stock price prediction. A random walk model is used as a reference to evaluate predictions of returns.

In addition to minimising risks to stock market investors, strategies based on price prediction may provide evidence against the EMH. Predictive studies such as that presented here contribute to the building of profitable strategies, especially risk-adjusted ones, as greater predictability can affect an investment portfolio's exposure level. Thus, greater accuracy in price forecasting may imply potential risk-adjusted profits for investors. Other studies, such as those of Yeh et al. (2011), Choudhury et al. (2014)[5,31] and Patel et al. (2015b)[24]; select just one frequency for prices used in predictions. Our study contrasts different frequencies, using the same model. Specifically, predictions are made using both daily prices and up-to-the-minute prices. Another important contribution of this study is that it contrasts the results of a training model based on a fixed period with those with dynamically updated training periods, thus comparing the predictive performances of these two strategies.

Regarding the up-to-the-minutes trading frequency, costs of obtaining and processing large prices database are usually high. However, for illustrating the robustness of a price prediction method, longer periods of testing are always

preferable. The same goes for the variety of the selected securities. In this context, this article brings results from SVR price predictions for 2 years of the 1-minute historical stock prices for Brazilian small and large capitalization companies. Finally, an analysis of SVR predictions and basics statistics of each stock is introduced, indicating the existence of a relationship between the predictions precision and volatility in prices.

The paper is organised as follows. The next section presents a brief theoretical framework, explaining the methods and data frequencies used in previous studies. Sec. 3 describes the method used to predict stock prices, the performance measures considered and the data used. Sec. 4 presents and discusses the results. Finally, Sec. 5 presents conclusions and limitations of the study and suggests areas of future research.

2. Brief literature review

Describing the EMH, Malkiel and Fama (1970) state that, on balance, prices reflect all relevant information available when pricing an asset. This hypothesis arises from empirical observations of changes in price time series that are very similar to a random walk process. According to these authors, even a system in which a number of buy and sell orders are generated in the short term is not profitable, due to transaction costs and commissions Malkiel and Fama (1970, p. 396). Despite the evidence obtained for market efficiency, Malkiel and Fama (1970, pp. 413–416) further encourage a search for more data confirming or disproving their hypothesis. Since then, academic papers have sought to show that stock market prices are, to some extent, predictable. Malkiel (2003, p. 80) concludes that not all market participants are rational and that there are irregular price formations, leading to exploitable return patterns over short time periods.[27]; p. 20) consider the possibility that a profitable predictive system may exist but only up to the point of its discovery. In this case, the performance of that system would deteriorate when more market participants begin to use it.

The development of consistently profitable systems may constitute evidence against the EMH, as suggested by Hsu et al. (2016, pp. 217–218). Such systems may benefit from computationally intensive techniques, such as those that exploit machine learning algorithms. Hsu et al. (2016, p. 229) show that machine learning algorithms commonly use financial time series to evaluate their predictive capabilities. In this context Ballings et al. (2015)[2] and Gerlein et al. (2016)[7] have obtained good results in predictions when applying classifiers such as SVM, k-Nearest Neighbours (KNN), neural networks and decision trees.

Zbikowski (2015)[32]; for example, utilised SVM and a predictive variable selection method within Technical Analysis (TA) indicators. In an attempt to develop an optimal market transactions strategy,[5] used k-means to predict market volatility and SVR to predict prices in the Indian stock market. The authors analysed daily data to estimate prices for two days.[2] studied the direction of the stock market and, in so doing, evaluated classifiers. In Ballings et al. (2015)[2]'s study, annual data from more than 5000 European companies were used in classifiers such as logistic regression, neural networks, KNN and SVM. Classifiers are used to determine the direction of the respective company stocks in the following year. The authors compared the results of those classifiers to ensemble approaches, such as Random Forests (RF), AdaBoost and kernel factory. The so-called ensemble techniques involve multiple classifiers, usually of the same type or algorithm, resulting in independent classifications but with some decision method for determining a single final classification. Applying these techniques,[2] calculated their predictive variables, considering companies' balance sheets and financial statements. Based on the prediction of the direction that a particular stock would take during the year, the authors showed how a profitable strategy could be built. Also seeking to predict the direction of stock prices, Kumar et al. (2016)[13] and Kim (2003)[12] applied SVM-based systems to (TA) indicators.

Criticising classification approaches based only on the direction of the stock market,[3] attempted to predict risks and returns of Tehran stocks. The authors proposed a predictor variable selection method and applied decision tree and neural network mechanisms. The predictive variables used by Barak and Modarres (2015)[3] were constructed from financial statements published by companies. Zbikowski (2015)[32] used TA indicator values as predictive variables in his short-term trends prediction model. Zbikowski (2015)[32] proposed a modified SVM, with variable selection determined by Fisher scores, a method used to rank definitions into classes according to certain predictor variables. The TA variables used by Zbikowski (2015)[32] included On Balance Volume (OBV), the Relative Strength Index (RSI) and the Williams oscillator. The results obtained through the SVM approach exceeded those based on a buy-and-hold strategy in Zbikowski (2015)[32]'s study. In turn, Yeh et al. (2011)[31] and Lu et al. (2009)[18] applied SVR-based systems to predict the TAIEX and Nikkei 225 indices, both using daily data.

Gerlein et al. (2016)[7] proposed using computationally simpler classifiers for intraday strategies in the Foreign Exchange (FOREX) market. Decision tree algorithms and lazy models were applied to TA variables for strategies on USDJPY, GBPUSD and EURUSD prices. Patel et al. (2015a)[23] also used TA variables as inputs in their model but in a different manner from other authors. Instead of using the indicator values directly, as Gerlein et al. (2016)[7] and Zbikowski (2015)[32] did, Patel et al. (2015a)[23] used the trend indication given by the indicators. A TA indicator can identify the market trend as bullish or bearish. Thus, Patel et al. (2015a)[23]'s model uses this information as a predictive variable in algorithms such as SVM, random trees and neural networks to predict trend rather than price.

Some authors have used hybrid machine learning algorithms to increase predictive performance. Xiao et al. (2013)[30]; for example, proposed integrating various neural networks with SVM to predict the daily values of stock indices. Using TA indicators as input variables, Nayak et al. (2015)[22] proposed a hybrid SVM and KNN system for predicting index values. The results were better than those of traditional neural networks. In turn, Patel et al. (2015b)[24] evaluated the efficiency of the daily closing price predictions performed by SVM, RF and neural network hybrids, comparing the results with the isolated use of these same algorithms. The authors concluded that hybrid uses of these algorithms offered better results. Finally, Dash and Dash (2016) investigated an approach to neural networks with TA compared to traditional machine learning algorithms in stock transaction decisions.

The development of computing has not only enabled the development of complex prediction algorithms but also quantitative analysis of high-frequency data, as in Brownlees and Gallo (2006)[4]'s study. This data type requires a different treatment from that of traditional lower frequency data. Brownlees and Gallo (2006)[4] suggest ways of treating high frequency data with regard to outliers and temporal irregularities. Once treated, high frequency data can be used to develop very fast market transaction strategies, known as High Frequency Trading (HFT). This type of transaction has become common in today's financial markets, and its effects have been studied by authors such as Lee (2013)[14] and Goldstein et al. (2014).[8]

Given the advances in the use of HFT discussed by Araújo et al. (2015, p. 4082), there is plenty of room for the development of models in this area. For example, Araújo et al. (2015)[1] developed a new mathematical model to forecast price changes, measured in seconds, of companies listed on the BM&F Bovespa. Meanwhile, (Manahov et al., 2014)[21] investigated the profitability of a genetic learning algorithm for the FOREX market, applied to price changes measured in minutes. SVR is applied to up-to-the-minute prices in the Chinese market in Qu and Zhang (2016)[25]'s study. Other studies have applied prediction techniques to even higher frequency data. Brownlees and Gallo (2006)[4]; for example, suggest the use of HFT on immediate price variations, that is, on ticks that vary according to each purchase or sale transaction performed.

The articles used in the construction of the following prediction models are shown in Table 1. The main prediction methods for each article and frequency of empirical data are also shown in the table. Note that most of the studies analysed apply a daily data model, highlighting the need to produce research with higher data frequencies. Therefore,

Table 1
Prediction methods used by this study's references and data frequency.

