Decision Support Systems 54 (2013) 1228–1244

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting

Ling-Jing Kao [a], Chih-Chou Chiu [a], Chi-Jie Lu [b,*], Chih-Hsiang Chang [c]

[a] Department of Business Management, National Taipei University of Technology, Taiwan
[b] Department of Industrial Management, Chien Hsin University of Science and Technology, Taiwan
[c] Institute of Commerce Automation and Management, National Taipei University of Technology, Taiwan

ARTICLE INFO

Article history:
Received 20 March 2012
Received in revised form 9 October 2012
Accepted 11 November 2012
Available online 20 November 2012

Keywords:
Stock index forecasting
Wavelet transform
Multivariate adaptive regression splines
Support vector regression
Feature extraction

ABSTRACT

Forecasting stock prices is a major activity of financial firms and private investors when they make investment decisions. Feature extraction is usually the first step of a stock price forecasting model development. Wavelet transform, used mainly for the extraction of information contained in signals, is a signal processing technique that can simultaneously analyze the time domain and the frequency domain. When wavelet transform is employed to construct a forecasting model, the wavelet basis functions and decomposition stages need to be determined first. However, because forecasting models constructed by different wavelet sub-series would exhibit different forecasting capabilities and yield varying forecast results, the selection of wavelet that can lead to an optimal forecast outcome is extremely critical in model construction. In this study, a new stock price forecasting model which integrates wavelet transform, multivariate adaptive regression splines (MARS), and support vector regression (SVR) (called Wavelet-MARS-SVR) is proposed to not only address the problem of wavelet sub-series selection but also improve the forecast accuracy. The performance of the proposed method is evaluated by comparing the forecasting results of Wavelet-MARS-SVR with the ones made by other five competing approaches (Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR and single ANFIS) on the stock price data of two newly emerging stock markets and two mature stock markets. The empirical study shows that the proposed approach can not only solve the problem of wavelet sub-series selection but also outperform other competing models. Moreover, according to the sub-series which are selected by the proposed approach, we can successfully identify the data of which sessions (or points in time) among past stock market prices exerted significant impact on the construction of the forecasting model.

1. Introduction

The stock market has long been one of the investment targets that investors pay most attention to. In order to gain profits or to avoid risks, investors usually conjecture the trend of stock indices and draw up strategies for future investments according to their anticipation. Therefore, the construction of an effective stock price forecasting model which can be used to reduce the personal bias and mistakes is the major concern for individual, institutional investors and researchers. For example, Atsalakis and Valavanis [4] and Bahrammirzaee [6] survey various computing methods in stock market forecasting, Chang and Fan [9] propose an integrated approach of wavelet and fuzzy system for stock price forecasting, Khansa and Liginlal [22] compare the performance of vector autoregression and time-delayer neural network in stock market forecasting, Lu et al. [26] use independent components analysis to perform feature extraction and support vector regression in financial time series forecasting, and Tsai and Hsiao [44] combine well-known feature selection methods, principal component analysis, genetic algorithms, and decision trees to identify more representative variables for better prediction. However, owing to the high-frequency, non-stationary and chaotic properties of the stock index data [19,24,48], a stock price forecasting model utilizing the original stock index data fails to provide satisfying forecast results. To solve this problem, before constructing a forecasting model, many studies would first utilize an information extraction technique to extract features (also called latent signals) contained in data, then use these extracted characteristics to construct the forecasting model [9,24,26,35].

Wavelet transform, used mainly for extracting information contained in signals, is a type of signal processing technique that can simultaneously analyze the time domain and the frequency domain. Conceptually, wavelet transform uses wavelet basis functions to decompose the original sign into different sub-series and highlight the eigenvalues hidden in the original signal. Traditionally, wavelet transform is primarily applied to image processing or signal processing

[*] Corresponding author at: Department of Industrial Management, Chien Hsin University of Science and Technology, Zhong-Li City, Taoyuan County 32097, Taiwan, R.O.C. Tel.: +886 3 4581196x6712; fax: +886 3 4683298.
E-mail addresses: jerrylu@uch.edu.tw, chijie.lu@gmail.com (C.-J. Lu).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2012.11.012

[36,39,43]; however, with its powerful feature extraction capability, wavelet transform has now been successfully applied to time-series studies [2,7,12,17,18,20,24,32,33,37].

For example, Bjorn [7] employed the wavelet transform to obtain financial time series featuring fractal and chaotic characteristics, and utilized these resulting series as input factors to carry out forecasting. Pan and Wang [32] adopted wavelet transform as the estimator of random, non-linear regression and incorporated it into the state space model to carry out an empirical study on the S&P 500 Index, discovering the forecastability of the stock market. Gonghui et al. [18] used redundant Haar wavelet transform, together with the dynamic recurrent neural network, to construct the forecasting model, obtaining experimental results that showed the wavelet transform's capability to effectively increase the forecast accuracy of the neural network. Furthermore, Shin and Han [38] adopted the genetic algorithm to optimize the multi-resolution analysis of wavelet transform and used the artificial neural network to forecast the Korean Won exchange rate. The results of Shin and Han [38] showed that the forecast accuracy of the optimized model was clearly better than the general artificial neural network.

Zhang et al. [50] extracted features out of financial time series by applying wavelet transform to the data and then constructed three different types of forecasting models by taking all the subsets as input variables for the artificial neural network, yielding results that indicated the wavelet transform-based model possessed the minimum prediction errors. Dai and Lu [13] utilized wavelet transform to break down closing prices on the Nikkei 225 into multiple sub-series and subsequently built the forecasting model by SVR (support vector regression), discovering that wavelet transform-treated stock price information can effectively enhance the forecasting capability of SVR. Zhao et al. [51] used Shanghai stock market data to compare the forecasting capabilities of the traditional ARIMA model, the ANN model, and the model that combines wavelet transform and artificial intelligence techniques. The results show that, in analyzing financial time series information, wavelet transform was able to extract more useful information. Chang and Fan [9] integrated Haar wavelet transform and Takagi–Sugeno–Kang (TSK) fuzzy rule-based systems for stock price forecasting. They applied wavelet transform to decompose the time series data into multiple sub-series, calculated the technical indices of stock price in the various sub-series, and ultimately employed the TSK fuzzy-rule-based system to predict stock price based on a set of selected technical indices. The empirical result showed that the wavelet-preprocessed TSK forecasting model outperformed the TSK forecasting model that did not apply wavelet transform.

Generally speaking, when wavelet transform is employed to construct a forecasting model, the wavelet basis functions and decomposition stages need to be determined first. The obtained wavelet sub-series are then applied to the forecasting model as input variables. However, because large amount of sub-series are generated from wavelet transform decompositions under different bases and stages, the excessiveness of input variables and the time-consumption of model construction are two problems often encountered when all sub-series are considered simultaneously. Also, forecasting models constructed by different wavelet sub-series would exhibit different forecasting capabilities and yield varying forecast results [13,17,33]. Therefore, how to identify wavelet sub-series that affect the forecast result is indeed an important task.

In this paper, we propose a new stock forecasting approach which integrates wavelet transform, multivariate adaptive regression splines (MARS) and support vector regression (SVR) approach (called Wavelet-MARS-SVR) to not only address the problem of variable selection but also improve the forecast capability. In this new approach, MARS is used to determine the importance of sub-series obtained from wavelet transform, and SVR is used to construct the stock forecasting model. We adopted MARS because it can compute the degrees of importance of variables from the numerous piecewise equations and is often used for identifying significant variables [3,27,47,52]. And SVR is chosen for the stock forecasting model because it has been widely applied to various financial time series forecasting problems and has delivered excellent results [6,11,26,31,41,42].

Our proposed approach consists of three stages. In the first stage, by utilizing wavelet transform, we decompose the predictor variable under different basis functions and decomposition stages to get the sub-series. In the second stage, we use MARS to identify the significant sub-series among all sub-series obtained from the wavelet transform. Finally, in the third stage, the identified sub-series containing the key factors that affect forecasting accuracy are applied in SVR as new input variables to build a forecasting model.

To evaluate the performance of the proposed method, two newly emerging stock market indices (SSEC & Bovespa) and two mature stock market indices (Dow Jones & Nikkei 225) are used in this study. The forecast accuracy of Wavelet-MARS-SVR is also compared with other approaches, such as Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR, and single adaptive neuro fuzzy inference system (ANFIS) models. The ARIMA and ANFIS models are selected as benchmarks in model comparison because the ARIMA model is an essential and important approach to forecast stock index [31,48] and the ANFIS model proposed by Jang [21] is a well-known and effective neuro-fuzzy system for stock price forecasting [5,8,10]. The result shows the proposed Wavelet-MARS-SVR approach can not only solve the problem of variable selection, identify the significant periods that affect the highs and lows of indices for the respective stock markets, but also have the best forecasting accuracy.

