Ouahilal *et al. J Big Data (2017) 4:31*
DOI 10.1186/s40537-017-0092-5

Journal of Big Data

RESEARCH Open Access

A novel hybrid model based on Hodrick–Prescott filter and support vector regression algorithm for optimizing stock market price prediction

Meryem Ouahilal[1*], Mohammed El Mohajir[2], Mohamed Chahhou[2] and Badr Eddine El Mohajir[1]

*Correspondence:
m.ouahilal@ieee.org
[1] Faculty of Science, Abdelmalek Essaadi University, Tetuan, Morocco
Full list of author information is available at the end of the article

Abstract

Predicting stock market price is considered as a challenging task of financial time series analysis, which is of great interest to stock investors, stock traders and applied researchers. Many machine learning techniques have been used in this area to predict the stock market price, including regression algorithms which can be useful tools to provide good performance of financial time series prediction. Support Vector Regression is one of the most powerful algorithms in machine learning. There have been countless successes in utilizing SVR algorithm for stock market prediction. In this paper, we propose a novel hybrid approach based on machine learning and filtering techniques. Our proposed approach combines Support Vector Regression and Hodrick–Prescott filter in order to optimize the prediction of stock price. To assess the performance of this proposed approach, we have conducted several experiments using real world datasets. The principle objective of this paper is to demonstrate the improvement in predictive performance of stock market and verify the works of our proposed model in comparison with other optimized models. The experimental results confirm that the proposed algorithm constitutes a powerful model for predicting stock market prices.

Keywords: Stock price prediction, Financial time series forecasting, Business analytics, Support vector regression, Noise filtering techniques, Hodrick–Prescott filter, Decision support

Introduction

This paper addresses the issue of predicting stock market price in financial time series. Specifically we focus on the closing price which is the most up-to-date valuation of a security until trading commences again on the next trading day. The closing prices provide a useful marker for investors to evaluate changes in stock market prices over time.

A financial time series consists of various components equivalent to short-term irregular and seasonal variations, a medium-term business cycle, and long-term trend movement. Most macroeconomic analysis is concerned with a medium-term business cycle and long-term trend movement. However these fundamental movements are hidden in the original financial data because of multiple irregular and seasonal variations are dominant in the data [1].

Consequently, it is often difficult to read directly from the original data the fundamental movement of a financial variable under study. The financial time series includes some noise that may influence the information of the dataset. For better understanding and analysis of the data, and improve the accuracy of stock price prediction, noise filtering is necessary before using the predictive model.

On the other hand, there have been many studies using machine learning techniques to predict the stock market price [2]. A large number of successful applications have shown that regression algorithms, in particular, the support vector regression models, can be very useful tools for financial time series modeling and predicting [3, 4].

Support Vector Regression is one of the most powerful algorithms in machine learning. The theory has been developed over the last three decades by Vapnik, Chervonenkis and others [5–7]. There have been countless successes in utilizing SVR algorithm for stock market prediction. To mention a few, the author in [8] predicts future direction of stock price index using SVM model. In this study, he investigated the effect of the parameters in SVM. The goal was to find the optimal value of the parameters in order to improve the prediction results. The author also compared SVM with BPN and CBR. The experimental results showed that SVM outperformed BPN and CBR. The authors in [9] used support vector regression algorithm together with the independent component analysis (ICA) which is a statistical signal processing technique to implementing financial time series forecasting. The experimental results showed that their proposed model outperformed the SVR model without the ICA filtering. Recently, researchers in [10] predicted the stock market price of real world datasets using a hybrid model based on support vector regression and modified harmony search algorithm. The proposed method was tested on two sets of reliable financial datasets and experimental results on time series data showed that the proposed model improved accuracy of prediction compared to other optimization methods.

However, all these researchers have raised the problem of high noise in the financial time series, as well as some financial time series only possess poor information and small sample size. Predicting them with SVR directly is probably sensitive to the noise and may lead to overfitting.

In order to overcome these limitations, a novel predictive hybrid approach based on machine learning and filtering techniques has been proposed, which combines support vector regression algorithm and Hodrick–Prescott filter, for improving the prediction of stock market price by learning the historical data of real world data sets using our proposed framework. The predictive framework based on our proposed hybrid approach is shown in Fig. 1.

The significance and novelty of this paper are summarized as follows:

A novel efficient algorithm for stock market price prediction

We develop an efficient algorithm for predicting stock market price based on machine learning and noise filtering techniques. Our proposed algorithm is based on a hybrid approach which combines support vector regression algorithm and Hodrick–Prescott filter, in order to improve the performance of the stock market price predictions.

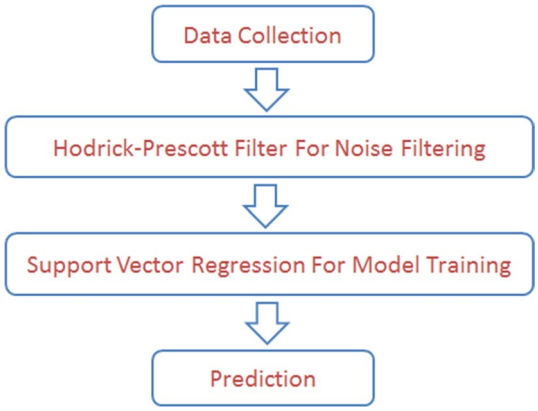

Fig. 1 Framework of stock market price prediction

Empirical demonstration of the effectiveness of our approach

We use real world datasets to compare our method with other existing financial time series predictive methods. To assess the performance of this proposed approach, we have conducted several experiments using real world datasets of different moroccan financial time series. The experimental results show that the proposed framework is a powerful predictive tool for stock market price.

The rest of this paper is organized as follow. Section II gives a short overview of financial time series. In the section III we present different filtering techniques used in economic field including: Hodrick–Prescott filter, Christiano Fitzgerald filter and Baxter king filter, and how to use them for noise filtering. Section IV provides a brief theoretical overview of some machine learning techniques which are regressive predictive algorithms including: Decision Tree Regression, Multiple Linear Regression and Support Vector Regression. Section V presents our methodology. Section VI focuses on prediction of stock market prices using our hybrid approach. Section VII evaluates our proposed model by conducting additional experiments on eight different financial time series with different sizes. Section VIII discusses the experimental results of the case study. Finally, section IX concludes this work and presents some direction for future research.

Related work: financial time series

A time series is a chronological sequence of observations on a particular variable. Usually the observations are taken at regular intervals (days, months, years), but the sampling could be irregular. A time series analysis consists of two phases:

- Constructing a model that represents a time series.
- Using the model to predict future values.

If a time series has a regular pattern, then a value of the series should be a function of previous values. If X is the target value that we are trying to model and predict, and X_t is the value of X at time t, then the goal is to create a model of the form:

$$X_t = f(X_{t-1}, X_{t-2}, X_{t-3}, ..., X) + e_t \quad (1)$$

Where X_{t-1} is the value of X for the previous observation, X_{t-2} is the value two observations ago, etc., and e_t represents noise that does not follow a predictable pattern (this is called a random shock). Values of variables occurring prior to the current observation are named lag values. If a time series follows a repeating pattern, then the value of X_t is usually highly correlated with $X_{t\text{-cycle}}$ where cycle is the number of observations in the regular cycle. For example, monthly observations with an annual cycle often can be modeled by

$$X_t = f(X_{t-12}). \tag{2}$$

The goal of building a time series model is the same as the goal for other types of predictive models which is to construct a model such that the error between the predicted value of the target variable and the observed value is as small as possible.