Reference	Method	Data frequency
Tay and Cao (2001)[26]	SVM	Daily data
Kim (2003)[12]	SVM	Daily data
Lu et al. (2009)	SVR	Daily data
Yeh et al. (2011)	SVR	Daily data
Choudhury et al. (2014)	SVR	Intraday data
Araújo et al. (2015)	Based on neural networks	Data in seconds
Ballings et al. (2015)	SVM, KNN, neural networks	Annual data
Nayak et al. (2015)	SVM, KNN	Daily data
Patel et al. (2015b)	SVR, RF, neural networks	Daily data
Patel et al. (2015a)	SVR, RF, neural networks	Daily data
Zbikowski (2015)[32]	SVM	Daily data
Dash and Dash (2016)[6]	Neural networks	Daily data
Gerlein et al. (2016)	Multiple classifiers	Daily data
Hsu et al. (2016)	SVM, neural networks	Intraday data
Qu and Zhang (2016)[25]	SVR	Intraday data
Kumar et al. (2016).	SVM	Daily data

as stated earlier, this study applies SVR prediction models not only to daily data but also to data with an up-to-the-minute frequency.

3. Method

The prediction of closing stock prices in this study is performed by SVR. The results are compared to returns obtained by the random walk model, assuming zero average returns on stocks and a given variance. The Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the adequacy of the models' price predictions. These measures have also been used by Nayak et al. (2015), Patel et al. (2015b), Araújo et al. (2015), Manahov et al. (2014)[1,21,22,24] and Choudhury et al. (2014).[5] Based on these authors' studies, the MAPE and RMSE can be calculated according to Eqs (1) and (2), respectively.

$$\text{MAPE} = \frac{1}{T}\sum_{i=1}^{T}\left|\frac{d_i - \widehat{d}_i}{d_i}\right| \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}\left(d_i - \widehat{d}_i\right)^2} \tag{2}$$

The following notations were adopted for the variables used in this study:

- N: total samples;
- T: total test samples;
- d: real sample value;
- $\widehat{d}$: value estimated by the model;
- P: period (minutes or days);
- Cl: closing price for the period;
- Hi: maximum price in the period;
- Lo: minimum price in the period;
- Up: number of price rises;
- Dw: number of price reductions;
- r: return.

3.1. Technical analysis indicators

The predictor variables commonly used in the literature for SVR and SVM models are TA indicators.[6,7,9,10,12,13,24] According to Nayak et al. (2015, p. 672), a TA indicator is composed of data derived from the application of a certain formula to the past prices of a stock. This paper considers the values of these indicators, detailed below, as predictive variables in the SVR model.

The simplest selected indicators are the a moving averages, which are easy to understand and calculate. The Simple Moving Average (SMA) is the arithmetic mean of T past prices Cl_i, according to Nayak et al. (2015, p. 672) and as described by Eq. (3).

$$\text{SMA} = \frac{1}{T}\sum_{i=1}^{T} Cl_i \tag{3}$$

One moving average variation, known as the Weighted Moving Average (WMA), used in this study, assigns higher weightings to more recent prices. A WMA of P periods is given by Eq. (4), as shown in Patel et al. (2015b, p. 2164). This study defines Cl_i as either daily closing or up-to-the-minute prices, depending on the frequency used in the analysis.

$$\text{WMA} = \frac{PCl_i + (P-1)Cl_{i-1} + \ldots + Cl_{i-P}}{P + (P-1) + \ldots + 1} \tag{4}$$

The Relative Strength Index (RSI) is a comparison indicator between losses and recent gains and determines an overbought or oversold market. At time t, it has the form of Eq. (5)[24]; p. 2164). In this study, the RSI is calculated for the stock's closing prices in the selected period.

$$\text{RSI} = 100 - \frac{100}{1 + \frac{\sum_{i=0}^{P-1} Up_{t-i}/n}{\sum_{P-1} Dw_{t-i}/n}} \tag{5}$$

The Accumulation/Distribution Oscillator (ADO) is an indicator of momentum, i.e., the strength of a price trend Hsu et al. (2016, p. 221), and is given by Equation (6).

$$\text{ADO} = \frac{Hi_t - Cl_{t-1}}{Hi_t - Lo_t} \tag{6}$$

The final TA indicator considered in this study is the Average True Range (ATR). This indicator seeks to measure bull and bear price trends, using the True Range (TR), defined in Eq. (7). According to Nayak et al. (2015, p. 672), the ATR represents a mean, given by Eq. (8), in which the first value of the period is given by $1/P\sum_{i=1}^{P}\text{TR}_i$.

$$\text{TR} = \max[Hi_t - Lo_t, |Hi_t - Cl_{t-1}|, |Lo_t - Cl_{t-1}|] \tag{7}$$

$$\text{ATR}_t = \frac{\text{ATR}_{t-1}(P-1) + \text{TR}_t}{P} \tag{8}$$

3.2. Support vector regression

A classification method based on SVM maps the independent variables of N samples available into a space of more dimensions and is typically used to classify observations between groups. This method, developed by Vapnik (1995)[28]; uses $\{(\mathbf{x}_k, y_k)\}_{k=1}^{N}$ training observations to build a linear model using non-linear classification thresholds, mapping variables on a greater number of dimensions. Separation between classes is achieved using an optimal hyperplane, calculated based on N observations, where $\mathbf{x}$ is the independent variable vector, and y is classification $y_k \in \{-1, 1\}$ for each sample. Thus, the classification hyperplane is given by Eq. (9), which satisfies the conditions of Eqs (10) and (11).

$$\mathbf{w}^T\phi(\mathbf{x}_k) + b = 0 \tag{9}$$

$$\mathbf{w}^T\phi(\mathbf{x}_k) + b \geq 1 \text{ para } y_k = 1 \tag{10}$$

$$\mathbf{w}^T\phi(\mathbf{x}_k) + b \leq -1 \text{ para } y_k = -1 \tag{11}$$

In the context of price prediction, the goal is not necessarily classification into groups but estimation of real values. This study therefore uses SVR, which is employed to obtain a regression model used to predict asset prices. The SVR model in this paper includes the following notations:

- $\mathbf{x}$: vector with predictor variables;
- y: sample classification;
- $\mathbf{w}$: weight vector;
- b: constant;
- C, c: model parameters.

The function $\phi(\bullet) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_k}$ is the said mapping of independent variables to a space with a greater number of dimensions, in which it is possible to linearly separate the samples according to each class. Eq. (9) and its conditions in Eqs. (10) and (11) are summarised in Eq. (12).

$$y_k[\mathbf{w}^T\phi(\mathbf{x}_k) + b] \geq 1 \tag{12}$$

The classification condition originally proposed by Vapnik (1995)[28] is $y(\mathbf{x}) = \text{sgn}(\mathbf{w}^T\phi(\mathbf{x}) + b)$. However, even in the new space mapped by the function $\phi(\bullet)$, there may not be a perfect separation of N samples into two classes $\{-1, 1\}$. Thus, we define a variable $\xi \geq 0$ as a tolerance margin in the classification thresholds, making the classifier more flexible in accepting possible errors. With this flexibilisation, the hyperplane condition in Eq. (12) becomes Eq. (13), and the problem of finding the optimal hyperplane becomes a convex optimisation problem given by Eq. (14). In this equation, C is the adjustment parameter for the edge of the hyperplane with the smallest possible misclassification, under the conditions of Eq. (13).

$$y_k[\mathbf{w}^T\phi(\mathbf{x}_k) + b] \geq 1 - \xi_k \tag{13}$$

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{k=1}^{N} \xi_k \tag{14}$$

The above formulated optimisation can be converted into Wolfe (1961)[29]'s dual, given by Eq. (15), in which $\mathbf{y}^T\boldsymbol{\alpha} = 0$ e $0 \leq \alpha_i \leq C, i = 1, \ldots, N$. In this model, $\mathbf{e} = [1, \ldots, N]^T$ represents a vector of unit values, and Q is a l by l matrix in which $Q_{ij} \equiv y_i y_j K(x_i, x_j)$. The function $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ is called the kernel function.