This paper contributes to the wavelet literature and stock forecasting in the following three aspects. First, basis functions and decomposition stages that were used to generate the selected sub-series can be investigated, and investigation results can be served as reference for the selection of appropriate wavelet basis functions in predicting a specific stock index. Secondly, the significance represented by the selected significant sub-series in respect to stock price data can be analyzed to subsequently figure out the data of which sessions (or points in time) among past stock market prices exerted significant impacts on the construction of the forecasting model. Thirdly, through the significant sub-series identified by MARS, the proposed Wavelet-MARS-SVR approach can construct the forecasting model more efficiently because the construction time for Wavelet-MARS-SVR is only two-thirds of Wavelet-SVR's.

The rest of this paper is organized as follows. Section II gives a brief introduction to wavelet transform, MARS and SVR. The proposed hybrid forecasting model is thoroughly described in Section III. Section IV presents the empirical results from the datasets including the SSEC, Bovespa, Dow Jones and Nikkei 225 indexes. The paper is concluded in Section V.

2. Research methodology

2.1. Wavelet transform

In this section, a brief description of wavelet transform is given. For a thorough review of wavelet transform we refer to [14,24,33]. Practical application of wavelet analysis is given in [17].

Wavelet transform is a strong mathematical tool that provides a time-frequency representation of an analyzed signal in the time domain [14,28]. It can be divided into continuous wavelet transform (CWT) and discrete wavelet transform (DWT) depending on their natures. The CWT $W(u, v)$ of signal $f(x)$ with respect to a mother wavelet $\psi(x)$ is given:

$$W(u,v) = u^{-1/2} \int_{-\infty}^{\infty} f(x)\psi\left(\frac{x-v}{u}\right)dx \qquad (1)$$

where the dilation (or scale) parameter u controls the spread of the wavelet and translation parameter v determines its central position.

Translation defines the time shift, and dilation defines the time scale. The $W(u, v)$ coefficient (called wavelet coefficient) represents how well the original signal $f(x)$ and the scaled/translated mother wavelet match. Thus, the set of all wavelet coefficients, associated to a particular signal, is the wavelet representation of the signal with respect to the mother wavelet.

Since the CWT is achieved by continuously scaling and translating the mother wavelet, substantial redundant information is generated, which is one of the main disadvantages of CWT [14,17,28]. To alleviate this redundancy problem, researchers always scaled and translated the mother wavelet based on powers of two [14,17,28]. This scheme, known as the DWT, is as accurate as the CWT and is one of the most adopted methods in literature [17,33]. For illustration, we can define the DWT as:

$$W(p,q) = 2^{-(p/2)}\sum_{t=0}^{T-1} f(t)\psi\left(\frac{t-q\cdot 2^p}{2^p}\right) \tag{2}$$

where T is the length of the signal $f(x)$. The scaling and translation parameters are functions of the integer variables p and q ($u=2^p$, $v=q\cdot 2^p$); t is the discrete time index.

In most practical applications, based on multi-resolution analysis (MRA) [28], DWT can be achieved by the straightforward filtering operation which directly implements a convolution of signal $f(t)$ and the wavelet at scale u. Thus, the wavelet plays a role of band-pass filter (the band corresponds to the scale). There is a fast algorithm for the DWT. In fast DWT, first, an original discrete signal $f(t)$ is decomposed into two components, A_1 and D_1, by convoluting the signal with a decomposition low-pass filter (D_LP) and a decomposition high-pass filter (D_HP), respectively. The decomposition low-pass and high-pass filters can be derived from mother wavelet. The A_1, named the approximation of the signal, contains the general trend (or low frequency components) of the signal $f(t)$, and the D_1, named the detail of the signal, is associated with the high frequency components of the signal $f(t)$. Then, the approximation A_1 is again decomposed into a new approximation A_2 and a detail D_2 by a larger scale and continuing to a third scale, fourth scale and so on, according to the application. (In Wavelet Transform, the scale parameter u is analogous to frequency and is a measure of the amount of detail in the signal. Therefore, a larger scale means that more of a time series is used in the calculation of the coefficients). By successive decomposition of the approximations, a multistage decomposition process can be achieved where the original signal is broken down into lower resolution components (or sub-series) in terms of the following expansion coefficients: $f(t) = A_L + \sum_{i=1}^{L} D_i$, where A_L is the approximation of the signal $f(t)$ at stage L, and D_i are the details of the signal $f(t)$ at stage $i=1, 2,..., L$. Fig. 1 shows a schematic of multi-resolution analysis of DWT decomposition.

There are many kinds of wavelets which can be used as a mother wavelet, such as Meyer wavelet, Daubechies wavelet, Morlet wavelet and so on [33,36]. These wavelets have different specificities. Daubechies wavelet is one of the most widely used wavelets and is compactly supported orthonormal wavelet and has nice performance in time series forecasting [9,24,35,50]. In this paper, the Daubechies wavelet is applied. The names of the Daubechies family wavelets are written as DB"N", where N is the order, and DB the "surname" of the wavelet.

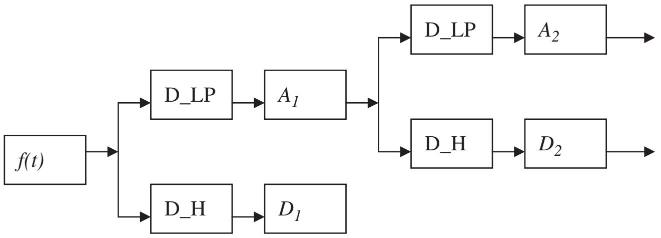


Fig. 1. Multi-resolution analysis of DWT decomposition.

In DWT, traditionally, the dyadic down-sampling process is imposed at each stage for efficiently compressing the original signal information into a compressed representation [28]. However, the major inconvenience of the DWT with down-sampling is that it is not translation-invariant [24,45]. That is, down-sampling has the undesirable effect: one cannot relate information at a given timing point at different scales in a simple manner. Moreover, while it is desirable in some applications (e.g. image compression) to remove the redundant information, in time series forecasting tasks, the redundant information can be used to improve the accuracy of the forecasting. Unser [45] proposed an overcomplete wavelet representation, namely the discrete wavelet frame transform (DWFT), to alleviate the problem caused by the dyadic down-sampling process of DWT. DWFT resembles the DWT counterpart and avoids the down-sampling operations in DWT. Thus, it guarantees aliasing degree and yields a shift invariant signal representation [24,45]. Like DWT, performing DWFT to L decomposition stages also results in a total of L details sub-series (i.e. D_i) and one approximation sub-series (i.e. A_L). Each sub-series in DWFT has the same size as the original signal.

2.2. Multivariate adaptive regression splines

Because, in our proposed Wavelet-MARS-SVR stock price forecasting model, we utilize MARS to identify significant sub-series among all sub-series obtained from the wavelet transform, a brief introduction of MARS is provided in this sub-section.

MARS is a nonlinear and non-parametric regression methodology proposed by Friedman [16]. The MARS modeling procedure is based on a divide-and-conquer strategy in which training data sets are partitioned into separate regions, each of which is assigned its own regression equation.

MARS essentially builds flexible models by fitting piecewise linear regressions; that is, the non-linearity of a model is approximated through the use of separate linear regression slopes in distinct intervals of the independent variable space. Therefore, the slope of the regression line is allowed to change from one interval to the other as the two 'knot' points are crossed. The variables to be used and the end points of the intervals for each variable are found through a fast but intensive search procedure. In addition to searching for variables one by one, MARS also searches for interactions between variables, allowing any degree of interaction to be considered as long as it can provide a better fit with the data.

The general MARS function is defined by the following equation [16,23].

$$f(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K\_m}\left[s_{k,m}\left(x(k,m) - t_{k,m}\right)\right], \tag{3}$$

where a_0 is a constant; a_m are the coefficients of the model, which are estimated to yield the best fit to the data; M is the number of basis functions; K_m is the number of splits that generate the m-th basis function; $s_{k,m}$ takes values of either 1 or -1 and indicates the right/left sense of the associated step function; $x(k,m)$ is the label of the independent variable; and $t_{k,m}$ indicates the knot locations.