Time series prediction is one of the most basic predictive analytics needs of several businesses. Many data elements are observed as time series. These may be product sales, stock market prices and so on. From a strategic perspective, managers and decision makers will regularly need to be able to predict trends and seasonal patterns for these elements.

All predicting time series techniques can be divided into two board categories: Qualitative and Quantitative [11].

- Qualitative techniques refer to a number of forecasting approaches based on subjective estimates from informed experts. Usually, no statistical data analysis is involved. Rather, estimates are based on a deliberative process of a group of experts, based on their past knowledge and experience. Examples are the Delphi technique and scenario writing. These approaches are useful when good data are not available, or we wish to gain general insights through the opinions of experts.
- Quantitative Techniques refer to forecasting based on the analysis of historical data using mathematical and statistical principles and concepts. The quantitative forecasting approach is further sub-divided into two parts: causal techniques and time series techniques.

  - Causal techniques are based on regression analysis that examines the relationship between the variable to be forecasted and other explanatory variables.
  - Time Series techniques usually use historical data for only the variable of interest to forecast its future values (see Table 1).

In this research work, we have implemented time series predictions using a causal technique which is the most sophisticated kind of forecasting tool. It expresses mathematically the relevant causal relationships between the factor to be predicted and other factors. It may also directly incorporate the results of a time series analysis.

The causal techniques will be more developed in the "Predictive analysis technique" of this research work.

An economic time series consists of several components corresponding to short-term irregular and seasonal variations, a medium-term business cycle, and a long-term trend movement. Most macroeconomic analysis is concerned with a medium-term business

Table 1 Time series forecasting techniques

Categories	Application	Specific techniques
Qualitative techniques	Useful when historical data are scare or non-existent	Delphi technique Scenario writing Visionary forecast Historic analogies
Causal techniques	Useful when historical data are available for both the dependent (forecast) and the independent variables	Regression models Econometric models Leading indicators Correlation methods
Time Series techniques	Useful when historical data exists for forecast variable and the data exhibits a pattern	Moving average Autoregression models Seasonal regression models Exponential smoothing Trend projection Cointegration models

cycle and a long-term trend movement. However these fundamental movements are hidden in the original economic data because multiple irregular and seasonal variations are dominant in the data [12, 13].

Therefore it is often difficult to read directly from the original data the important movement of an economic variable under study. The financial time series contains some noise that may influence the information of the dataset. For better understanding and analysis of the trend, and improve the accuracy of stock price prediction, noise filtering is necessary before using the predictive model.

Noise filtering techniques

It is expected for a time series to contain some noise that may influence the whole information of the dataset. In the stock market, the volume of stocks vary every day and don't show any signs for prediction in the stock market, therefore resulting in difficulty to understand the trend of the change in it.

However, for a macroeconomic perspective of the stock market, the long-term trend should be predicted and analyzed. Although this long-term trend cannot give an clear indication which specific stock will rise tomorrow, it reveals nonetheless the performance of the whole investment environment and to a certain extent gives important hints helping make decisions on the stock market [11, 14].

To better understand and analyze the trend, noise filtering is essential. In this research, we will evaluate three different noise filtering techniques and compare their effectiveness on the financial time series analysis.

Hodrick–Prescott filter

The Hodrick–Prescott filter is a mathematical tool used in macroeconomics, specifically in real business cycle theory, to eliminate the cyclical component of a time series from raw data.

It is used to obtain a smoothed-curve representation of a time series, one that is more sensitive to long-term than to short-term fluctuations. The adjustment of the sensitivity of the trend to short-term fluctuations is achieved by modifying a multiplier λ.

The objective of Hodrick–Prescott filter is to decompose the time series into several series with common frequencies. Let y_t be our data at a specific time t. we want to decompose the data into growth component, τ_t, and the cyclical component, c_t

$$y_t = \tau_t + c_t \quad \text{for } t = 1,\ldots,T \tag{3}$$

One advantage of the Hodrick–Prescott filter is that it may be applicable to non-stationary time series, which is a relevant concern for many macroeconomic and financial time series.

The HP filter removes a smooth trend τ_t from a time series y_t by solving the minimization problem.

$$min \sum_{t=1}^{T} \left[\left(y_t - \tau_t\right)^2 + \lambda((\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}))^2 \right] \tag{4}$$

With respect to τ_t. The residual, or deviation from trend $z_t = y_t - \tau_t$ is commonly referred to as the business cycle component, and is the object of economic interest. In this sense the HP filter is a highpass filter, removing the trend and returning high-frequency components in z_t.

The parameter λ penalizes fluctuations in the second differences of y_t, and must be specified by the user of the HP filter [15–17].

Baxter king filter

Baxter-King band pass filter is a method of smoothing the time series, which is a modification of the Hodrick–Prescott filter that provides wider opportunities for removing cyclical component from a time series.

The filter method consists of singling out the repeated component of a time series by setting the width for oscillations of periodic component. Baxter-King filter is a band pass filter that removes the cyclical component from the time series based on weighted moving average with specified weights.

Baxter and King proposed a real symmetric filter with a finite length, which is called the BK filter. Suppose that the filter has a finite length (2K + 1): K leads and K lags. Then, the weights are obtained by solving the following minimization problem:

$$\min \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| B\left(e^{-i\omega}\right) -^{\wedge} B^{KK}\left(e^{-i\omega}\right) \right|^2 d\omega \tag{5}$$

Where '*i*' indicates the imaginary number. The solution gives filter weights:

$$B_0 = \frac{b-a}{\pi}, \quad B_j = \frac{\sin\left(jb\right) - \sin\left(ja\right)}{\pi j}, -K \le j \le K \tag{6}$$

The BK filter weights are obtained by the following normalization:

$$a_j = B_j - \frac{\sum_{h=-K}^{K} B_h}{2K+1} \tag{7}$$

Then, the approximation with T observations is computed as

$$y_t = \sum_{h=-K}^{K} a_h x_{t-h}, \quad K+1 \le t \le T-K \tag{8}$$

The result of smoothing is the source series with removed seasonal cyclic component.

A generalized Baxter-King filter is applied to non-stationary time series. Non-stationarity is accounted for in the matrix of weights that depend on the observation number in generalized model [18, 19].