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \tag{15}$$

The solution of Wolfe (1961)[29]'s dual function turns the classification decision function into the form of Eq. (16).

$$\text{sgn}(\boldsymbol{w}^T\phi(\boldsymbol{x}) + b) = \text{sgn}\left(\sum_{i=1}^{N} y_i\alpha_i K(x_i, x) + b\right) \tag{16}$$

As already indicated, SVR uses principles similar to SVM, but the response variable is a continuous value $y \in \mathbb{R}$. However, as shown by Huang and Tsai (2009, p. 1530) and Patel et al. (2015b, p. 2164), instead of seeking the hyperplane in Eq. (13), SVR seeks the linear regression function, given by Eq. (17). To achieve this, a threshold error ε is defined to be minimised in the expression in Equation (18). This expression is called the ε-insensitivity loss error function. The SVR regression process therefore seeks to minimise ε in Eq. (18) and $||\mathbf{w}||^2$ in the expression of R, defined in Eq. (19).

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x} + b \tag{17}$$

$$|y - f(\mathbf{x}, \mathbf{w})|_\varepsilon = \begin{cases} 0, & \text{case} |y - f(\mathbf{x}, \mathbf{w})| \leq \varepsilon \\ |y - f(\mathbf{x}, \mathbf{w})| - \varepsilon, & \text{otherwise} \end{cases} \tag{18}$$

$$R = \frac{1}{2}||\mathbf{w}||^2 + c\left(\sum_{i=1}^{N} |y_i - f(\mathbf{x}_i, \mathbf{w})|_\varepsilon\right) \tag{19}$$

Tolerance variables are again introduced, defining ζ as the value in excess of ε and ζ^* to limit the value to the regression target. Thus, the minimisation of Eq. (19) becomes Eq. (20), under the conditions of Eqs. (21) and (22) for ζ_i and $\zeta_i^* \geq 0$ and $i = 1, 2, \ldots, N$.

$$R = \frac{1}{2}\|\mathbf{w}\|^2 + c\sum_{i=1}^{N} \left(\zeta_i + \zeta_i^*\right) \tag{20}$$

$$(\mathbf{w}^T\mathbf{x}_i + b) - y_i \leq \varepsilon + \zeta_i \tag{21}$$

$$y_i - (\mathbf{w}^T\mathbf{x}_i + b) \leq \varepsilon + \zeta_i^* \tag{22}$$

This paper considers common forms of kernel functions from the literature[2,10,22,24]; given explicitly by the linear, radial and polynomial functions in Eqs (23)–(25), respectively.

$$K(x_i, x_j) = x_i^T x_j \tag{23}$$

$$K(x_i, x_j) = e^{-\gamma\left|\left|x_i - x_j\right|\right|^2}, \text{para } \gamma > 0 \tag{24}$$

$$K(x_i, x_j) = \left(x_i^T x_j + 1\right)^d \tag{25}$$

Note that the shape of the kernel function directly influences the values obtained by the SVR regression. Similarly, the constant c in Eq. (19) and the parameters γ and d in Eqs. (24) and (25) should be optimised. For this purpose, a training data set is divided into two new sets: the first is used to choose the optimal parameters, and the second is used to validate the smallest error possible, given these choices. This process, called k-fold cross validation, selects the parameters c, γ and d, according to the lowest RMSE.

To illustrate the general idea of price prediction, Fig. 1 visually demonstrates the difference between returns prices observed in the market and those predicted by the SVR model. The data used in this illustration are daily and are not used in the following experiments. As seen in Fig. 1, the SVR model implies considerable prediction error, and the literature seeks to minimise these errors, as in this study.

Fig. 1. Real returns and those predicted by the SVR model for daily prices. The continuous curve represents returns observed in the market, and the dashed curve represents returns predicted by the model.

3.3. Random walk

According to Malkiel and Fama (1970, p. 383), prices in efficient markets should reflect all information available at a particular time. Any price changes are difficult to predict and should be independent of previous values. Following a basic approach to efficiency analysis, returns must be independent and identically distributed. It should be noted that this study only investigates price unpredictability, using a random walk model. Thus, if the price of a stock market asset at time t follows a random walk, as described in Araújo et al. (2015, pp. 4083–4084), the price behaviour must be based on Eq. (26), with return r at time t, and follow a normal distribution, with zero mean and a variance of σ^2 ($r \sim \mathcal{N}(0, \sigma^2)$) , i.e., white noise. With this return distribution, the market would be unpredictable and efficient enough not to allow the formation of profitable transactional strategies on the long-term.

$$Cl_t = Cl_{t-1} + r_t \tag{26}$$

This study uses random walk predictions as a reference model for comparison with SVR predictions, as set out by.[18] According to Eq. (26), the model predicts that the next stock price will be the current price plus a random value taken from the distribution $\mathcal{N}(0, \sigma^2)$.

3.4. Data

This study considers Brazilian, American and Chinese stocks, with three blue chip and three small cap stocks for each country, totaling 18 assets. The stocks were chosen to obtain a distribution of companies of different sizes in different markets of both developed and developing nature. The stocks selected in this study are shown in Table 2. The time period selected for daily prices comprises 15 years. It should be noted, as Table 2 shows, some of the stocks do not have all those historical prices available, specially some small capitalization companies. However, most stocks selected have more than 10 years of historical daily prices, including bull and bear periods.

Due to processing costs and data availability, 1-min historical prices were constrained as detailed in Table 3. The data period for 1-min prices for all stocks is from March 1 to May 26, 2017, i.e., three months of up-to-the-minute prices, a window of analysis selected following[21] high frequency study. This period is used to compare up-to-the-minutes predictions across markets and stocks regarding their capitalisation, in both fixed trained SVR models and dynamically updated ones. A longer period of historical 1-min prices is also considered for six stocks, limited to one market because of processing costs limitations. Brazilian stocks, therefore, are selected for this longer period of testing, as a means to evaluate SVR predictions stability over the long run in high frequency trading. As shown by

Table 2
Selected stocks and historical periods considered in this article for daily prices.

Code	Company name	Country	Classification	Start date	End date
BBAS3	*Banco do Brasil*	Brazil	Blueship	01/05/2002	31/05/2017
PETR4	*Petrobrás*	Brasil	Blueship	02/05/2002	31/05/2017
VALE5	*Vale do Rio Doce*	Brazil	Blueship	01/05/2002	30/05/2017
ALPA4	*Alpargatas*	Brazil	Small cap	02/01/2007	29/05/2017
DIRR3	*Direcional Engenharia*	Brazil	Small cap	19/11/2009	26/05/2017
LEVE3	*Metal Leve*	Brazil	Small cap	01/05/2002	31/05/2017
BAC	*Bank of America*	USA	Blueship	01/05/2002	30/05/2017
GOOGL	*Google*	USA	Blueship	19/08/2004	30/05/2017
XOM	*Exxom Mobil*	USA	Blueship	01/05/2002	30/05/2017
ANGI	*Angie's List*	USA	Small cap	17/11/2011	26/05/2017
HL	*Hecla Mining*	USA	Small cap	01/05/2002	30/05/2017
PZZA	*Papa Johns*	USA	Small cap	01/05/2002	30/05/2017
600028	*China Petroleum*	China	Blueship	01/05/2002	31/05/2017
601318	*Ping an Insurance*	China	Blueship	08/05/2007	31/05/2017
601668	*China State Construction Eng*	China	Blueship	29/07/2009	26/05/2017
1432	*China Shengmu Organic Milk*	China	Small cap	15/07/2014	31/05/2017
1970	*IMAX China Holding*	China	Small cap	08/10/2015	31/05/2017
2030	*Cabbeen Fashion*	China	Small cap	28/10/2013	29/05/2017

Table 3
Fixed training and testing periods for the 1-min prices.

Set	Period	Minutes	Days
3-months cases (all stocks):			
Train	1/3/2017 to 2/5/2017	23555	39
Test	5/5/2017 to 26/5/2017	10006	16
2-years cases (Brazilian stocks):			
Train	10/2/1016 to 4/7/2017	123972	350
Test	5/7/2017 to 8/2/2018	53132	150

Tables 3 and 2 years of 1-min prices are used for that case. Data were obtained from Reuters©, Yahoo!Finance and BM&F Bovespa. Minutes without information were considered to have the same prices as the previous minute.