The optimal MARS model is determined by a two-stage process. First, MARS initially constructs a very large number of basis functions to overfit the data, where variables are allowed to enter as continuous, categorical, or ordinal, and they can interact with one another or be restricted to entry as additive components only. In the second stage, basis functions are deleted in the order of least contributions using the generalized cross-validation (GCV) criterion [16] which is defined as

$$\mathrm{GCV}(M) = \frac{1}{N}\sum_{i=1}^{N} \frac{[y_i - f_M(x_i)]^2}{\left[1-\frac{C(M)}{N}\right]^2}, \tag{4}$$

where N denotes the number of observations; $C(M)$ is the cost-penalty measures of a model containing M basis functions; the numerator measures the lack of fit on the M basis function model $f_M(x_i)$ and the denominator denotes the penalty for model complexity $C(M)$); y_i is the target outputs. In other words, the purpose of $C(M)$ is to penalize model complexity, to avoid overfitting, and to promote model parsimony. To do so, $C(M)$ introduces a cost incurred per basis function to the model. This is similar to the adjusted R^2 in least-squares regression. It is usually defined as $C(M) = M$ in linear least-squares regression.

The importance of a variable is assessed by observing the decrease in the calculated GCV when this variable is removed from the model. This process continues until the remaining basis functions all satisfy the pre-determined requirements.

After creating a MARS model, one can estimate the relative importance of a variable based on its contribution to the fit of the model on a scale of 0–100. To do this, MARS deletes all terms containing the selected variable, refits the model and then calculates the fit's reduction, called score. Thus, the score corresponds to the ratio of the reduction in fit produced by these variables to that of the most important variable. The most important variable which has the highest score is the one that reduces the fit of the model most after being deleted, and vice versa. MARS is capable of tracking very complex data structures which are often concealed in high-dimensional data. Please refer to Friedman [16] for more details regarding the model building process.

2.3. Support vector regression

SVR, built upon statistical learning theory, is a novel neural network algorithm technique that has received increasing attention as a method for solving nonlinear regression estimation problems. SVR is derived from the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error [46].

According to Vapnik [46], the SVR model is expressed as:

$$f(x) = (\mathbf{z}\cdot\phi(x)) + b, \tag{5}$$

where $\mathbf{z}$ is a weight vector, b is bias, and $\phi(x)$ is a kernel function which is usually defined as a non-linear function to transform non-linear inputs to a linear mode in a high-dimensional feature space. Unlike the traditional regression model whose coefficients are estimated by minimizing the square loss, SVR applies so called ε-insensitivity loss function to estimate its parameters. ε-insensitivity loss function is defined as:

$$L_\varepsilon(f(x)-y) = \begin{cases} |f(x)-y|-\varepsilon \ \mathit{if} \ |f(x)-y| \geq \varepsilon \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases}, \tag{6}$$

where y is the desired(target) output; ε is defined as the region of ε-insensitivity. When the predicted value falls into the band area, the loss is zero. In contrast, if the predicted value falls outside the band area, then the loss is equal to the difference between the predicted value and the margin.

When empirical risk and structure risk are considered together, the SVR model can be constructed to minimize the following quadratic programming problem.

$$\text{Min}: \frac{1}{2}\mathbf{z}^T\mathbf{z} + C\sum_i (\xi_i + \xi_i^*)$$

$$\text{Subject to} \begin{cases} y_i - \mathbf{z}^T x_i - b \leq \varepsilon + \xi_i \\ \mathbf{z}^T x_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}, \tag{7}$$

where $i = 1,\ldots, n$ is the number of training data; $(\xi_i + \xi_i^*)$ is the empirical risk; $\frac{1}{2}\mathbf{z}^T\mathbf{z}$ is the structure risk preventing over-learning and lack of applied universality; and C is a modifying coefficient representing the trade-off between empirical risk and structure risk. With an appropriate modifying coefficient C, band area width ε, and kernel function K, the optimum value of each parameter can be solved by Lagrange. We follow Vapnik [46] and adopt the general form of the SVR-based regression function defined as

$$f(x, \mathbf{z}) = f(x, \alpha, \alpha^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x, x_i) + b, \tag{8}$$

where α_j and α_j^* are Lagrangian multipliers which satisfy the equality $\alpha_j\alpha_j^* = 0$.

Any function that meets Mercer's condition can be adopted as the kernel function. The candidates include polynomial kernel and radial basis function (RBF) kernel [8]. Among many choices, the RBF kernel is one of the most widely applied kernel function in SVR [11,41,42] and is defined as $K(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right)$, where σ denotes the width of the RBF. According to Cherkassky and Ma [11], SVR has the best performance in most forecasting programs when the value of σ is set between 0.1 and 0.5. In this paper, we use the RBF as kernel function and set σ = 0.2.

As indicated in Eq. (7), SVR model is also affected by the values of parameters C and ε. In literature, there are no general rules for the choice of C and ε. In this research, we adopt the method of grid search proposed by Lin et al. [25] in which exponentially growing sequences of C (for example, $C = 2^{-15}, 2^{-3}, 2^{-1}, \ldots, 2^{15}$) and ε are used to determine the best parameter set of C and ε which can generate the minimum forecasting mean square error.

3. Research scheme and model interpretation

3.1. Research scheme

The proposed Wavelet-MAR-SVR approach is illustrated in Fig. 2. After data preprocessing, we use wavelet transform to decompose the data into separate sub-series. Then, these sub-series were used as input variables in MARS to perform the variable selection using generalized cross-validation (GCV). Finally, the selected sub-series were integrated into the SVR approach to construct a forecasting model. The detailed illustration of each step in the research scheme is provided as follows:

Step 1 Data Preprocessing

The daily closing stock prices collected are ordered by time and the non-stationarity of the stock price data is removed by taking Log (Return) of the stock price as shown in Eq. (9):

$$\mathit{Log}(\mathit{Return})_t = \mathit{Log}\left(\frac{x_t}{x_{t-1}}\right) \tag{9}$$

where t is today's closing price and $t-1$ is yesterday's closing price.

Step 2 Wavelet transform

Wavelet transform is employed to decompose the preprocessed stock price data into wavelet sub-series. Because DB1 (a.k.a. Haar), DB2, DB3 and DB4 of the Daubechies wavelet basis functions are the ones most often applied in the previous studies [9,18,38,49,50], we have used DB1–DB4 as the wavelet basis functions for the wavelet transformation when executing

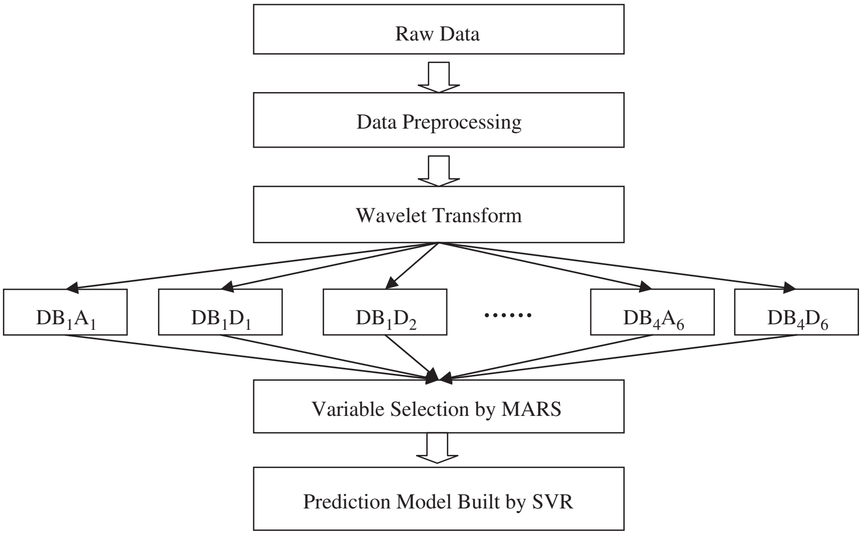


Fig. 2. The proposed Wavelet-MAR-SVR approach.

wavelet transform. Regarding the coefficients of DB1 to DB4, we simply adopted the values (shown in Table 1) proposed by Daubechies [14]. Basically, these coefficients are derived by reversing the order of the scaling function (low-pass filter) coefficients and then reversing the sign of every second one [11,29]. In addition, the maximum stage to apply the wavelet transform depends on how many data points are contained in a data set. According to the literature [43], for the decomposition of any data, it is sufficient to decompose and obtain all kinds of hidden information when the stage to apply the wavelet transform equals 6. Therefore, in this study, we will take 6 stages as the basis of decomposition. Under the settings of 6 decomposition stages and DB1–DB4 as the wavelet basis functions, we will generate 48 sub-series.