Christiano Fitzgerald filter

The Christiano–Fitzgerald random walk filter is a band pass filter that was constructed on the same principles as the Baxter and King (BK) filter. These filters formulate the detrending and smoothing problem in the frequency domain. Should we have continuous and/or infinitely long time series the frequency filtering could be an exact procedure. However the granularity and finiteness of real world time series do not permit perfect frequency filtering. Both the BK and CF filters approximate the ideal infinite band pass filter. The Baxter and King version is a symmetric approximation, with no phase shifts in the resulting filtered series. But symmetry and phase correctness comes at the expense of series trimming. Depending on the trim factor a certain number of values at the end of the series cannot be calculated. There is a trade-off between the trimming factor and the precision with which the optimal filter can be approximated. On the other hand the Christiano–Fitzgerald random walk filter uses the whole time series for the calculation of each filtered data point. The advantage of the CF filter is that it is designed to work well on a larger class of time series than the BK filter, converges in the long run to the optimal filter, and in real time applications outperforms the BK filter.

The CF filter has a steep frequency response function at the boundaries of the filter band (i.e. low leakage); it is an asymmetric filter that converges in the long run to the optimal filter. It can be calculated as follows:

$$c_t = B_0 y_t + B_1 y_{t+1} + \cdots + B_{T-1-t} y_{T-1} + B_{T-t} y_T + \cdots + B_{t-2} y_2 + B_{t-1} y_1 \tag{9}$$

where

$$B_j = \frac{\sin(jb) - \sin(ja)}{\pi j}, j \ge 1 \quad and \quad B_0 = \frac{b-a}{\pi}, a = \frac{2\pi}{P_u}, b = \frac{2\pi}{P_l}, B_k = -\frac{1}{2}B_0 - \sum_{j=1}^{k-1} B_j \tag{10}$$

The parameters P_u and P_l are the cut-off cycle length in month. Cycles longer than P_l and shorter than P_u are preserved in the cyclical term c_t [20, 21].

Predictive analytics techniques

Predictive modelling is the process by which a model is created to predict an outcome. If the outcome is categorical it is called classification and if the outcome is numerical it is called regression. Descriptive modelling or clustering is the assignment of observations into clusters so that observations in the same cluster are similar. Finally, association rules can find interesting associations amongst observations. Figure 2 represents all the existing predictive analytics techniques classified by four categories.

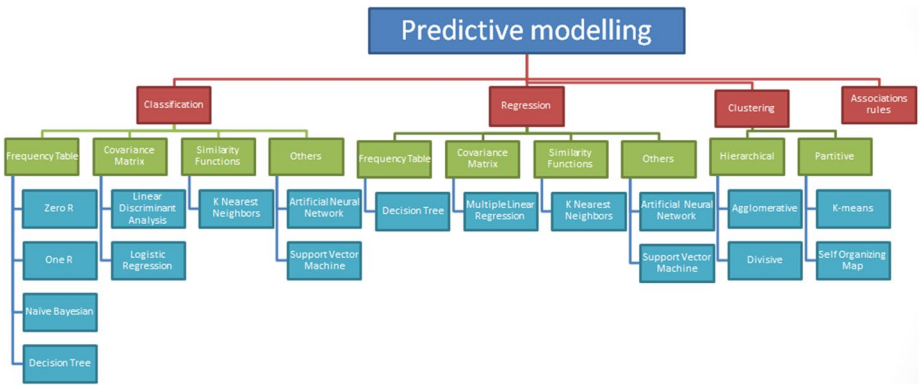

Fig. 2 Predictive modelling algorithms

Predictive analytics determines what is likely to happen in the future. This analysis is based on machine learning and statistical techniques as well as other more recently developed techniques that fall under the general category of data mining. The objective of these techniques is to be capable to provide predictions and forecasts about the future of the businesses activities.

There have been many studies using machine learning techniques to predict the stock market price. A large number of successful applications have shown that regression algorithms can be very useful tools for financial time series modeling and forecasting [2, 22, 23].

Regression is a data mining function that predicts a number. A regression task begins with a data set in which the target values are known. Regression models are tested by computing various statistics that calculate the difference between the predicted values and the expected values. The historical data for a regression project is typically divided into two data sets: one for building the model, the other for testing the model.

Regression modelling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, and environmental modelling. There are different families of regression algorithms and different ways of measuring the error [23, 24].

Decision tree regression

Decision tree builds regression or classification models in the form of a tree structure. It decomposes a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The highest decision node in a tree which corresponds to the best predictor named root node. Decision trees can manipulate both categorical and numerical data.

The core algorithm for building decision trees named ID3 by Quinlan which engages a top-down, greedy search through the space of possible branches with no backtracking. The ID3 algorithm can be used to build a decision tree for regression by changing Information Gain with Standard Deviation Reduction.

A decision tree is built top-down from a root node and includes partitioning the data into subsets that contain instances with similar values (homogenous). We use standard deviation to calculate the homogeneity of a numerical sample. If the numerical sample is totally homogeneous its standard deviation is zero.

$$S = \sqrt{\frac{\sum(x-\mu)^2}{n}} \tag{11}$$

$$S(T,X) = \sum_{c \in X} P(c)S(c) \tag{12}$$

Equation (11) represents the standard deviation of one attribute and Eq. (12) represents the standard deviation of two attributes.

The standard deviation reduction is based on the reduction in standard deviation after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest standard deviation reduction (i.e., the most homogeneous branches) [24, 25].

Multiple linear regression

A regression model is a compact mathematical representation of the relationship between the response variable and the input parameters in a given design space. Linear regression models are commonly used to obtain estimates of parameter significance as well as forecasting of the response variable at random points in the design space. One of the simpler forms of such models is

$$y = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \varepsilon \tag{13}$$

where y is the dependent or response variable, $\{xi|\ 1 \leq i \leq m\}$ are the independent or regressor variables and ε is the residual - the error due to lack of fit. β_0 is interpreted as the intercept of the response surface with the y-axis and $\{\beta i|\ 1 \leq i \leq m\}$ are known as the partial regression coefficients. The coefficient values represent the expected change in the response y per unit change in x_i and indicate the relative significance of the corresponding terms. It is frequent the case that the regressor variables interact i.e. the effect of a change in x_i on y depends on the value of x_j.

In such cases, the simple model in Eq. 13 is not sufficient. It is necessary to introduce terms that explicitly model two-factor interactions as shown below.

$$y = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m} \sum_{j=1+1}^{m} \beta_{i,j} x_i x_j + \varepsilon \tag{14}$$

Equation 15 represents a generic model that includes three-factor, four-factor and all higher order interactions. There are 2^m terms in this model and an equal number of unknown regression coefficients

$$y = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m} \sum_{j=1+1}^{m} \beta_{i,j} x_i x_j + \sum_{i=1}^{m} \sum_{j=i+1}^{m} \sum_{k=j+1}^{m} \beta_{i,j,k} x_i x_j x_k + \cdots + \beta_{1,2,\ldots m} x_1 x_2 x_m. \tag{15}$$

The linear regression models we represent in this session can be represented as a sum of k terms from this complete linear model, developed in a generic form as.

$$y = \beta_0 + \beta_1 x_{i^1} + \beta_2 x_{i^2} \cdots + \beta_{k-1} x_{i^{k-1}} + \varepsilon \tag{16}$$

where each $x_i{}^j$ is a distinct term from the generic model, and can be single factor, two factor, three factor or of any higher order. The collection of terms chosen for a given linear model will be mentioned to as the model terms.