To compare the SVR prediction results with regard to different price frequencies, daily and up-to-the-minute data are used for the periods described above. Daily prices are widely used in academic studies.[5,10,12,18,24,26,31] The results obtained with daily prices can be compared to the use of higher frequency, up-to-the-minute prices. It is worth mentioning that valid prices in this study include only those obtained during the respective sessions in each market considered. Thus, the data were limited in advance to the official opening and closing times of each market. It should be noted that high frequency data, for example, those expressed in milliseconds, are not within the scope of this study, due to the method employed.

Two strategies are considered when using SVR. The first is to separate the data into training observations, for the optimisation of the SVR model, and test observations, with prediction errors calculated on the closing prices, as adopted by Gerlein et al. (2016), 296 Manahov et al. (2014), Nayak et al. (2015)[7,21,22] and Patel et al. (2015b).[24] For daily prices, each period in Table 2 is split in a training set, with approximately 70% contiguous days, and a test set, with the remaining 30% days. For the 1-min historical prices, Table 3 shows the divisions into training and testing periods for all studied stocks, following the same strategy of separating 70% of data for training the models and 30% for testing them.

The second SVR optimisation strategy involves updating the model as new information becomes available in the market, as suggested by Lessmann et al. (2011[15]; p. 2122) and Hsu et al. (2016, p. 223). It is a dynamic optimisation that uses periodically updated training data, known as a sliding or moving window. It is expected that this procedure will capture new market conditions as soon as possible.

Regarding the up-to-the-minute prices, although three months seems a short interval, it should be noted that period comprises more than 33000 data points, making the task of obtaining and processing the prices for all 18 selected securities a challenging effort. However, a longer period of analysis is desirable, mainly because three months may not include all possible market conditions for real tests. In this context, as stated before, this article brings yet another SVR evaluation for price prediction, using 2 whole years of up-to-the-minutes prices. Brazilian stocks are selected for this simulation over the long run, with data gathered directly from BM&F Bovespa. Such a long period, in terms of 1-min prices, contains short-term bull and bear markets for this timeframe, being suitable for evaluating the models' predictions stability.

4. Analysis and results

Before applying SVR to the prices described in this paper, prices and the TA indicators highlighted above − namely, SMA, WMA, RSI, ADO and ATR − were calculated and normalized. A calculation period of $P = 10$ was fixed for all indicators. The SVR prediction model transactions for the up-to-the-minute data prices thus start just 10 min after the beginning of each trading session. The analysis of results are organised into two sections, one dedicated to the use of a fixed training period as described before and other section dedicated to constantly updating the model in a moving training window. Both these sections consider the 15 years historical daily prices, the 3 months period for up-to-the-minutes prices and the results of predicting prices using 2 years of 1-min Brazilian stock data. Finally, a dedicated section introduces a correlation study between average returns and volatility and the SVR prediction precision.

4.1. Fixed training

Once the TA indicators were calculated and normalized, the SVR models were optimised for the fixed training prices described in Table 3, for the 1-min prices, and the 70% first contiguous available daily prices shown in Table 2. The optimum parameters for each kernel function, according to Eqs. (19), (24) and (25), are given in Table 4, in the case of daily prices, in Table 5, in the case of up-to-the-minute prices for the 3-months period and finally in Table 6 for the 2-years 1-min prices period. Each table also shows the RMSE associated with optimisation, using training data.

Applying the parameters in Tables 4–6, the SVR models with each kernel function were run on daily and up-to-the-minute price test sets, respectively, according to the periods shown in Tables 2 and 3. The resulting error in RMSE and MAPE for each stock is given in Table 7 for daily prices, in Table 8 for up-to-the-minute prices during the 3-months data period and in Table 9 for the 2-years period of 1-min prices. In most cases, almost all of the RMSE errors are greater in the test data. This behaviour is expected, as the models are optimised for the training data. However, there are exceptions, especially with the use of the linear kernel for all periods and price timeframes considered.

Comparison of Tables 4 and 7 reveals that the daily test data SVR had smaller errors, compared to those RMSE errors on the optimization phase on the daily training data for the following stocks: PETR4, VALE5, DIRR3, BAC, ANGI, HL, 601318, 1970 and 2030. When using daily prices, the use of radial kernel resulted in smaller errors for the test data set only in the case of BAC stock. Usage of polynomial kernels did not, for any stock, result in smaller errors in the test data than in the training data.

Up-to-the-minute test data frequencies, for the 3-months period data, are compared in Tables 5 and 8. In this case, the application of SVR with a linear kernel to the test data produces smaller errors than the same SVR applied to training data for the following stocks: BBAS3, PETR4, VALE 5, ALPA4, DIRR3, LEVE3, XOM, 600028, 601318, 601668, 1432, 1970, and 2030. Interestingly, the application of the linear kernel to the test data resulted in a smaller error than its application to the training data for all Brazilian and Chinese sample stocks. When using the longer period of training/testing of 2-years of 1-minute prices, that behaviour is not observed for BBAS3, DIRR3 and LEVE3, which can be noted by comparing Tables 6 and 9.

The results in Tables 7–9 are indicative of the SVR's superior predictive power when using a linear kernel compared with radial and polynomial kernels, for this study's selected stocks. To measure the significance of these results, the errors produced were compared with the errors produced by a random walk model. To that end, random predictions were generated according to the model given by Eq. (26). To model white noise in that equation, the

Table 4
Optimal parameters and RMSE errors for each stock and kernel with daily prices.

Stock	Linear kernel		Radial kernel			Polynomial kernel		
	RMSE	c	RMSE	c	γ	RMSE	c	d
BBAS3	0.05745	1.00	0.05603	1.00	0.37132	0.05605	1.00	2.00
PETR4	0.07704	1.00	0.07856	1.00	0.35509	0.07343	1.00	2.00
VALE5	0.07368	1.00	0.07488	1.00	0.48782	0.06912	1.00	2.00
ALPA4	0.05421	1.00	0.05792	1.00	0.53674	0.05179	1.00	2.00
DIRR3	0.09526	1.00	0.09024	1.00	0.26466	0.09122	1.00	3.00
LEVE3	0.06949	1.00	0.10048	1.00	5.82473	0.07177	1.00	1.00
BAC	0.07108	1.00	0.07053	1.00	0.67116	0.06965	1.00	3.00
GOOGL	0.02864	1.00	0.02810	1.00	0.28154	0.02756	1.00	2.00
XOM	0.05099	1.00	0.05163	1.00	0.41009	0.04896	1.00	3.00
ANGI	0.13171	1.00	0.12078	1.00	0.45076	0.11700	1.00	3.00
HL	0.13238	1.00	0.12264	1.00	0.28910	0.12342	1.00	3.00
PZZA	0.01458	1.00	0.01452	1.00	0.47304	0.01370	0.25	3.00
600028	0.09759	1.00	0.09259	1.00	0.62462	0.08162	1.00	2.00
601318	17.3848	1.00	0.08261	1.00	0.63188	0.08073	1.00	3.00
601668	0.04796	1.00	0.05288	1.00	0.54471	0.04295	0.50	3.00
1432	0.18971	1.00	0.15882	1.00	0.35996	0.17126	1.00	3.00
1970	0.25036	1.00	0.22665	1.00	0.43222	0.21895	0.50	3.00
2030	0.10831	1.00	0.11371	1.00	1.07655	0.09542	0.50	3.00

Note: c: required parameter for all kernels; γ: Radial kernel parameter; d: Polynomial kernel parameter.

Table 5
Optimal parameters and RMSE errors for each stock and kernel with up-to-the-minute prices (3-months cases).