Step 3 Variable selection by MARS

To identify significant sub-series as our input variables, the MARS approach was utilized herein. In MARS, the variable selection is conducted by calculating generalized cross-validation (GVC) in Eq. (4). As shown in the equation, N calculates observation number; $\sum_{i=1}^{N}[y_i-f_M(x_i)]^2$ calculates lack-of-fit of the sum of squared residuals $f_M(x_i)$ of BF in M number which is found for data set; $\left[1-\frac{C(M)}{M}\right]^2$ is the penalty term to apply to M number of BF. The penalty term is applied for reducing the number of BFs, which tend to increase in the model and for restricting the ideal model number. Finally, the ideal MARS model is represented by an equation estimated by the lowest GCV obtained from Eq. (4). The 48 sub-series decomposed by Wavelet transformation are used as input variables in MARS. And an MARS input variable which has the level of significance greater than 5% is considered as input variable to SVR approach in this study.

Step 4 Forecasting Model Built by SVR

Finally, we build the forecasting model by regarding the significant wavelet sub-series identified by MARS as input variables in SVR. As previously mentioned, we will adopt the grid search as the parameter setting method when carrying out the construction of the forecasting model by SVR.

3.2. Model interpretation

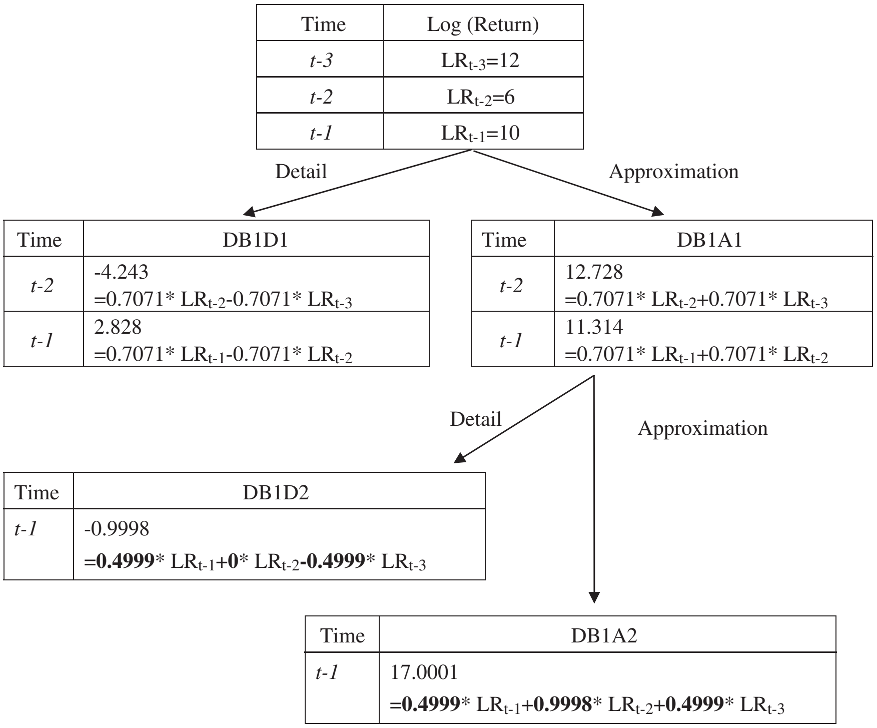
Because the analytical result of wavelet transform would vary with varying selections of basis function and varying stages of decomposition, in this study, we specifically define the basis functions employed and the stages as the following: DBiDj denotes the detailed sub-series for the j^{th} decomposition stage after having carried out the transformation using the DBi basis function; DBiAj denotes the approximated sub-series for the j^{th} decomposition stage after having carried out the transformation using the DBi basis function. Next, we use Fig.3 as an example to explain how to determine the significance of each sub-series generated under different basis function and stages.

In Fig. 3, we take 10, 6 and 12 to represent the Log(Return) value of time at $t-1$, $t-2$ and $t-3$ respectively and plan to conduct the wavelet transform at time t using the basis function DB1. Based on the high-pass and low-pass filter coefficients for the DB1 basis

Table 1
Decomposition filter coefficients of DB1 to DB4.

Basis function	Decomposition low-pass filter coefficients (approximation)							
DB1	0.7071	0.7071						
DB2	−0.1294	0.2241	0.8365	0.483				
DB3	0.0352	−0.0854	−0.135	0.4599	0.8069	0.3327		
DB4	−0.0106	0.0329	0.0308	−0.187	−0.028	0.6309	0.7148	0.2304

Basis function	Decomposition high-pass filter coefficients (detail)							
DB1	−0.7071	0.7071						
DB2	−0.4830	0.8365	−0.2241	−0.1294				
DB3	−0.3327	0.8069	−0.4599	−0.1350	0.0854	0.0352		
DB4	−0.2304	0.7148	−0.6309	−0.0280	0.1870	0.0308	−0.0329	−0.0106



Time	Log (Return)
t-3	LR_{t-3}=12
t-2	LR_{t-2}=6
t-1	LR_{t-1}=10

Time	DB1D1
t-2	-4.243 =0.7071* LR_{t-2}-0.7071* LR_{t-3}
t-1	2.828 =0.7071* LR_{t-1}-0.7071* LR_{t-2}

Time	DB1A1
t-2	12.728 =0.7071* LR_{t-2}+0.7071* LR_{t-3}
t-1	11.314 =0.7071* LR_{t-1}+0.7071* LR_{t-2}

Time	DB1D2
t-1	-0.9998 =**0.4999*** LR_{t-1}+**0*** LR_{t-2}**-0.4999*** LR_{t-3}

Time	DB1A2
t-1	17.0001 =**0.4999*** LR_{t-1}+**0.9998*** LR_{t-2}+**0.4999*** LR_{t-3}

Fig. 3. The illustration of weight computation.

function as shown in Table 1, we are able to obtain the DB1D1 sub-series, 2.828 and −4.243 respectively, and the DB1A1 sub-series, 11.314 and 12.728 respectively. In other words, DB1D1 and DB1A1, respectively, are the resulting detailed sub-series and approximated sub-series generated based on the previous two sessions' Log(Return) values. Similarly, we are able to obtain DB1D2 and DB1A2 sub-series as −0.9998 (=0.7071*[0.7071*LR_{t-1}+0.7071*LR_{t-2}]−0.7071*[0.7071*LR_{t-2}+0.7071*LR_{t-3}]) and 17.0001 (=0.7071*[0.7071*LR_{t-1}+0.7071*LR_{t-2}]+0.7071*[0.7071*LR_{t-2}+0.7071*LR_{t-3}]) respectively with the values representing the detailed sub-series and the approximated sub-series generated based on the Log(Return) values for the previous three sessions. In a wavelet transform, the approximation coefficients can be viewed as the weighted moving average, and the details coefficients can be taken as representing the degrees of data volatility, i.e. the moving difference. Therefore, if, by the MARS method, DB1A2 was selected as a significant variable, this indicates that the weighted moving average of the previous three sessions exerts an influence over the model's forecasting with the effect of the ($t-2$) session being the most significant.

Aside from the significance of DB1D1, DB1A1, DB1D2 and DB1A2, we also sort out the impact of DBiDj and DBiAj sub-series (i=1–4, j=1–6) on the forecast result. Respectively, Tables 2 and 3 list the weights of DB1A1 to DB1A4 and the weights of DB1D1 to DB1D4. The weights of DB2Aj, DB2Dj, DB3Aj, DB3Dj, DB4Aj and DB4Dj, are summarized in the Appendix A. The weights of DB2, DB3 and DB4 are respectively shown in Figs. A1, A2 and A3. From Tables 2 and 3, we discovered that, for DB1A2, DB1A3, DB1A4, DB1A5 and DB1A6 sub-series, the sessions more influential on the forecast result are ($t-2$), ($t-2$ and $t-3$), ($t-3$), ($t-3$ and $t-4$) and ($t-5$) respectively and that, for DB1D2, DB1D3, DB1D4, DB1D5 and DB1D6 sub-series, ($t-1$ and $t-3$), ($t-1$ to $t-4$), ($t-2$ and $t-4$), ($t-2$ and $t-5$) and ($t-3$ and $t-5$) are the sessions exerting greater influences on the forecast result. Moreover, on the more influential sessions for the rest of the sub-series that affect the forecast result, one can refer to the sessions that the higher line in the figure of the Appendix A corresponds to.

4. Empirical study

4.1. Datasets and performance criteria

To evaluate the performance of the proposed Wavelet-MARS-SVR forecasting model, two emerging daily stock market indexes (SSE Composite index of China (called SSEC) and Bovespa index of Brazil) and two mature daily stock market indexes (Dow Jones index of US (called DJ) and Nikkei 225 index of Japan (called N225)) are used herein. All of the data collected in this study are cash closing indexes.