In matrix terms, Eq. 16 can be written as

$$y = X\beta + \varepsilon \tag{17}$$

where β is the vector of regression coefficients and X is the model matrix. The model matrix has columns corresponding to the regressor variables x_1, x_2, ..., x_m, columns for interaction terms of any order, and a column of one's defining the intercept [24, 26].

Support vector regression

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Support Vector Machine is one of the most powerful algorithms in machine learning. The theory has been developed over the last three decades by Vapnik, Chervonenkis and others. When support vector machines were used to solve the regression problem they were usually called support vector regression.

In SVR, the fundamental idea is to map nonlinearly the original data X into the high-dimensional feature space and then do linear regression in this feature space. Therefore, suppose a set of data

$$S = \left\{ (x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_l, y_l) \right\} \in (X \times Y)^l \tag{18}$$

is given, where $x_i \in X = R^n$ is the input vector, $y_i \in Y = R$ is the corresponding out value and l is the total number of data set, the Support Vector Regression function is

$$f(x) = w \cdot \varphi(x) + b \tag{19}$$

Where $\phi(x)$ is the nonlinear mapping function, w is the weight vector and b is the bias value. They can be assessed by minimizing the regularized risk function

$$R(w) = \frac{1}{2}||w||^2 + C \sum_{i=1}^{l} L_e\left(y_i, f(x_i)\right) \tag{20}$$

Where $\frac{1}{2}||w||^2$ is used as a measurement of function flatness. C is the punishment parameter, which determines the trade-off between the training error and the generalization performance, $L_e(y_i, f(x_i))$ is called the insensitive loss function which is defined as

$$L_e(y_i, f(x_i)) = \begin{cases} \left|y_i - f(x_i)\right| - \varepsilon, & \left|y_i - f(x_i)\right| \geq \varepsilon \\ 0, & \left|y_i - f(x_i)\right| < \varepsilon \end{cases} \tag{21}$$

Where $\left|y_i - f(x_i)\right|$ is the error of predicting value and ε is the loss function.

When the error of estimation is taken into account, introduction of two positive slack variables ζ and ζ^* is to represent the distance between the actual value and the corresponding boundary values.

Different kernel functions are nominated; in fact, there is different network structure in support vector machines. The selection of kernel function is important to the effectiveness of Support Vector Regression. However, there is no mature theory in the selection of kernel function of SVR [24, 27].

Methodology

In a previous research work, we have made a literature survey of predictive analytics for business decision support and we have run an empirical comparative study of predictive algorithms to provide financial time series predictions. Three regression algorithms have been conducted which are multiple linear regression, support vector regression and decision tree regression to evaluate their effectiveness of making forecasts. To assess the performance of the overall three algorithms, several experiments have been conducted using real world financial time series. According to our comparative study we could clearly conclude that the support vector regression is the best prediction algorithm that can be execute to provide the financial time series forecasting [24, 28]. The architecture of our proposed framework is shown in Fig. 3.

In this paper, we propose a novel hybrid approach based on the combination of the Hodrick–Prescott filter (HP) and the Support Vector Regression algorithm (SVR). Therefore, we propose a new Framework of financial time series prediction based on our hybrid approach [28].

The objective of this approach is to improve and optimize SVR model predictions with the help of HP filter that will parse and normalize our data by filter and remove all existing noise in our financial time series.

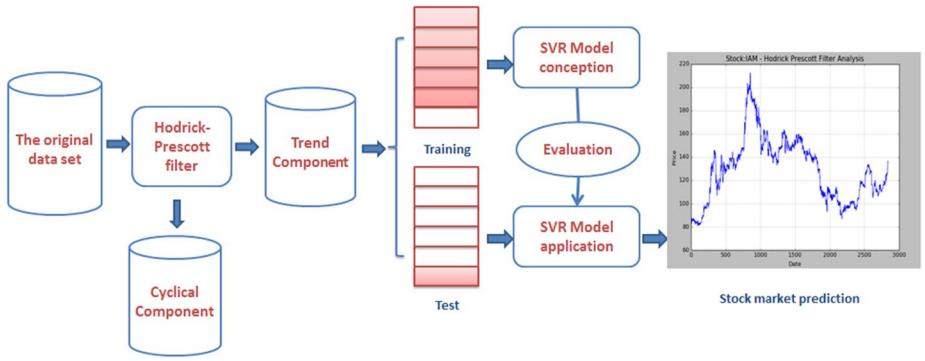

Fig. 3 Framework of stock price prediction using SVR model and HP filter

Stock market prediction

Dataset description

In the stock market, the closing price is the final price at which a security is traded on a given trading day. The closing price represents the most informed valuation of a security until trading commences again on the next trading day (Fig. 4).

The closing prices provide a useful marker for investors to assess changes in stock prices over time—the closing price of one day can be compared to the previous closing price in order to measure market sentiment for a given security over a trading day [29].

Thus, the closing price is selected as our prediction target of the original data set. The data were daily collected by IAM during the period from 2004 to 2016. Our data set has 6 attributes and 2840 samples with a size of 112 ko. They are Date, Open price, Close price, High price, Low price, and Volume. The goal is to predict Close price for different amount of time in the future.

Our hybrid approach

The regression analysis focuses on the Close price on the *(t + 1)-th* day changes when the Open price, Close price, High price, Low price and Volume on the *i-th* day vary.

Our objective is to fit the following relationship by regression analysis:

$$Close_{t+1} = f\left(Open_t,\ Close_t,\ High_t,\ Low_t,\ Volume_t\right)$$

Before performing the regression, we need to use Hodrick–Prescott filter to filter noise and normalize the data value on each attribute separately.

The goal of Hodrick–Prescott filter is to decompose the time series into several series with common frequencies. We want to decompose the data into the trend and the cyclical components.

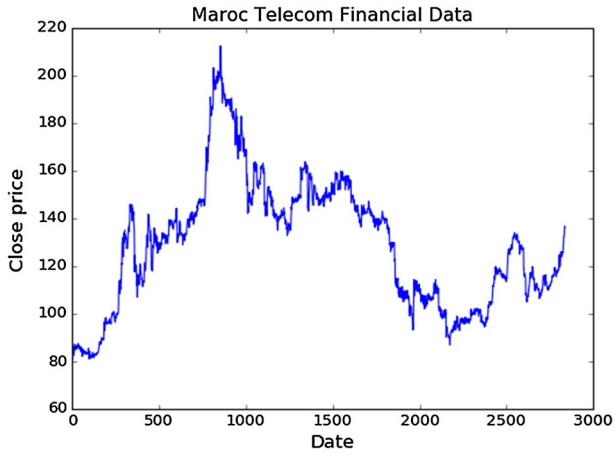

Fig. 4 Maroc Telecom Financial Tme Series

Above Fig. 5 shows the Stock IAM price and HP Filter components trend and cycle component. It's clearly visible that trend component is ultra-smooth and very good in predicting the future of IAM price direction. And the Cycle Component extreme values suggest a possible trend reversal.