Stock	Linear kernel		Radial kernel			Polynomial kernel		
	RMSE	c	RMSE	c	γ	RMSE	c	d
BBAS3	0.05232	1.0	0.02098	1.0	0.37626	0.01644	1.00	2.0
PETR4	0.05704	1.0	0.03171	1.0	0.43427	0.02073	0.50	3.0
VALE5	0.05935	1.0	0.03744	1.0	0.42226	0.02168	0.50	3.0
ALPA4	0.05435	1.0	0.02998	1.0	0.51963	0.02147	0.50	3.0
DIRR3	0.05152	1.0	0.02748	1.0	0.42272	0.01988	1.00	3.0
LEVE3	0.05453	1.0	0.02376	1.0	0.33712	0.01893	0.25	3.0
BAC	0.03005	1.0	0.02766	0.5	1.36658	0.01473	0.25	2.0
GOOGL	0.00970	1.0	0.01770	1.0	5.80610	0.01017	1.00	2.0
XOM	0.00666	1.0	0.00715	1.0	2.37655	0.00375	1.00	3.0
ANGI	0.00953	1.0	0.02128	1.0	3.84291	0.00594	0.50	2.0
HL	0.03343	1.0	0.03927	1.0	1.43961	0.02188	0.25	1.0
PZZA	0.01143	1.0	0.01683	1.0	1.42533	0.01143	1.00	1.0
600028	0.08102	1.0	0.02319	1.0	0.40770	0.01705	1.00	2.0
601318	0.04970	1.0	0.01157	1.0	0.68792	0.00695	0.50	3.0
601668	0.06899	1.0	0.04563	1.0	0.48394	0.02585	1.00	2.0
1432	0.04145	1.0	0.02251	1.0	0.75254	0.01464	0.25	3.0
1970	0.07106	1.0	0.03555	1.0	0.42316	0.02500	1.00	3.0
2030	0.05315	1.0	0.04332	1.0	0.56395	0.02638	0.25	2.0

Note : c: required parameter for all kernels; γ: Radial kernel parameter; d: Polynomial kernel parameter.

Table 6
Optimal parameters and RMSE errors for each stock and kernel with up-to-the-minute prices (2-years cases).

Stock	Linear kernel		Radial kernel			Polynomial kernel		
	RMSE	c	RMSE	c	γ	RMSE	c	d
BBAS3	0.02521	1.00	0.03139	1.00	0.39859	0.02019	1.00	2.00
PETR4	0.02808	1.00	0.03247	1.00	0.41870	0.02278	1.00	2.00
VALE5	0.02279	1.00	0.03458	1.00	0.49505	0.02063	1.00	2.00
ALPA4	0.01361	1.00	0.01749	1.00	0.47601	0.01135	2.00	1.00
DIRR3	0.04192	1.00	0.04371	1.00	0.33309	0.03758	1.00	3.00
LEVE3	0.03084	1.00	0.03370	1.00	0.26088	0.02863	1.00	2.00

Note: c: required parameter for all kernels; γ: Radial kernel parameter; d: Polynomial kernel parameter.

Table 7
SVR results, using the optimal parameter in Table 4 on the daily price test data.

Stock	Linear kernel		Radial kernel		Polynomial kernel	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
BBAS3	0.16427	0.20141	1.13716	1.06158	19.1481	32.3253
PETR4	0.06727	0.40963	0.57590	1.00977	2.18556	7.60178
VALE5	0.05786	0.57993	0.86932	2.96455	3.46965	29.2500
ALPA4	0.08676	0.31555	0.40642	0.54263	1.07965	2.92223
DIRR3	0.08743	0.06276	1.15706	0.64556	1.52443	0.67216
LEVE3	0.14270	0.10383	1.61960	1.25777	0.12765	0.09086
BAC	0.04395	0.12505	0.06929	0.26302	0.08167	0.22308
GOOGL	0.06248	0.03752	1.97066	1.30933	1.18256	0.72193
XOM	0.07517	0.06040	1.08478	0.87865	6.87381	4.13189
ANGI	0.10194	0.42575	0.53587	1.59145	0.78755	1.11141
HL	0.08209	0.29055	0.32289	0.52430	1.33855	1.53013
PZZA	0.07511	0.09601	1.97670	1.49839	30.7060	10.3271
600028	0.17402	3.40886	0.16198	3.27387	2.20295	17.3848
601318	0.31341	0.31624	0.63316	0.45609	3.23618	2.20723
601668	0.16790	0.11452	1.49302	0.94396	69.1139	29.1619
1432	0.19716	0.72379	0.30898	0.93350	0.82217	3.02214
1970	0.19135	0.27265	0.46560	0.52747	0.33288	0.36864
2030	0.07756	0.14436	1.74701	2.57181	3.24215	4.18583

Table 8
SVR results, using the optimal parameter in Table 5 on the up-to-the-minute price test data (3-month cases).

Stock	Linear kernel		Radial kernel		Polynomial kernel	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
BBAS3	0.04280	0.10205	1.17903	1.11416	1.22156	5.78860
PETR4	0.04412	0.71739	0.20380	0.75658	2.41968	3.48905
VALE5	0.02790	0.15494	0.33634	0.36753	1.008094	8.495684
ALPA4	0.04380	0.25884	0.67884	2.11752	230.1143	2004.279
DIRR3	0.04566	0.03532	0.64196	0.28137	4.52389	1.81663
LEVE3	0.02787	0.10794	0.69095	1.73337	3.52646	4.06868
BAC	0.04049	0.56743	0.02367	0.20404	0.43350	3.00627
GOOGL	0.02330	0.02525	1.03981	1.21231	0.47182	0.29781
XOM	0.00575	0.04100	0.01217	0.07729	0.03574	0.20269
ANGI	0.03928	0.02649	2.00272	1.34247	4.05219	1.29621
HL	0.04158	0.12863	0.07357	0.17985	0.02613	0.07963
PZZA	0.01866	0.06666	0.20137	0.36305	0.01872	0.06700
600028	0.02321	0.07113	0.52052	0.65335	6.36125	9.70272
601318	0.03883	0.04743	1.58233	1.00422	35.7510	11.4795
601668	0.03725	0.44597	0.03686	0.30912	0.39956	2.63612
1432	0.01575	0.00806	1.27500	0.76444	0.94059	0.52970
1970	0.03319	0.22744	0.09702	0.36440	0.33793	1.79328
2030	0.03708	0.21539	0.50188	0.55486	1.52496	4.67396

Table 9
SVR results, using the optimal parameter in Table 6 on the up-to-the-minute price test data (2-years cases).

Stock	Linear kernel		Radial kernel		Polinomial kernel	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
BBAS3	0.02364	0.10063	0.38569	0.18316	0.38382	0.63054
PETR4	0.04348	0.18961	0.30784	0.28624	0.72348	1.04578
VALE5	0.03699	0.04743	0.09120	0.09199	0.53417	0.52307
ALPA4	0.03879	0.02882	1.81449	1.08270	4.89478	2.76474
DIRR3	0.03397	0.20610	0.03820	0.19452	0.38056	0.83559
LEVE3	0.03067	0.09856	0.55004	0.23597	1.17608	1.91663

distribution variance was estimated according to price frequency. The variance in daily prices was therefore estimated using the closing prices of the most recent 7 days. Similarly, the 90 most recent minutes were used to estimate the variance in the distribution of returns for up-to-the-minute prices. Based on this procedure, the RMSE and MAPE results from obtaining the random walk model's predictions of daily closing and up-to-the-minute prices are as shown in Table 10.

Comparison of errors produced by the SVR and random walk models for the fixed training period with respect to daily prices, shown in Tables 7 and 10, respectively, reveals that the SVR model with a linear kernel has superior predictive power to the random walk model only for the following stocks: DIRR3, BAC, HL, 1970, and 2030. This means that the RMSE and MAPE error measures are smaller for the linear kernel SVR model for these stocks than those obtained for daily price predictions generated by a random walk model. Such stocks are exemplars and have common characteristics, most being classified as small caps in the three studied countries, as shown in Table 2, exception being the BAC stock. For the SVR with radial and polynomial kernels, no stocks had the prices predicted better than with the random model in the case of daily prices. Although by few stocks, these results also indicate superior predictive power using a simple linear kernel.

The errors of the SVR models with a fixed training period with regard to up-to-the-minute prices, shown in Tables 8 and 9, were also compared with those of the random walk model, shown in Table 10. The fixed training SVR results for up-to-the-minute prices reveal that this model does not have superior predictive power to a random prediction model for almost any of the selected stocks. Except for the application of the linear kernel to the Chinese stock, 1432, all the errors produced in the SVR predictions, measured by RMSE, were greater than the errors produced by the random

Table 10
Random walk model results for daily closing and up-to-the-minute prices.