Table 2
The weight table of DB1A1 to DB1A4.

Time	Stages					
	DB1A1	DB1A2	DB1A3	DB1A4	DB1A5	DB1A6
$t-7$						0.1250
$t-6$					0.1768	0.7500
$t-5$				0.2500	0.8838	1.8749
$t-4$			0.3535	1.0000	1.7677	2.4999
$t-3$		0.4999	1.0606	1.4999	1.7677	1.8749
$t-2$	0.7071	0.9998	1.0606	1.0000	0.8838	0.7500
$t-1$	0.7071	0.4999	0.3535	0.2500	0.1768	0.1250

Table 3
The weight table of DB1D1 to DB1D4.

Time	Stages					
	DB1D1	DB1D2	DB1D3	DB1D4	DB1D5	DB1D6
$t-7$						−0.1250
$t-6$					−0.1768	−0.5000
$t-5$				−0.2500	−0.5303	−0.6250
$t-4$			−0.3535	−0.5000	−0.3535	0.0000
$t-3$		−0.5000	−0.3535	0.0000	0.3535	0.6250
$t-2$	−0.7071	0.0000	0.3535	0.5000	0.5303	0.5000
$t-1$	0.7071	0.5000	0.3535	0.2500	0.1768	0.1250

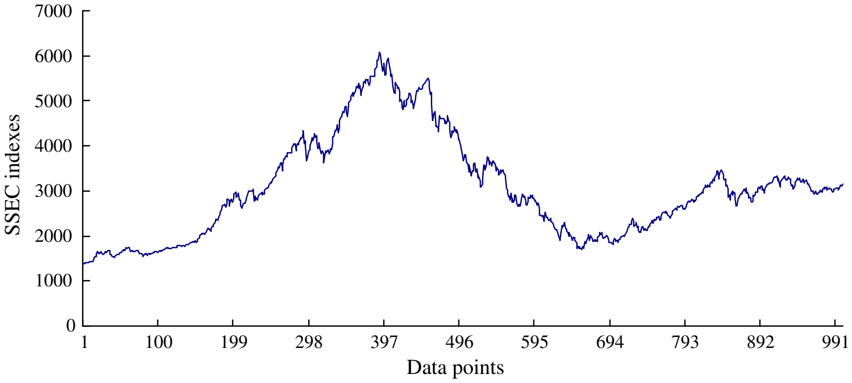


Fig. 4. The daily SSEC closing indexes from 2006/4/18 to 2010/4/1.

The time period for each closing index is summarized and shown in Figs. 4–7. There are a total of 1000 data points for each dataset. The first 800 data points (80% of the total sample points) are used as the training sample while the remaining 200 data points (20% of the total sample points) are used as the testing sample.

The forecasting results of the proposed model are compared to integrated wavelet and SVR model without using MARS (called Wavelet-SVR model), integrated wavelet and MARS model without using SVR (called Wavelet-MARS model), single ARIMA, single SVR and single ANFIS models. The forecasting performance is evaluated using the following performance measures: the root mean square error (RMSE), mean absolute difference (MAD), mean absolute percentage error (MAPE), and root mean square percentage error (RMSPE). The definitions of these criteria were summarized in Table 4. RMSE, MAD, MAPE and RMSPE are measures of the deviation between actual and predicted values. The smaller the deviation, the better the accuracy.

4.2. SSEC index

For Wavelet-SVR model, first, the preprocessed data was passed to the wavelet transform model for decomposition. Since the study adopts four basis functions, DB1–DB4, and sets the number of decomposition stages to 6, 48 sub-series are generated following the decomposition process. All these 48 sub-series are directly used as input variables in Wavelet-SVR model. The grid search method is used for searching the best parameter set for Wavelet-SVR model. The testing results of Wavelet-SVR model with combinations of different parameter sets are summarized in Table 5. Table 5 shows that the parameter set ($C=2^{-15}$, $\varepsilon=2^{-5}$) gives the best forecasting result (minimum testing MSE) and is the best parameter set for Wavelet-SVR model in forecasting SSEC index.

For Wavelet-MARS model, the 48 sub-series decomposed by wavelet transform are used as input variables. Table 6 summarizes the obtained significant variables and their relative importance. Among the 48 sub-series, 7 significant sub-series are selected by MARS, being DB4A1, DB4A4, DB2A4, DB4D1, DB1D3, DB3A1 and DB2A5, respectively. We find that DB4A1 (approximate function of closing prices for the previous 8 sessions) is the most important percentage in the table, followed by DB4A4 (approximate function for the previous 30 sessions). This finding suggests that, as it is in a newly emerging market, the stock prices of the SSEC Index are still quite volatile, and therefore, moving averages over longer periods of time is necessary for making forecasting of trends in stock prices.

For the proposed Wavelet-MARS-SVR model, seven important sub-series (DB4A1, DB4A4, DB2A4, DB4D1, DB1D3, DB3A1 and DB2A5) identified by MARS are used as input variables in SVR approach. The testing results of Wavelet-MARS-SVR model with combinations of different parameter sets are summarized in Table 7. It can be observed from Table 7 that the parameter set ($C=2^{-13}$, $\varepsilon=2^{-5}$) gives the best forecasting result and hence is the best parameter setup for the proposed Wavelet-MARS-SVR model.

For developing single SVR model, the closing indices of the previous 1 day ($t-1$), 2 days ($t-2$) and 3 days ($t-3$) are directly used as input variables. Table 8 summarizes the model selection results of the single SVR model. As shown in the table, the parameter set ($C=2^{-11}$, $\varepsilon=2^{-7}$) is the best parameter setup for the single SVR model.

The ANFIS toolbox of MATLAB software (MATLAB software, R2012 version, The MathWorks Inc.) is used to analyze the data with the single ANFIS model in this study. The input variables of the single

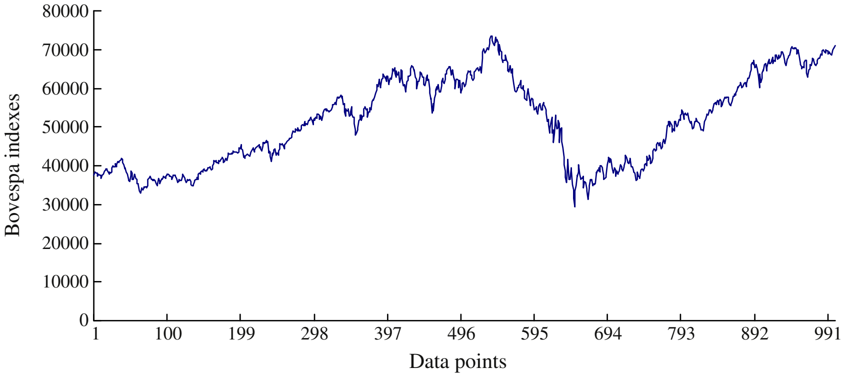


Fig. 5. The daily Bovespa closing indexes from 2006/3/14 to 2010/4/1.

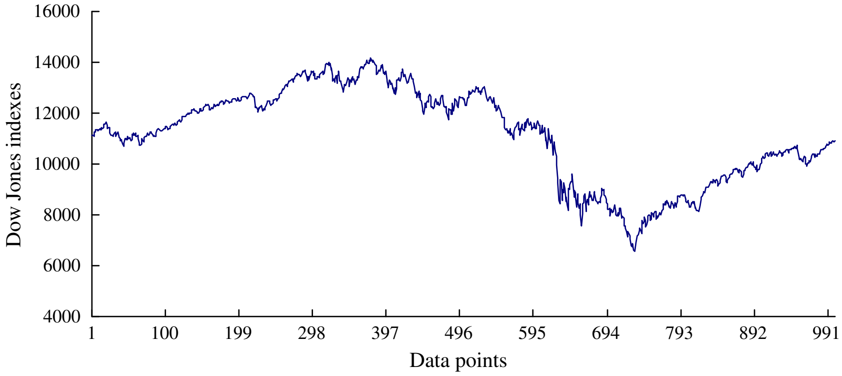


Fig. 6. The daily DJ closing indexes from 2006/4/12 to 2010/4/1.