Experimental results

The regression performances vary with different selection of four important parameters: (1) the kernel function (2) penalty parameter c (3) kernel parameter g and (4) degree of the kernel function d. Four-stage grid search is used to find the best combination of parameters for each filter.

In prediction experiments, the data are divided into two subsets. The data from December 2004 to December 2012 was employed as training set used for training the models of the algorithms.

We have selected four folds of testing set decomposed as follow:

- The data from January 2013 to December 2013 are employed as first fold of testing set.
- The data from January 2013 to December 2014 are employed as second fold of testing set.
- The data from January 2013 to December 2015 are employed as third fold of testing set.
- The data from January 2013 to July 2016 are employed as fourth fold of testing set.

Figure 6 shows the results of our regression by plotting the original data and regressive data together for different amount of time in the future.

Table 2 shows the kernel components value selected by the grid search that helps to produce good prediction result from this analysis.

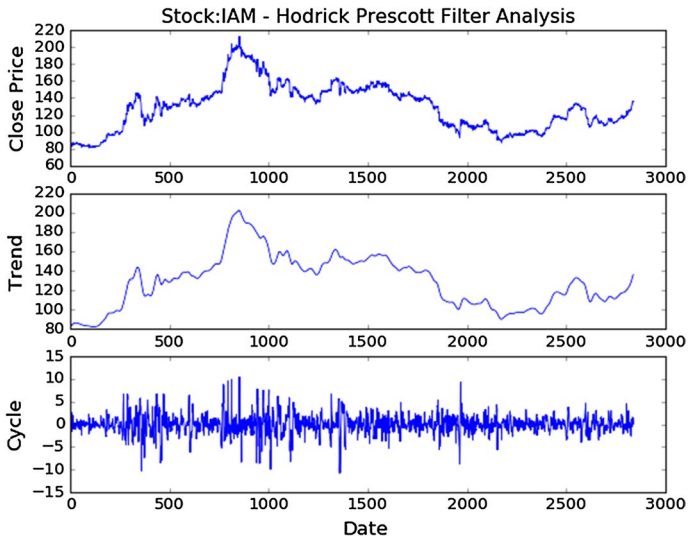

Fig. 5 Hodrick Prescott filter analysis

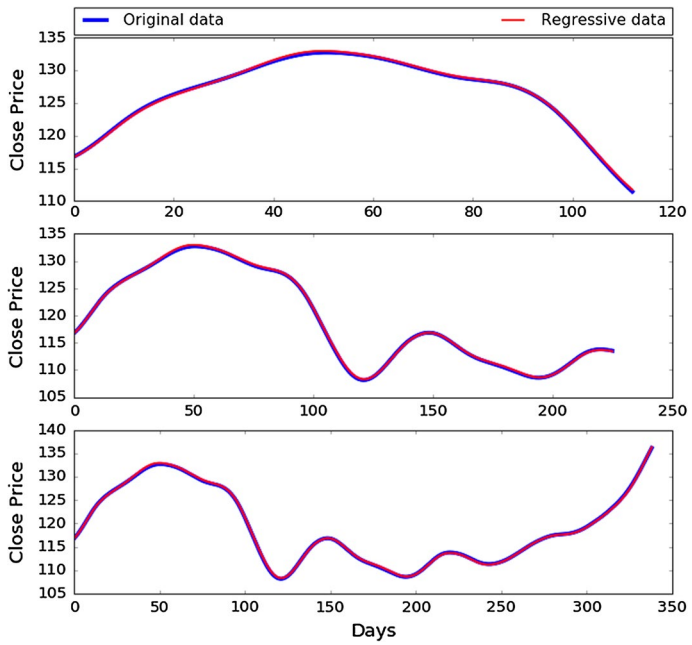

Fig. 6 Stock: IAM - Original data VS Regressive data

Table 2 Kernel parameters setting

Regressive model	Kernel	C	G	D
SVR	rbf	250	0.01	3
SVR + HP	rbf	275	0.1	3
SVR + CF	rbf	150	0.01	3
SVR + BK	rbf	250	0.1	3

The error rate is computed between the actual and predicted stock prices come from the experiments. To calculate the error rate, Mean average percentage error (MAPE) is used in this study.

It's defined in the following:

$$MAPE(y, y') = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - y_i'|}{|y_i|} \times 100\% \tag{22}$$

Where y' and y represent the predicted result and observed value respectively and N is the sum of training samples.

Figure 7 and Table 3 show the error (average MAPE) committed in different SVR models using different kinds of filters. The calculation of MAPE was done for the testing dataset and the training dataset.

Where HP is the Hodrick–Prescott filter, CF is the Christiano Fitzgerald filter and BK is the Baxter-King filter. These filters are the most known and used in financial time series analysis.

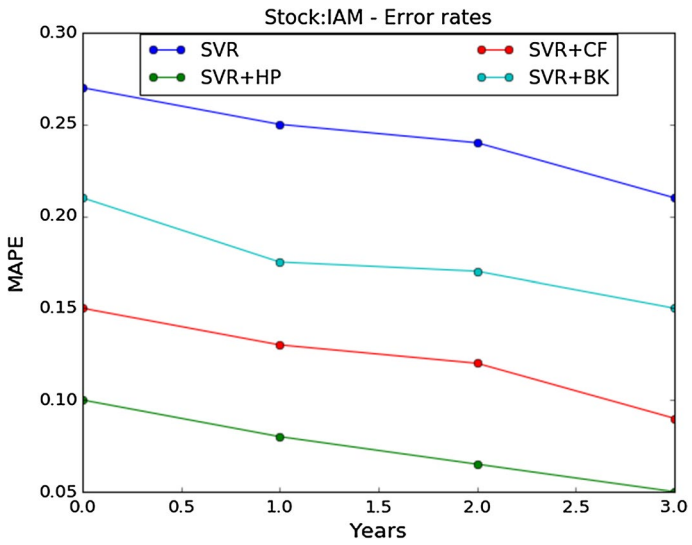

Fig. 7 MAPE (Error rate) of the predictive models

Table 3 MAPE (Error rate) of the predictive models

Years	SVR	SVR + HP	SVR + CF	SVR + BK
2013	0.27	0.10	0.15	0.21
2013–2014	0.25	0.08	0.13	0.175
2013–2015	0.24	0.065	0.12	0.17
2013–2016	0.21	0.05	0.09	0.15

Table 4 Data sets description

Dataset	Period	Attributes	Samples	Size (ko)
Addoha holding	10/2006–08/2017	6	2750	108
Auto haul	11/2007–08/2017	6	2021	71
BCP bank	07/2004–08/2017	6	3210	121
BMCE bank	07/2000–08/2017	6	4126	149
Delta holding	05/2008–08/2017	6	2227	82
Total macro	05/2015–08/2017	6	510	20
Wafa assurance	11/2000–08/2017	6	3250	122
CIH bank	09/2000–08/2017	6	3808	141

Evaluation of our model

In this section, we evaluate the effectiveness of our model by conducting additional experiments based on our hybrid approach on different financial time series with different sizes.

Datasets description

In order to evaluate our model and to prove its works, we conducted several experiments on eight different financial time series with different sizes (Table 4) (Fig. 8).