Stock	Daily prices		Up-to-the-minute prices	
	RMSE	MAPE	RMSE	MAPE
BBAS3	0.06982	0.31553	0.01214	0.13480
PETR4	0.06299	0.41335	0.01511	0.57371
VALE5	0.05713	0.36126	0.01454	0.12852
ALPA4	0.05624	0.76789	0.01976	0.18754
DIRR3	0.09596	0.74089	0.02238	0.13788
LEVE3	0.07263	0.10570	0.01681	0.16462
BAC	0.04598	0.45303	0.00832	0.09031
GOOGL	0.03656	0.10691	0.00484	0.01745
XOM	0.05310	0.40824	0.00444	0.11338
ANGI	0.09856	0.47464	0.00712	0.01068
HL	0.11783	0.77834	0.01443	0.08547
PZZA	0.02499	0.13904	0.00866	0.12760
600028	0.06476	0.71341	0.02195	0.23547
601318	0.08782	0.33966	0.00773	0.04590
601668	0.07430	0.29177	0.01767	0.10858
1432	0.17248	0.69774	0.01621	0.02047
1970	0.25802	0.57245	0.02284	0.21137
2030	0.08892	0.26102	0.02759	0.15310

walk model for the up-to-the-minute closing prices with fixed training and test periods. In the case of a fixed training period, therefore, the SVR model's predictive power is only evidenced for daily prices, while the model is ineffective for higher frequency, up-to-the-minute periods.

4.2. Moving training window

Having recorded the closing price prediction errors for the SVR models with fixed training and test sets, we turn to an examination of the strategy of constantly updating the models. In this study, the frequency selected for updating the models was the extreme case of an update made whenever a new price becomes available. Thus, for daily prices, the model was updated every day, and the next day's closing price served as a test observation. The daily prices of the 7 most recent days were selected for training, leaving the closing price on the 8th day for the prediction test. Similarly, for up-to-the-minute prices, the model was updated every minute, and the next minute's closing price served as a test observation. The previous 90 min' prices were used for training, leaving the 91st closing price for the prediction test. The SVR parameters, i.e., c, γ and d, were fixed as the optimum values obtained in Table 4 for daily prices, Table 5 for up-to-the-minute prices in the 3-months historical data period and Table 6 for the 2-years period.

The results obtained when periodically updating the model are recorded in Table 11 (for daily prices) and 12 (for up-to-the-minute prices in the 3-months period). The Brazilian cases selected for the 2-years 1-min price prediction study using the moving training window SVR strategy are reported in Table 13. These results were measured in RMSE and MAPE and compared with the results obtained by the random walk model. Therefore, comparing the results in Tables 10 and 11, smaller errors were observed in the SVR model predictions, using linear and radial kernels for virtually all selected stocks, regardless of country of origin or capitalisation, for daily prices. The exception was LEVE3 stock, which presented odd results, possibly due to data errors. Moreover, on average, errors measured by MAPE were reduced when the linear kernel was used.

Tables 12 and 10 allow for a comparison of the SVR model with the random walk model for up-to-the-minute prices during the 3-months period proposed. In this case, attention is drawn to the errors obtained using the linear kernel for the US stocks, GOOGL, XOM and ANGI. In these cases, the SVR model had no predictive power for the up-to-the-minute prices selected for this study possibly due to data errors. However, for the other stocks, the linear kernel returned smaller errors than the random walk model. Use of the radial kernel returned smaller errors than the random model for all stocks, regardless of the country of origin or capitalisation, as shown in Table 2. Finally, the use of SVR with a polynomial kernel had predictive power only for the American small cap stock, PZZA.

Table 11
Results of SVR models updated periodically for daily prices.

Stock	Linear kernel		Radial kernel		Polynomial kernel	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
BBAS3	0.05687	0.25892	0.05894	0.26003	0.09175	0.44025
PETR4	0.05270	0.36937	0.05351	0.31293	0.20236	0.61956
VALE5	0.04403	0.24956	0.04783	0.22266	0.07257	0.32720
ALPA4	0.04652	0.74441	0.04960	0.46956	0.07247	0.96709
DIRR3	0.07727	0.52189	0.07677	0.52011	0.18988	0.97750
LEVE3	319.044	886.815	0.05974	0.10358	70.3598	185.093
BAC	0.03658	0.32727	0.03993	0.38629	0.11997	0.60985
GOOGL	0.02980	0.09312	0.02987	0.11268	0.05404	0.18005
XOM	0.04224	0.36616	0.04199	0.33754	0.10866	0.63326
ANGI	0.08268	0.35844	0.08508	0.37845	0.50205	1.59174
HL	0.09456	0.69319	0.09481	0.71560	0.23611	1.66588
PZZA	0.02067	0.10521	0.02126	0.08679	0.04162	0.21016
600028	0.05120	0.54468	0.05669	0.52282	0.09285	0.74604
601318	0.06868	0.23194	0.07748	0.26111	0.18146	0.76056
601668	0.05779	0.16590	0.06226	0.21833	0.14252	0.44786
1432	0.14203	0.56543	0.13364	0.54647	0.66641	1.30121
1970	0.22315	0.68089	0.22564	0.65871	0.49861	1.53855
2030	0.06776	0.17963	0.07659	0.23437	0.22052	0.52482

Table 12
Results of SVR models updated periodically for up-to-the-minute prices (3-months cases).

Stock	Linear kernel		Radial kernel		Polinomial kernel	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
BBAS3	0.00787	0.08743	0.01083	0.11790	0.02364	0.25295
PETR4	0.00970	0.40789	0.01369	0.59740	0.01806	0.69828
VALE5	0.00926	0.08162	0.01317	0.12653	0.01756	0.15890
ALPA4	0.01187	0.10699	0.01687	0.16579	0.03964	0.26407
DIRR3	0.01300	0.08051	0.01751	0.11830	0.02477	0.15939
LEVE3	0.00973	0.09519	0.01309	0.11936	0.01870	0.16123
BAC	1656531	7910363	0.00916	0.09557	2.2e+25	1.5e+26
GOOGL	617306.3	722994.3	0.00497	0.01878	3.7e+22	4.3e+22
XOM	8163009	83056969	0.00431	0.06861	1.8e+37	2.5e+38
ANGI	4853022	12952911	0.00708	0.01052	5.8e+24	1.6e+25
HL	0.00773	0.04365	0.01388	0.07773	43769951	190231018
PZZA	0.00419	0.06079	0.00746	0.10127	0.00419	0.06071
600028	0.01321	0.14705	0.01586	0.17251	0.10413	0.60028
601318	0.00475	0.03128	0.00753	0.05383	0.01444	0.07176
601668	0.01127	0.07303	0.01610	0.10458	0.02678	0.15542
1432	1.92099	4.00243	0.01150	0.01531	830235	1.7e+12
1970	0.01208	0.10058	0.01729	0.14582	0.06873	0.49109
2030	0.01259	0.07017	0.01811	0.10148	0.04341	0.21062

Table 13
Results of SVR models updated periodically for up-to-the-minute prices (2-years cases).

Stock	Linear kernel		Radial kernel		Polinomial kernel	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
BBAS3	0.00369	0.03717	0.00542	0.04813	0.01119	0.19015
PETR4	0.00416	0.02895	0.00598	0.04254	0.01224	0.09986
VALE5	0.00292	0.00797	0.00456	0.01167	0.00904	0.02352
ALPA4	0.00619	0.04543	0.00931	0.07025	0.01824	0.13474
DIRR3	0.01983	0.13016	0.02756	0.17858	0.06632	0.32251
LEVE3	0.01728	0.08113	0.02326	0.11044	0.04923	0.23288