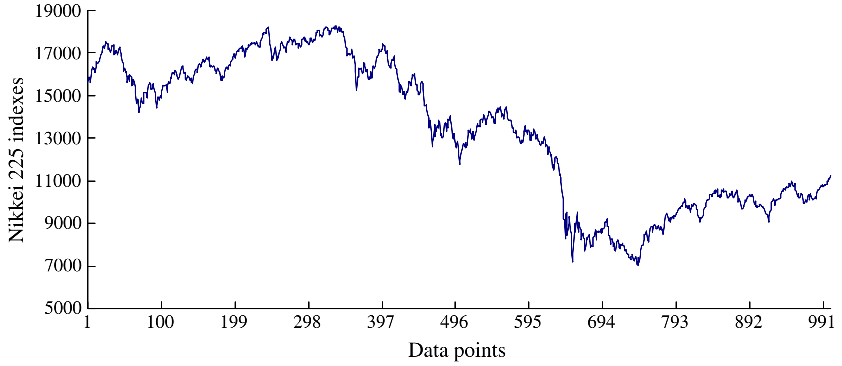


Fig. 7. The daily N225 closing indexes from 2006/3/3 to 2010/4/1.

ANFIS model are the same as the inputs of the single SVR model. The default settings of the ANFIS toolbox are used for forecasting SSEC index.

For single ARIMA model, the data not decomposed by wavelet transform are directly used as input variables. The SPSS statistical package (SPSS software, PC version 12; SPSS Inc.) is used to build ARIMA model in this study. The estimated results are summarized in Table 9.

The SSEC index forecasting results using Wavelet-MARS-SVR, Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR and single ANFIS models are computed and listed in Table 10. Table 10 depicts that RMSE, MAD, MAPE and RMSPE of the proposed Wavelet-MARS-SVR model are 52.92204, 39.01709, 1.255% and 1.7116%, respectively. It can be observed that these values are smaller than those of the five comparison models. It indicates that there is a smaller deviation between the actual and predicted values when the proposed model is applied. Thus, the proposed Wavelet-MARS-SVR model provides a better forecasting result than Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR and single ANFIS models in terms of forecasting error for SSEC index.

4.3. Bovespa, Dow Jones, and Nikkei 225 Stock Indexes

The stock index forecasting for DJ, N225 and Bovespa is conducted using similar modeling process illustrated in Section 4.2. Respectively, Tables 11–13 show the results of MARS-based variable selection for Bovespa, Dow Jones and Nikkei 225.

Table 4
Performance measures and their definitions.

Metrics	Calculation*
RMSE	$\text{RMSE} = \sqrt{\dfrac{\sum_{i=1}^{n}(T_i - P_i)^2}{n}}$
MAD	$\text{MAD} = \dfrac{\sum_{i=1}^{n}\lvert T_i - P_i\rvert}{n}$
MAPE	$\text{MAPE} = \dfrac{\sum_{i=1}^{n}\left\lvert\dfrac{T_i - P_i}{T_i}\right\rvert}{N}$
RMSPE	$\text{RMSPE} = \sqrt{\dfrac{\sum_{i=1}^{n}\left(\dfrac{T_i - P_i}{T_i}\right)^2}{n}}$

* Note that T and P represent the actual and predicted value, respectively, n is total number of data points.

Table 5
The model selection results of Wavelet-SVR model-SSEC.

C	ε	Training MSE	Testing MSE
$\mathbf{2^{-15}}$	2^{-9}	0.0003046	0.0005085
	2^{-7}	0.0003029	0.0005064
	$\mathbf{2^{-5}}$	**0.0003010**	**0.0005052**
	2^{-3}	0.0003105	0.0005164
	2^{-1}	0.0003175	0.0005245
2^{-13}	2^{-9}	0.0003046	0.0005081
	2^{-7}	0.0003029	0.0005064
	2^{-5}	0.0003039	0.0005055
	2^{-3}	0.0003105	0.0005163
	2^{-1}	0.0003175	0.0005245
2^{-11}	2^{-9}	0.0003046	0.0005084
	2^{-7}	0.0003029	0.0005063
	2^{-5}	0.0003010	0.0005052
	2^{-3}	0.0003105	0.0005167
	2^{-1}	0.0003175	0.0005246

Table 6
Results of MARS-based variable selection – SSEC.

Variables	Connotation of variable	Std. dev	GCV value	Importance (%)
DB4A1	Weighted average of closing prices of previous 8 sessions	0.000485	0.003	100.00
DB4A4	Weighted average of closing prices of previous 30 sessions	0.000483	0.003	87.97
DB2A4	Weighted average of closing prices of previous 7 sessions	0.000483	0.003	80.59
DB4D1	Weighted deviation of closing prices of previous 8 sessions	0.000487	0.003	72.06
DB1D3	Weighted deviation of closing prices of previous 5 sessions	0.000484	0.003	67.36
DB3A1	Weighted average of closing prices of previous 6 sessions	0.000479	0.002	45.90
DB2A5	Weighted deviation of closing prices of previous 26 sessions	0.000480	0.002	29.46

Table 7
The model selection results of the proposed Wavelet-MARS-SVR model-SSEC.

C	ε	Training MSE	Testing MSE
$\mathbf{2^{-13}}$	2^{-9}	0.0003047	0.0005061
	2^{-7}	0.0003029	0.0005047
	$\mathbf{2^{-5}}$	**0.0003005**	**0.0005037**
	2^{-3}	0.0003090	0.0005347
	2^{-1}	0.0003246	0.0005222
2^{-11}	2^{-9}	0.0003047	0.0005061
	2^{-7}	0.0003029	0.0005047
	2^{-5}	0.0003011	0.0005038
	2^{-3}	0.0003090	0.0005147
	2^{-1}	0.0003246	0.0005222
2^{-9}	2^{-9}	0.0003047	0.0005061
	2^{-7}	0.0003029	0.0005047
	2^{-5}	0.0003011	0.0005039
	2^{-3}	0.0003090	0.0005147
	2^{-1}	0.0003246	0.0005322

From Table 11, it can be seen that six MARS-selected significant sub-series (DB1A6, DB2D2, DB1D1, DB4D2, DB4A4, DB1D4) have been identified for the Brazil. Among them, DB1A6 (weighted average function of closing prices of previous 6 sessions) exhibits the highest degree of importance, followed by DB2D2 (deviation function of closing functions of previous 9 sessions). Moreover, since the majority of significant sub-series shown in Table 11 are deviation functions, we can infer that, as a result of the Brazil market's higher degree of volatility, stock price deviations would need to be considered in order to make effective predictions on trends in stock prices.

Table 8
The model selection results of the single SVR model-SSEC.

C	ε	Training MSE	Testing MSE
2^{-13}	2^{-9}	0.0003691	0.0005780
	2^{-7}	0.0003706	0.0005768
	2^{-5}	0.0003788	0.0005811
	2^{-3}	0.0003602	0.0005302
	2^{-1}	0.0003944	0.0005923
$\mathbf{2^{-11}}$	2^{-9}	0.0003519	0.0005448
	$\mathbf{2^{-7}}$	**0.0003203**	**0.0005189**
	2^{-5}	0.0003993	0.0005708
	2^{-3}	0.0003305	0.0005328
	2^{-1}	0.0004103	0.0005479
2^{-9}	2^{-9}	0.0003427	0.0005201
	2^{-7}	0.0003618	0.0005230
	2^{-5}	0.0004460	0.0005802
	2^{-3}	0.0004780	0.0005998
	2^{-1}	0.0005248	0.0006115

Table 9
Single ARIMA model-SSEC.

Parameter	Estimate	Standard error	*T*-value	Approx Pr>| t |	Lag
MU	−5.50E−06	4.96E−06	−1.11	0.2678	0
MA1,1	1.01136	0.03539	28.58	<.0001	1
MA1,2	−0.0063	0.05035	−0.13	0.9005	2
MA1,3	−0.0668	0.05029	−1.33	0.1845	3
MA1,4	−0.05112	0.05038	−1.01	0.3105	4
MA1,5	0.11286	0.03538	3.19	0.0015	5
Model parameter					
Constant estimate	−0.0000055				
Variance estimate	0.000502				
Std error estimate	0.02241				
AIC	−3796.13				
SBC	−3768.03				
Number of residuals	799				

Table 10
Summary of forecast results by six models on SSEC closing prices.

Models	RMSE	MAD	MAPE	RMSPE
Wavelet -SVR	52.93402	39.05813	1.268%	1.7218%
Wavelet -MARS	54.98807	40.37382	1.310%	1.7965%
Wavelet-MARS-SVR	**52.92204**	**39.01709**	**1.255%**	**1.7116%**
Single ARIMA	54.90912	40.17590	1.331%	1.7357%
Single SVR	53.95342	39.87542	1.291%	1.7294%
Single ANFIS	54.21242	39.91241	1.293%	1.7307%

Table 11
Results of MARS-based variable selection – BVSP.