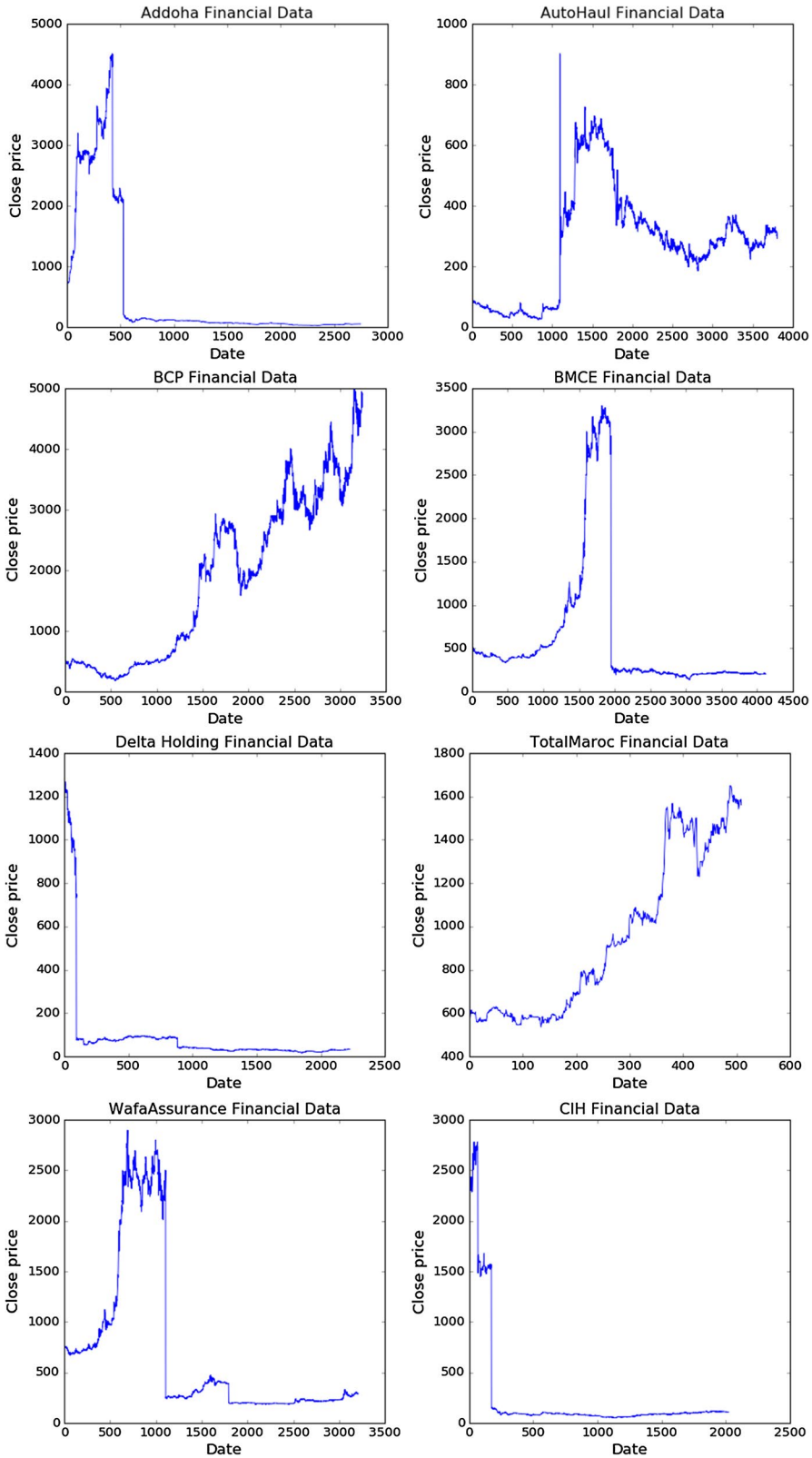

Fig. 8 8 different financial time series

The time series data were daily collected by different companies during different periods. Our datasets has six attributes. They are Date, Open price, Close price, High price, Low price, and Volume. The goal is to predict the Close price using 4 different models including: SVR, SVR + HP, SVR + CF and SVR + BK.

Methodology

To evaluate the performance of our model, we conduct several experiments on different financial time series using our hybrid approach which includes two steps:

- Decomposing the time series into trend and cyclical components using three different filter techniques.
- Implementing the regressive model using the support vector regression algorithm on the trend component of the time series.

The regression analysis focuses on the Close price on the (t + 1)-th day changes when the Open price, Close price, High price, Low price and Volume on the i-th day vary.

Our objective is to fit the following relationship by regression analysis:

$$\mathrm{Close}_{t+1} = f\left(\mathrm{Open}_t,\ \mathrm{Close}_t,\ \mathrm{High}_t,\ \mathrm{Low}_t,\ \mathrm{Volume}_t\right)$$

Experimental results

Before performing the regression, we need to use a different filter technique for each experiment to filter the noise and normalize the data values on each attribute separately. These filter techniques include: Hodrick Prescott filter, Christiano Fitzgerald filter, Baxter King filter.

The goal of these filters is to decompose the time series into several series with common frequencies. We want to decompose the data into the trend and the cyclical components

We have implemented our regressive models using Python as a programming language with different Python libraries and packages including: Scikit-learn and Statesmodels.

The regression performances vary with different selection of four important parameters:

- The kernel function f
- The penalty parameter c
- The kernel parameter g and
- The degree of the kernel function d.

Table 5 Kernel parameters setting

Regressive model	Kernel	C	G	D
SVR	rbf	175	0.01	3
SVR + HP	rbf	275	0.01	3
SVR + CF	rbf	250	0.1	3
SVR + BK	rbf	200	0.01	3