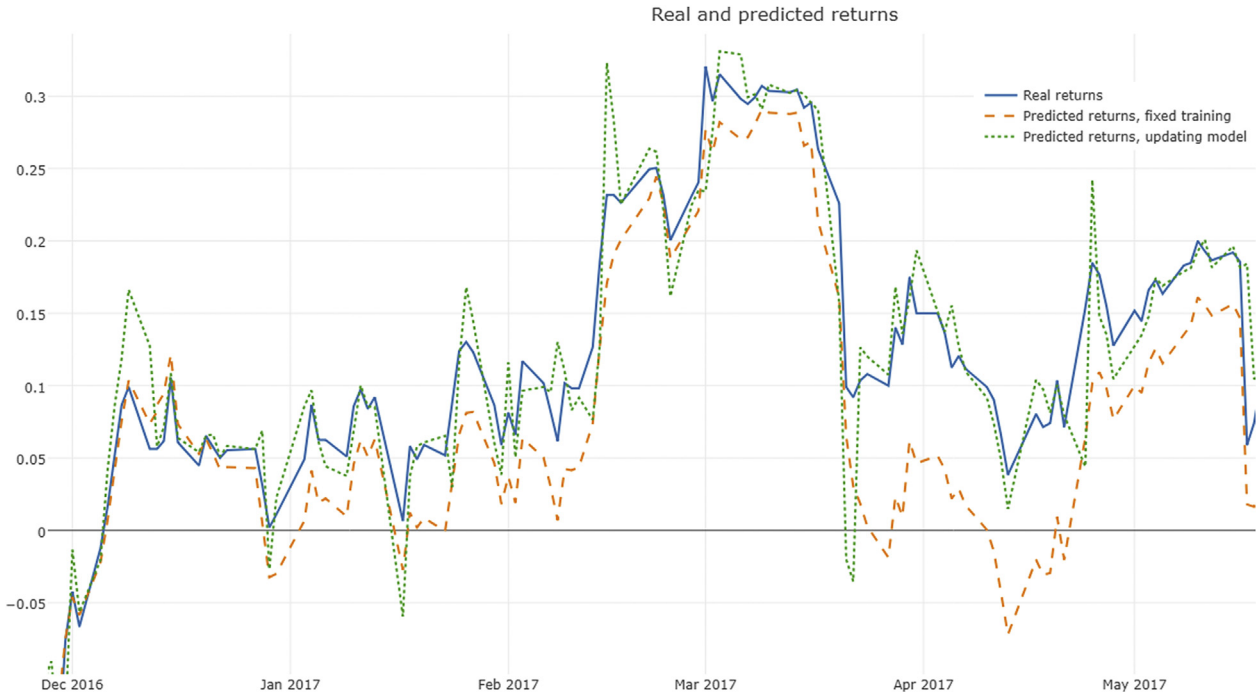

Fig. 2. Real returns and those predicted by the SVR model for the American stock BAC daily prices. The continuous curve represents returns observed in the market, the dashed curve represents the returns predicted by the fixed training model and finally the dotted curve represents the returns predicted by the moving training window model.

Results of 1-min price prediction in a training moving window fashion for Brazilian stock for a 2-years period are shown in Table 13. They are used as indicative SVR prediction power over the long run in the up-to-the-minutes prices timeframe. Comparing those results with the predictions obtained with the random walk model of Table 10, SVR updated regularly confirms its predictive power observed in the 3-months cases for almost all stocks, specially using the linear kernel. LEVE3 stands out as having results worse than using the random walk model. To illustrate SVR price prediction capabilities using both strategies, fixed trained model and moving training window, the daily returns for the American BAC stock are plotted in Fig. 2 as a continuous line. The returns for the stock are shown for roughly 5 months. The returns predicted by a fixed trained SVR model are plotted as a dashed line, whereas a dotted line represents the returns predicted by a moving training window SVR model. Both models track the real returns without the typical lag present in most technical indicators used alone. However, the moving training window SVR model tracks the real returns more closely in general, resulting in smaller RMSE for most of the curve.

4.3. SVR prediction models and stocks volatility

After registering SVR prediction results and comparing them with the random walk model-generated predictions in the previous paragraphs, this section verifies possible relationships between stocks basic statistics and the predictions themselves. Specifically, the daily SVR predictions using the moving training window are split yearly for evaluation against prices statistics. Two measures are examined for each of the 15-years historical prices per stock: average returns and volatility, calculated as the standard deviation of closing prices. Then we tabulate the correlations between yearly volatility and RMSE values obtained by SVR prediction models. Correlations are also calculated between the RMSE values and average daily return. Results are given in Table 14.

Examining the correlation values from average returns and RMSE errors, i.e. the left-most columns of Table 14, it is not possible to draw any conclusions about trends and relationships. Most of those correlation values are close to zero, exception being BAC and 1970. The American blue ship stock presented a relatively negative correlation between

Table 14
Correlation values between RMSE errors resulting from SVR daily prediction using the moving training window strategy for each kernel and two stock prices statistics: average return and volatility (calculated as the standard deviation of closing prices).

Stock	Average return			Volatility		
	Linear	Radial	Polynomial	Linear	Radial	Polynomial
BBAS3	0.167	0.084	0.244	0.996	0.971	0.963
PETR4	−0.360	−0.390	0.076	0.973	0.988	−0.347
VALE5	−0.146	−0.186	−0.238	0.999	0.996	0.957
ALPA4	0.379	0.368	0.361	0.989	0.983	0.916
DIRR3	0.049	−0.153	0.117	0.834	0.671	0.431
LEVE3	0.395	0.084	0.399	−0.055	−0.261	−0.052
BAC	−0.807	−0.800	−0.561	0.998	0.996	0.780
GOOGL	0.130	0.322	0.225	0.992	0.957	0.850
XOM	−0.201	−0.164	−0.229	0.991	0.938	0.930
ANGI	−0.241	−0.146	−0.315	0.897	0.929	0.277
HL	−0.208	−0.271	−0.019	0.988	0.987	0.919
PZZA	0.238	0.192	0.394	0.998	0.993	0.932
600028	−0.001	−0.054	−0.038	0.994	0.982	0.599
601318	0.055	0.167	0.192	0.988	0.965	0.944
601668	−0.024	−0.002	0.078	0.982	0.996	0.924
1432	−0.612	−0.857	−0.533	0.953	0.928	0.959
1970	0.983	0.985	0.952	0.966	0.964	0.991
2030	0.411	0.282	0.621	0.975	0.968	0.581

average return and SVR predicted price. That can indicate SVR predictions, considering constantly updated models, may be more precise during periods with larger daily returns for the BAC stock only. In the case of Chinese small cap 1970, we refrain from any comment, since the historical period considered for that specific stock, as shown by Table 2, is too short in aggregate years data.

Although the average returns do not seem related to the SVR predictions, there are indications of a strong relationship between the prediction errors and volatility. That can be observed for almost all the cases in the right-most columns of Table 14. Specially for the SVR predictions using linear kernel, correlation values are close to one for virtually all stocks, exception being LEVE3, which seems to contain too many historical data inconsistencies. A closer look in our data for that stock reveals heavy occurrences of missing data. However, apart from that case, SVR predictions, considering constantly updated models using linear and radial kernels, seem to be more precise during periods with lower volatility in prices.

5. Conclusion

Developing predictive price models for the stock market is challenging, but it is an important task when building profitable financial market transaction strategies. Computationally intensive methods, using past prices, are developed to facilitate better management of market risk for investors and speculators. Of the machine learning techniques available, this study uses SVR and measures its performance on various Brazilian, American and Chinese stocks with different characteristics, for example, small cap or blue chip. The predictive variables are calculated using TA indicators on asset prices. The results show the magnitude of the mean squared errors for the three common kernels in the literature, using specific algorithm training strategies with different price frequencies of days and minutes. The results are contrasted with those of a random walk-based model.

This study shows that using a fixed training set on daily prices, it is possible to obtain smaller prediction errors in the test set than in the training set when using a linear kernel. Moreover, this kernel was more adequate for price predictions than the radial and polynomial kernels in the case of daily prices and fixed training models and outperformed the random model for some stocks classified as blue chips and small caps in the three studied countries. However, increasing the price frequency to minutes reduced the model's predictive power using a fixed training period. In particular, SVR obtained inferior predictive results relative to a random walk model for almost all stocks studied in up-to-the-minute prices, using fixed training, regardless of the adopted kernel function.

The periodically updated models provided important evidence. In these cases, the use of linear and radial kernels resulted in smaller errors that the random walk model for almost all daily stock prices. The only exception was a stock with a high missing data rate. Constant model updating was also beneficial in the up-to-the-minute price frequency, and SVR models with linear and radial kernels achieved better results than the random walk model when this strategy was used. To emphasize the stability of the predictions over the long run, we processed a 2-years up-to-the-minutes prices period for the selected Brazilian stocks, confirming better results with a constantly updated model. The analyses presented in this study suggest that periodically updating the SVR model reduces the mean square error compared to using a rigid model without periodic updating. This result contrasts with that of Hsu et al. (2016)[9]; who did not achieve better performance when using a sliding window on the training data.