Variables	Connotation of variable	Std. dev	GCV value	Importance (%)
DB1A6	Weight average function of closing prices of previous 6 sessions	0.005	0.000527	100.00
DB2D2	Deviation function of closing prices of previous 9 sessions	0.005	0.000524	87.97
DB1D1	Deviation function of closing prices of previous 2 sessions	0.005	0.000524	80.59
DB4D2	Deviation function of closing prices of previous 15 sessions	0.004	0.000516	72.06
DB4A4	Weighted average function of closing prices of previous 29 sessions	0.003	0.000514	67.36
DB1D4	Deviation function of closing prices of previous 5 sessions	0.003	0.000512	45.90

Table 12
Results of MARS-based variable selection – Dow Jones.

Variables	Connotation of variable	Std. dev	GCV value	Importance (%)
DB3D5	Weighted average function of closing prices of previous 26 sessions	0.004	0.000256	100.00
DB1A1	Weighted average function of closing prices of previous 2 sessions	0.003	0.000255	97.92
DB2A2	Weighted average function of closing prices of previous 7 sessions	0.003	0.000254	89.60
DB3D4	Deviation function of closing prices of previous 21 sessions	0.003	0.000253	86.26
DB2D3	Deviation function of closing prices of previous 13 sessions	0.003	0.000252	76.77
DB4D1	Deviation function of closing prices of previous 8 sessions	0.002	0.000250	60.71
DB2A4	Weighted average function of closing prices of previous 13 sessions	0.002	0.000249	52.10
DB1A5	Deviation function of closing prices of previous 6 sessions	0.002	0.000249	48.57

Table 13
Results of MARS-based variable selection – Nikkei 225.

Variables	Connotation of variable	Std. dev	GCV value	Importance (%)
DB2A1	Weighted average function of closing prices of previous 4 sessions	0.007	0.000364	100.00
DB4D1	Deviation function of closing prices of previous 8 sessions	0.005	0.000351	77.33
DB3A2	Weighted average function of closing prices of previous 11 sessions	0.004	0.000345	63.15
DB4A1	Weighted average function of closing prices of previous 8 sessions	0.004	0.000344	61.56
DB2D4	Deviation function of closing prices of previous 17 sessions	0.004	0.000341	52.84
DB3D5	Deviation function of closing prices of previous 26 sessions	0.003	0.000338	43.16
DB3D6	Deviation function of closing prices of previous 31 sessions	0.002	0.000336	35.15
DB1A4	Weighted average function of closing prices of previous 5 sessions	0.005	0.000333	17.30

Table 14
Summary of forecast results by six models on BVSP, DJ and N225 closing prices.

Stock indexes	Models	RMSE	MAD	MAPE	RMSPE
BVSP	Wavelet-SVR	878.7561	642.7191	1.044%	1.430%
	Wavelet-MARS	881.0468	651.4920	1.055%	1.437%
	Wavelet-MARS-SVR	**876.9092**	**641.1467**	**1.043%**	**1.428%**
	Single ARIMA	883.9282	652.8783	1.072%	1.439%
	Single SVR	880.0151	649.6584	1.053%	1.436%
	Single ANFIS	882.0765	651.6157	1.057%	1.437%
DJ	Wavelet-SVR	32.2070	24.6752	0.742%	0.983%
	Wavelet-MARS	33.0173	25.4042	0.762%	1.010%
	Wavelet-MARS-SVR	**32.2058**	**24.6743**	**0.741%**	**0.983%**
	Single ARIMA	33.6321	25.3284	0.753%	1.053%
	Single SVR	32.9112	24.9462	0.748%	0.997%
	Single ANFIS	33.2513	25.2615	0.751%	1.021%
N225	Wavelet-SVR	133.2766	104.8005	1.014%	1.321%
	Wavelet-MARS	138.8247	110.5228	1.093%	1.381%
	Wavelet-MARS-SVR	**132.3721**	**103.4128**	**1.012%**	**1.303%**
	Single ARIMA	140.3164	111.2556	1.130%	1.380%
	Single SVR	137.6541	109.3261	1.081%	1.368%
	Single ANFIS	135.5629	107.6568	1.066%	1.352%

With respect to the Dow Jones data, the MARS-selected significant sub-series are summarized in Table 12. As shown in the table, there are eight significant sub-series (DB3D5, DB1A1, DB2A2, DB3D4, DB2D3, DB4D1, DB2A4 and DB1A5) that have been selected. DB3D5 (weighted average function of closing prices of previous 26 sessions) is the most important one, followed by DB1A1 (weighted average function of closing prices of previous 2 sessions). Furthermore, most of these significant sub-series are weighted average functions, indicating the smaller volatility of the U.S. market and the higher reference value of its weighted average information.

Table 13 summarizes the significant sub-series selected by MARS for the Nikkei 225 stock index. As shown in the table, these eight selected, significant sub-series are DB2A1, DB4D1, DB3A2, DB4A1, DB2D4, DB3D5, DB3D6 and DB1A4. Among them, the most important sub-series is DB2A1 (weighted average function of closing prices of previous 4 sessions), and the next most important one is DB4D1 (deviation function of closing prices of previous 8 sessions). Since most sub-series in Table 13 exhibiting higher significance are weighted average functions, we therefore infer and predict that the Japanese stock market is a more stable market and that, for forecasting stock prices, weighted averages can provide more information.

Table 14 summarizes Dow Jones, Nikkei 225, and Bovespa stock index forecasting results using Wavelet-MARS-SVR, Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR and single ANFIS models, respectively. It can be observed that the proposed Wavelet-MARS-SVR model has the smallest RMSE, MAD, MAPE and RMSPE in comparison with the five competing models in every stock market index. (Wavelet-MARS-SVR model only slightly better than Wavelet-SVR model in forecasting error, but the model construction time for Wavelet-MARS-SVR is 2/3 shorter than Wavelet-SVR's.) Thus, the proposed Wavelet-MARS-SVR can produce lower forecasting errors and outperforms the five competing models in forecasting SSEC, Bovespa, DJ and N225 stock indexes.

4.4. Robustness evaluation

To evaluate the robustness of the proposed method, the performance of the Wavelet-MARS-SVR, Wavelet-SVR, Wavelet-MARS,

Table 15
Robustness evaluation.

Relative ratio	Models	China (SSEC) Testing MAPE	Brazil (BVSP) Testing MAPE	USA (DJ) Testing MAPE	Japan (N225) Testing MAPE
60%	Wavelet-SVR	1.561%	1.892%	1.493%	1.673%
	Wavelet -MARS	1.701%	1.951%	2.021%	2.030%
	Wavelet -MARS-SVR	**1.521%**	**1.800%**	**1.421%**	**1.611%**
	Single ARIMA	1.712%	2.312%	1.950%	1.801%
	Single SVR	1.673%	1.940%	1.711%	1.778%
	Single ANFIS	1.698%	1.951%	1.721%	1.769%
70%	Wavelet-SVR	1.321%	1.279%	1.051%	1.242%
	Wavelet -MARS	1.412%	1.343%	1.182%	1.291%
	Wavelet -MARS-SVR	**1.302%**	**1.235%**	**1.019%**	**1.203%**
	Single ARIMA	1.510%	1.752%	1.184%	1.261%
	Single SVR	1.351%	1.302%	1.094%	1.257%
	Single ANFIS	1.355%	1.331%	1.096%	1.255%
80%	Wavelet-SVR	1.268%	1.044%	0.742%	1.014%
	Wavelet -MARS	1.310%	1.055%	0.762%	1.093%
	Wavelet -MARS-SVR	**1.255%**	**1.043%**	**0.741%**	**1.012%**
	Single ARIMA	1.331%	1.072%	0.753%	1.130%
	Single SVR	1.291%	1.053%	0.748%	1.081%
	Single ANFIS	1.293%	1.057%	0.751%	1.066%
90%	Wavelet-SVR	1.021%	0.918%	0.606%	0.940%
	Wavelet -MARS	1.032%	0.948%	0.632%	1.232%
	Wavelet -MARS-SVR	**1.002%**	**0.891%**	**0.581%**	**0.902%**
	Single ARIMA	1.032%	0.962%	0.623%	1.122%
	Single SVR	1.029%	0.945%	0.615%	0.985%
	Single ANFIS	1.029%	0.946%	0.619%	0.977%

single ARIMA, single SVR and single ANFIS models was tested using different ratios of training and testing sample sizes. The testing experiment is based on the relative ratio of the size of the training dataset size to complete dataset size. In this section, four relative ratios, 60%, 70%, 80%, and 90% are considered. The forecasting results for the four indexes by the six methods are summarized in Table 15 in terms of RMSE, MAD, MAPE and RMSPE. In Table 15, it can be observed that the proposed Wavelet-MARS-SVR method outperforms the other benchmarking tools under all four different ratios in terms of the four different performance measures. It therefore indicates that Wavelet-MARS-SVR approach indeed provides better forecast accuracy than the other five approaches.