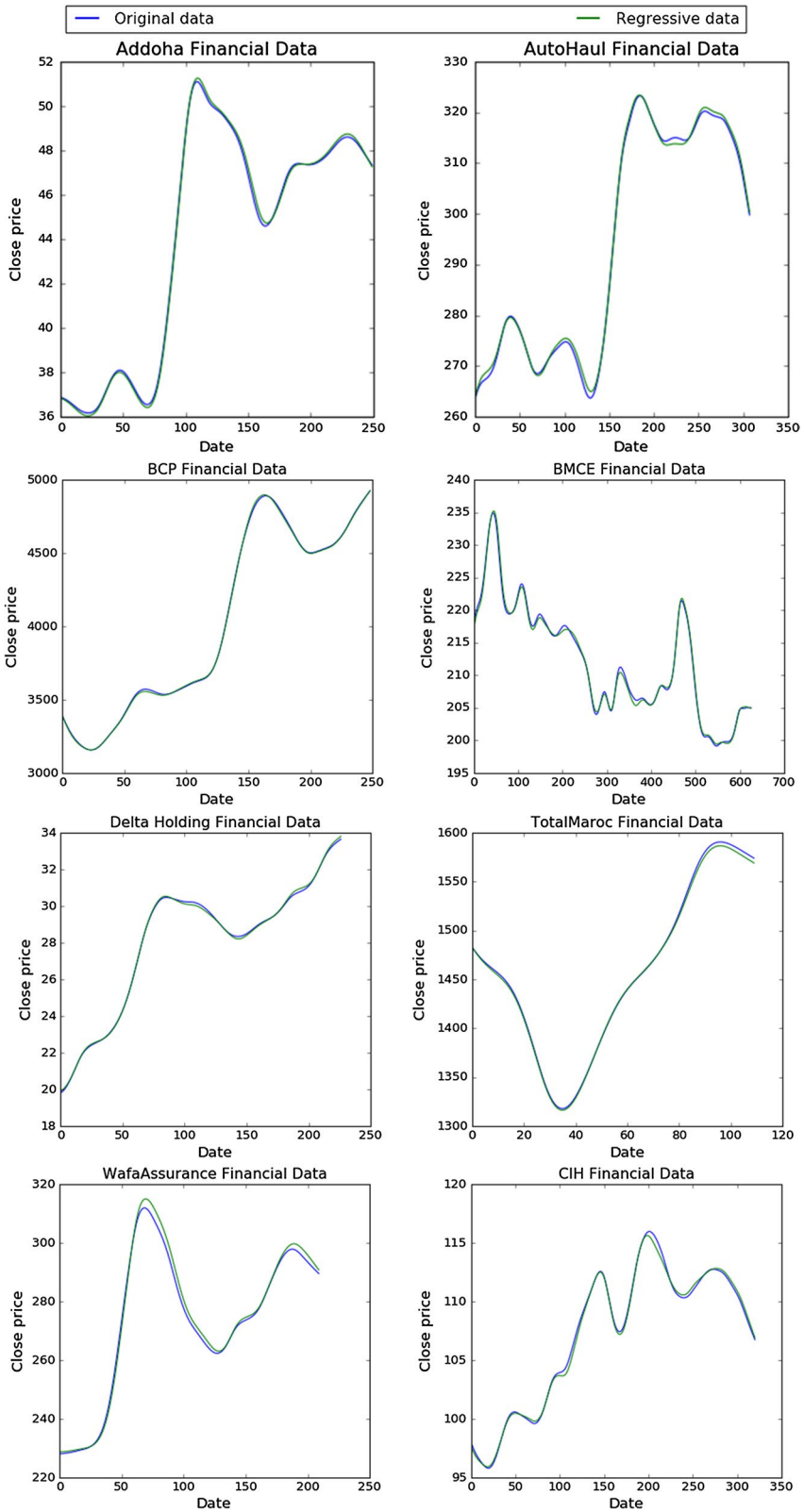

Fig. 9 Stock market price prediction: Original data VS Regressive data

One of the problems in using support vector regression model is to determine the parameter values of the proposed model. We have implemented a four-stage grid search to find the optimum values for each parameter and the best combination of parameters for each case. In prediction experiments, each dataset is divided into two subsets: training set and test set.

Table 5 shows the kernel components value selected for the first dataset by the grid search that helps to produce good prediction result from this analysis. Theses parameters differ from a dataset to another.

Figure 9 shows the results of our regression by plotting the original data and regressive data together for different time series.

The error rate is computed between the actual and predicted stock prices come from the experiments. To calculate the error rate, Mean average percentage error (MAPE) is used in this study.

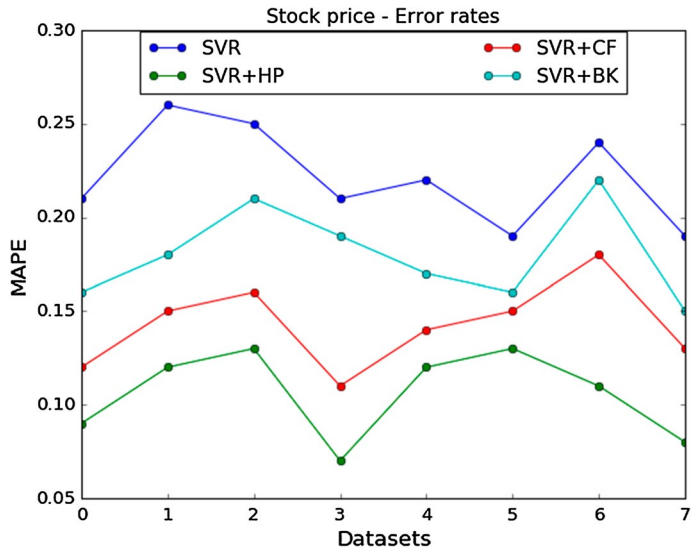

Fig. 10 MAPE (Error rate) of the predictive models

Table 6 MAPE (Error rate) of the predictive models

Dataset	Model			
	SVR	SVR + HP	SVR + CF	SVR + BK
Addoha holding	0.21	0.09	0.12	0.16
Auto haul	0.26	0.12	0.15	0.18
BCP bank	0.25	0.13	0.16	0.21
BMCE bank	0.21	0.07	0.11	0.19
Delta holding	0.22	0.12	0.14	0.17
Total macro	0.19	0.13	0.15	0.16
Wafa assurance	0.24	0.11	0.18	0.22
CIH bank	0.19	0.08	0.13	0.15

It's defined in the following:

$$MAPE(y, y') = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - y'_i|}{|y_i|} \times 100\% \quad (22)$$

Where y' and y represent the predicted result and observed value respectively and N is the sum of training samples.

Figure 10 and Table 6 show the error (average MAPE) committed in different SVR models using different kinds of filters. The calculation of MAPE was done for the testing dataset and the training dataset.

Where HP is the Hodrick–Prescott filter, CF is the Christiano Fitzgerald filter and BK is the Baxter-King filter. These filters are the most known and used in financial time series analysis.

Discussion of the experimental results

To assess the performance of our proposed approach we have conducted several experiments for predicting stock price with different Support Vector Regression models using different kind of filters.

The HP method is a two-sided filter capable of providing a smooth estimate of the long-term trend component of a series, as well as the corresponding cyclical component. The HP method minimizes the variance of a series around a parameter that approaches a linear trend [16, 17]. The CF and BK methods are both band-pass or frequency filters and are capable of isolating the cyclical component of a time series. These linear filters utilize a two-sided weighted moving average of the data in which the cycles, within some "band", are extracted and remaining cycles are filtered out [19–21].

After our filtering analysis we have found that the trend component produced by the Hodrick Prescott filter preserves the time series curve, unlike the trend component produced by the CF filter or the BK filter which tends to slightly modify the time series curve. That is why we have noticed the small difference in performance produced by these filters.

On the other hand, the prediction results change from a dataset to another due to the sample size and the time series structure. Therefore, we have found different optimum parameters of the support vector regression model suited to each case of study.

The objective was to verify that the combination of Support Vector Regression model and Hodrick–Prescott filter provide the best results of stock price prediction compared to the other filters.

Effectively, the combination of Support Vector Regression model and Hodrick–Prescott filter provide the best results since the MAPE error given by this model is the lowest among all proposed error rate.

According to our experimental results we can clearly conclude that the proposed framework using our hybrid approach which combines the Support Vector Regression model and the Hodrick–Prescott filter is a powerful predictive tool for stock market price and financial time series.