An important contribution of this study is a comparison of price prediction results of the presented SVR models with those of the random walk model, according to which markets are unpredictable in the long term. In this respect, the results presented here show that some SVR models, with periodic or fixed updates, may achieve better than random predictive performance, especially with the use of the linear kernel. Another result which prompts further investigation is the indication of a strong relationship between SVR price prediction and volatility, considering a moving training window.

Importantly, despite the evidence of asset price predictability presented here, this article does not propose transactional strategies applicable to the stock market. The results therefore do not directly refute the EMH. Given that the focus of the analysis is not the identification of purchasing or sales strategies that allow for extraordinary gains, the study does not address issues such as transaction costs or portfolio risk levels.

As the focus of the study is the analysis of asset price prediction errors, it is possible to build risk management models using SVR-based estimates. Exposure limits may be obtained by evaluating model errors. This study therefore provides a basis for the construction of systems that, while not directly evaluating the EMH, make possible the study of market efficiency and risk analysis. This study obtained results using SVR that were better than those of a null mean return random model.

Despite comparing daily rates with the use of high frequency up-to-the-minute trading, this paper considers only a predictive algorithm based on machine learning. Furthermore, the SVR model allows for testing of many kernel functions, while this study is limited to only the three most common in the literature. It should be noted that, for a more robust simulation of high frequency stock market strategies, it would be necessary to include transaction costs, communications network delays, differentiation between the market price and actual value of a purchase or sale transaction (slippage) and transaction liquidity.

This article has the limitation of not considering data quality assurance methods. Some results presented here suffer from poor data inputs and future studies should consider data treatment before usage. The present research approach equates missing minute prices to the previous values, not considering contiguous missing minutes or interpolating values. Some of the selected stocks illustrate the influence over the results of long streaks of missing data as well as outliers.

Another limitation of these research results is the length of the periods of historical prices considered, specially the 3-months up-to-the-minutes prices data. Although the 3-months period seem short compared to the 15 years daily historical prices data, it should be noted it contains nearly 33000 data points per stock, compared to the approximately 3700 data points for the daily period selected, posing a challenging processing task. For future reference, all processing of the 15 years daily data, 3-months 1-min data and the 2-years 1-min Brazilian stock data took about 15 h per stock in a powerful machine, with 24 processors type Intel® Xeon® CPU E5-2650 v4 @ 2,20 GHz with 226GB of RAM, running Linux 3.10.0, distribution CentOS 7.3. The implementations of SVR and related functions used in this research are scripted in the R® statistical language, version 3.4.1, using functions from *e1071* and *caret* packages. To reduce computation time, we took advantage of the parallel capabilities of the computer environment, allocating each stock simulations to an exclusive processor. Therefore, careful considerations are necessary for any real trading implementation attempts.

Future studies may include a larger number of test stocks and markets other than those selected here. Other predictive models could also be compared, including classifiers of the directions of asset prices. Independent variables may include other TA indicators, trend predictors or past prices. In addition, fundamental analysis indicators, such as company size, liquidity, indebtedness, profitability and activity measures, could be included. The inclusion of such data could improve the machine learning mechanism. It is also recommended that other model updating periodicities be tested, especially those with higher frequency than up-to-the-minute prices.

References

1. Araújo R, Oliveira AL, Meira S. A hybrid model for high-frequency stock market forecasting. *Expert Syst Appl*. 2015;42(8):4081−4096.
2. Ballings M, Van den Poel D, Hespeels N, Gryp R. Evaluating multiple classifiers for stock price direction prediction. *Expert Syst Appl*. 2015;42(20):7046−7056.
3. Barak S, Modarres M. Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Syst Appl*. 2015;42(3):1325−1339.
4. Brownlees C, Gallo G. Financial econometric analysis at ultra-high frequency: data handling concerns. *Comput Stat Data Anal*. 2006;51(4):2232−2245.
5. Choudhury S, Ghosh S, Bhattacharya A, Fernandes KJ, Tiwari MK. A real time clustering and SVM based price-volatility prediction for optimal trading strategy. *Neurocomputing*. 2014;131(1):419−426.
6. Dash R, Dash PK. A hybrid stock trading framework integrating technical analysis with machine learning techniques. *J Finance Data Sci*. 2016;2(1):42−57.
7. Gerlein EA, McGinnity M, Belatreche A, Coleman S. Evaluating machine learning classification for financial trading: an empirical approach. *Expert Syst Appl*. 2016;54(1):193−207.
8. Goldstein MA, Kumar P, Graves FC. Computerized and high-frequency trading. *Financ Rev*. 2014;49(2):177−202.
9. Hsu M-W, Lessmann S, Sung M-C, Ma T, Johnson JE. Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Syst Appl*. 2016;61(1):215−234.
10. Huang C-L, Tsai C-Y. A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Syst Appl*. 2009;36(2):1529−1539.
11. Kao L-J, Chiu C-C, Lu C-J, Yang J-L. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. *Neurocomputing*. 2013;99(1):534−542.
12. Kim K. Financial time series forecasting using support vector machines. *Neurocomputing*. 2003;55(1−2):307−319.
13. Kumar D, Meghwani SS, Thakur M. Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. *J Comput Sci*. 2016;17(1):1−13.
14. Lee EJ. High frequency trading in the Korean index futures market. *J Futures Market*. 2013;35(1):31−51.
15. Lessmann S, Sung M-C, Johnson JEV. Towards a methodology for measuring the true degree of efficiency in a speculative market. *J Oper Res Soc*. 2011;62(12):2120−2132.
16. Li H, Sun J, Li J-C, Yan X-Y. Forecasting business failure using two-stage ensemble of multivariate discriminant analysis and logistic regression. *Expet Syst*. 2012;30(5):385−397.
17. Li S, Shiue W, Huang M. The evaluation of consumer loans using support vector machines. *Expert Syst Appl*. 2006;30(4):772−782.
18. Lu C-J, Lee T-S, Chiu C-C. Financial time series forecasting using independent component analysis and support vector regression. *Decis Support Syst*. 2009;47(2):115−125.
19. Malkiel BG. The efficient market hypothesis and its critics. *J Econ Perspect*. 2003;17(1):59−82.
20. Malkiel BG, Fama EF. Efficient capital markets: a review of theory and empirical work. *J Finance*. 1970;25(2):383−417.
21. Manahov V, Hudson R, Gebka B. Does high frequency trading affect technical analysis and market efficiency? And if so, how? Journal of International financial markets. *Inst Money*. 2014;28(1):131−157.
22. Nayak RK, Mishra D, Rath AK. A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices. *Appl Soft Comput*. 2015;35(1):670−680.
23. Patel J, Shah S, Thakkar P, Kotecha K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Syst Appl*. 2015a;42(1):259−268.
24. Patel J, Shah S, Thakkar P, Kotecha K. Predicting stock market index using fusion of machine learning techniques. *Expert Syst Appl*. 2015b;42(4):2162−2172.
25. Qu H, Zhang Y. A new kernel of support vector regression for forecasting high-frequency stock returns. *Math Probl Eng*. 2016;2016(1):1−9.
26. Tay FE, Cao L. Application of support vector machines in financial time series forecasting. *Omega*. 2001;29(4):309−317.
27. Timmermann A, Granger CW. Efficient market hypothesis and forecasting. *Int J Forecast*. 2004;20(1):15−27.
28. Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer; 1995.
29. Wolfe P. A duality theorem for non-linear programming. *Q Appl Math*. oct 1961;19(3):239−244.
30. Xiao Y, Xiao J, Lu F, Wang S. Ensemble ANNs-PSO-GA approach for day-ahead stock e-exchange prices forecasting. *Int J Comput Intell Syst*. 2013;6(1):96−114.
31. Yeh C-Y, Huang C-W, Lee S-J. A multiple-kernel support vector regression approach for stock market price forecasting. *Expert Syst Appl*. 2011;38(3):2177−2186.
32. Zbikowski K. Using Volume Weighted Support Vector Machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Syst Appl*. 2015;42(4):1797−1805.
33. Zhou L, Lai KK, Yen J. Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation. *Int J Syst Sci*. 2012;45(3):241−253.