Conclusion and direction for future research

Predicting stock market prices is a major factor in stock market prediction and has been paid much attention. Therefore, the applications of regression model in financial field are a meaningful attempt.

In this research work, we proposed a novel hybrid approach of predicting stock price based on machine learning and filtering techniques which combines Support Vector Regression algorithm and Hodrick–Prescott filter in order to optimize the stock price prediction.

To assess the performance of this proposed approach, several experiments have been conducted using real world datasets. The objective was to verify that the combination of Support Vector Regression model and Hodrick–Prescott filter provide the best results of stock price prediction compared to the other filters. The experimental results show that compared with the SVR, SVR + CF and SVR + BK models, the proposed model is an effective method for predicting stock price, which greatly improves the accuracy of forecasting. Therefore we can confirm that the proposed model is a powerful predictive solution for the stock market prices.

Our proposed model provides a very good accuracy of stock market price prediction with a very minimalist execution time. The proposed model which combines the SVR model and the HP filter outperforms the standard SVR model and the other optimized model of SVR checked in this research work.

However, the stock market price not only depends on historical data but also greatly influenced by the macroeconomic factors and important news in the world. These limitations lead us to some problems to be solved.

Going forward, we have several additional avenues which we would like to explore. We plan to study the impact of the macroeconomic factors and some big news on the stock market performance in the Moroccan context. This study will allow us to identify the most relevant factors that can be incorporated into our predictive model in order to improve our financial prediction results. We also plan to test our methodology for different industries and check the results on different sets of real world data.

Authors' contributions
Meryem Ouahilal carried out the conception and design of the research, performed the implementations and drafted the manuscript. Mohammed El Mohajir, Mohamed Chahhou and Badr Eddine El Mohajir provided reviews on the manuscript. All authors read and approved the final manuscript.

Author details
[1] Faculty of Science, Abdelmalek Essaadi University, Tetuan, Morocco. [2] Faculty of Science, LIMS, Sidi Mohamed Ben Abdallah University, Fez, Morocco.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 17 July 2017 Accepted: 20 September 2017
Published online: 04 October 2017

References

1. Mahalakshmi G, Sridevi S, Rajaram S. A survey on forecasting time series data. In: Proceeding of the IEEE international conference on computing technologies and intelligent data engineering, Kovilpatti, India, 7–9 January 2016.
2. Anandhi V, Chezian RM. Support vector regression in forecasting. Int J Adv Res Comp Commun Eng. 2013;2(10):4148–51.
3. Badics MC. Stock market time series forecasting with data mining methods. Finan Econ Rev. 2014;13(4):205–25.
4. Li Z, Li Y, Yu F, Ge D. Adaptively weighted support vector regression for financial time series prediction. In: Proceeding of the IEEE international joint conference on neural networks. Beijing, China, 6–11 July 2014.
5. Vapnik V. The nature of statistical theory. NY: Springer; 2000.
6. Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. Neural information processing systems. Cambridge: MIT Press; 1997.
7. Basak D, Pal S, Chandra D. Support vector regression. Neural Inf Process-Lett Rev. 2007;11(10):203–24.
8. Kim KJ. Financial time series forecasting using support vector machines. Neurocomputing. 2003;55(1–2):307–19.
9. Lu CJ, Lee TS, Chiu CC. Financial time series forecasting using independent component analysis and support vector regression. Decis Support Syst. 2009;47(2):115–25.
10. Misaghi S, Sheijani OS. A hybrid model based on support vector regression and modified harmony search algorithm in time series prediction. In: The 5th IEEE Iranian joint congress on fuzzy and intelligent systems. Qazvin, Iran, 7–9 March 2017.
11. Brockwell PJ, Davis RA. Introduction to time series and forecasting. Berlin: Springer; 2016.
12. Varshney R, Mojsilovic A. Business analytics based on financial time series. IEEE Signal Process Mag. 2011;28(5):83–93.
13. Okkels CB. Financial forecasting: Stock market prediction. Master's thesis. Faculty of science, University of Copenhagen. 2014.
14. Higo M, Nakada SK. How can we extract a fundamental trend from an economic time series? IMES Discussion Paper Series (98-E-5). Bank of Japan: Institute for monetary and economic studies; 1998.
15. Hodrick RJ, Prescott EC, Postwar US. Business cycles: an empirical investigation. J Money Credit Bank. 1997;29(1):1–16.
16. Peter C, Phillips B, Jin S. Business cycles, trend elimination, and the HP filter. Cowles Fondation Discussion paper (2005). Cowles Fondation for Research in Economics, Yale University. 2015.
17. Robert de Jong M, Sakarya N. The Econometrics of the Hodrick–Prescott filter. Review of economics and statistics. Cambridge: The MIT Press Journals; 2015.
18. Baxter M, King RG. Measuring business cycles: approximate bandpass filters. Rev Econ Stat. 1999;81(4):575–93.
19. Baxter M, King RG. Measuring business cycles approximate band-pass filters for economic time series. The national Bureau of Economic Research (5022). 1995.
20. Christiano JL, Fitzgerald JT. The Band Pass Filter. NBER working paper series, 1999.
21. Nilson R, Gyomai G. Cycle Extraction: A comparison of the phase-average trend method, the Hodrick–Prescott and Christiano Fitzgerald filters, OECD statistical working papers. 2011.
22. Krollner B, Vanstone B, Finnie G. Financial time series forecasting with machine learning techniques: A survey. In: The proceeding of the European symposium on artificial neural networks—computational intelligence and machine learning. Bruge, Belgium. 28–30 April 2010. p. 25–30.
23. Ouahilal M, Jellouli I, El Mohajir M. A Comparative study of predictive algorithms for time series forecasting. In: The proceeding of the third IEEE international colloquium in information science and technology. Tetuan. 20–22 October 2014. p. 68–73.
24. Ouahilal M, El Mohajir M, Chahhou M, El Mohajir B. A comparative study of predictive algorithms for business analytics and decision support systems: finance as a case study. In: The Proceeding of the international conference on information technology for organizations development. Fez. March 30–April 1st 2016. p. 1–6.
25. Lai RK, Fan CY, Huang WH, Chang PC. Evolving and clustering fuzzy decision tree for financial time series data forecasting. Expert Syst Appl. 2009;36(2):3761–73.
26. Ismail Z, Yahya A, Shabri A. Forecasting gold price using multiple linear regression method. Am J Appl Sci. 2009;6(8):1509–14.
27. Meesad P, Rasel RI. Predicting stock market price using support vector regression. In: The proceeding of the international conference on informatics, electronics and vision. Dhaka. 17–18 May 2013. p. 1–6.
28. Ouahilal M, El Mohajir M, Chahhou M, El Mohajir B. Optimizing stock market price prediction using a hybrid approach based on HP filter and support vector regression. In: The proceeding of the IEEE international colloquium in information science and technology. Tangier. 24–26 October 2016. p. 290–294.
29. Tsay RS. Analysis of financial time series. Wiley Series in probability and statistics. New Jersey: Wiley; 2005.