4.5. Significance test

In order to test whether the proposed Wavelet-MAR-SVR model is superior to Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR and single ANFIS models in financial stock index forecasting, the Wilcoxon signed-rank test is applied. The Wilcoxon signed-rank test is a distribution-free, non-parametric technique which determines whether two models are different by comparing the signs and ranks of prediction values. The Wilcoxon signed-rank test is one of the most popular tests in evaluating the predictive capabilities of two different models [15,34,40]. For the details of the Wilcoxon signed-rank test, please refer to Diebold and Mariano [15] and Pollock et al. [34].

We employ the test to evaluate the predictive performance of the proposed method and the five competing models under different ratios of the size of the training data set to the complete data set. Tables 16 and 17 present the Z statistic values of the two-tailed Wilcoxon signed-rank test for RMSE values between the proposed Wavelet-MARS-SVR model and other five competing models in four stock markets. It can be observed from Tables 16 and 17, under different ratios, that the RMSE values of the proposed Wavelet-MARS-SVR model are significantly different from Wavelet-MARS, single ARIMA, single SVR and single ANFIS models. We can therefore conclude that the proposed Wavelet-MARS-SVR model is significantly better than Wavelet-MARS, single ARIMA, single SVR and single ANFIS models in financial stock index forecasting.

Even though the RMSE values of the proposed model has no significant difference from Wavelet-SVR, the proposed Wavelet-MARS-SVR model uses less variables to achieve the same forecast accuracy of Wavelet-SVR, which implies that Wavelet-MARS-SVR model is more efficient in model construction, and its explanation to the predictor variables provides a solid foundation in efficient decision-making.

4.6. Investigation of variables in the various markets

The historical sessions are screened from data of different markets. To further illustrate how the MARS selected significant sub-series can be used to explain the influence of different sessions on each stock indexes, we sort out the resulting MARS-selected significant sub-series generated from four different training and testing data ratios and summarize the results in Table 18. From the table, we find that the DB4D1 sub-series (weighted deviation of previous 8 sessions) appeared more than twice in SSEC, DJ and N225 stock indices and that DB4D2 (weighted deviation of previous 15 sessions) showed up three times in the Bovespa stock index (Figs. A1 and A2). Such results indicate that the degree of volatility in the previous eight sessions has a significant influence over the forecasting of the stock market. In addition, based on the weight of coefficient of DB4D1 as shown in Fig. A3, we are also able to infer further that, for SSEC, DJ and N225 stock indices, the closing price of the $(t-7)$ session has a greater significance on the index forecast than all the ones in the other eight sessions, whereas for the Bovespa stock index, the impact of the closing prices of previous 15 sessions on the index forecast is more significant with the importance of information

Table 16
Wilcoxon signed-rank test between Wavelet-MAR-SVR model, Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR and single ANFIS models by different relative ratios -SSEC and BVSP.

Models	Relative ratio	Wavelet-SVR		Wavelet-MARS		Single ARIMA		Single SVR		Single ANFIS	
		SSEC	BVSP	SSEC	BVSP	SSEC	BVSP	SSEC	BVSP	SSEC	BVSP
Wavelet-MAR-SVR	60%	−1.091 (.275)	−1.33 (.183)	−2.936 **(.003)	−1.644 *(.096)	−4.061 **(.000)	−7.06 **(.000)	−1.936 **(.041)	−1.571 *(.099)	−2.612 **(.010)	−1.642 *(.096)
	70%	−0.162 (.871)	-.817 (.414)	−2.408 **(.016)	−1.736 *(.083)	−4.849 **(.000)	−5.659 **(.000)	−1.708 *(.076)	−1.412 *(.097)	−1.751 *(.073)	−1.993 **(.039)
	80%	−0.89 (.374)	-.558 (.577)	−1.223 **(.021)	−1.752 *(.080)	−1.206 **(.022)	−1.956 **(.000)	−1.011 *(.082)	−1.731 *(.084)	−1.063 *(.068)	−1.912 **(.007)
	90%	−1.375 (.169)	-.602 (.547)	−7.028 **(.000)	−1.485 *(.093)	−4.70 **(.000)	−8.661 **(.050)	−4.912 **(.000)	−1.406 *(.096)	−4.915 **(.000)	−1.405 *(.096)

Note: The numbers in parentheses are the corresponding p-value; *: $p<0.1$; **: $p<0.05$.

Table 17
Wilcoxon signed-rank test between the six models by different relative ratios-DJ and N225.

Models	Relative ratio	Wavelet-SVR		Wavelet-MARS		Single ARIMA		Single SVR		Single ANFIS	
		DJ	N225	DJ	N225	DJ	N225	DJ	N225	DJ	N225
Wavelet-MAR-SVR	60%	-.960 (.337)	-.407 (.684)	−4.378 **(.000)	−4.153 **(.000)	−3.899 **(.000)	−1.836 **(.000)	−1.861 **(.000)	−1.719 **(.000)	−1.903 **(.000)	−1.516 **(.002)
	70%	-.876 (.381)	-.583 (.560)	−3.832 **(.000)	−3.870 **(.000)	−4.561 **(.000)	−5.572 **(.000)	−2.091 **(.000)	−4.097 **(.000)	−2.103 **(.000)	−3.921 **(.000)
	80%	-.350 (.726)	-.985 (.325)	−1.807 *(.071)	−2.578 **(.010)	−2.263 **(.000)	−2.365 **(.018)	−1.587 *(.088)	−2.162 **(.031)	−1.867 *(.057)	−1.932 *(.054)
	90%	−1.489 (.137)	−1.76 *(.086)	−3.452 **(.001)	−3.37 **(.001)	−1.987 **(.047)	−2.094 **(.036)	−1.637 *(.077)	−1.968 *(.057)	−1.712 *(.062)	−1.912 *(.061)

Note: The numbers in parentheses are the corresponding *p*-value; *: $p<0.1$; **: $p<0.05$.

Table 18
Summary of significant variables selected by the proposed method under different training ratios for the various markets.

SSEC		BVSP		DJ		N225	
Sub-series	Freq.	Sub-series	Freq.	Sub-series	Freq.	Sub-series	Freq.
DB2A4	4	DB1A6	3	DB2D3	3	DB4D1	3
DB4A4	4	DB1D1	3	DB3D4	3	DB1A4	2
DB4D1	4	DB4D2	3	DB1A5	2	DB2A1	2
DB4A1	3	DB1D4	1	DB2A2	2	DB3A2	2
DB1D3	2	DB2D1	1	DB3A6	2	DB3D5	1
DB2A5	2	DB2D3	1	DB3D5	2	DB1A1	1
DB3A1	2	DB3D1	1	DB4D1	2	DB1A2	1
DB1D5	1	DB3D3	1	DB1A1	1	DB1A4	1
DB2A1	1	DB3D4	1	DB1A5	1	DB1A5	1
DB2A3	1	DB4A2	1	DB1A6	1	DB1D2	1
DB3A2	1	DB4A4	1	DB1D3	1	DB1D3	1
DB3A5	1			DB2A3	1	DB1D6	1
				DB2A4	1	DB2D3	1
				DB2A6	1	DB2D4	1
				DB2D5	1	DB2D5	1
				DB3D2	1	DB3A5	1
				DB4A3	1	DB3D2	1
				DB4A4	1	DB4A1	1
				DB4D3	1	DB4D2	1
				DB4D5	1	DB4D4	1
						DB3D6	1
						DB3D6	1

revealed, with $(t-7)$ and $(t-9)$ sessions being higher than the other 13 sessions. Furthermore, of the selected significant sub-series, the DB4 basis function is the most frequent, which also verifies indirectly the conclusion made in the literature [1,30].

5. Concluding remarks

This paper proposed a three-stage forecasting model by integrating wavelet transform, MARS and SVR for financial time series. The proposed Wavelet-MARS-SVR method first uses Wavelet transform to decompose the financial time series data. Then, the decomposed sub-series are used as input variables in MARS for variable selection. Finally, the identified sub-series containing the key factors that affect forecasting accuracy are applied in SVR as the new input variables to build up a forecasting model.

Four datasets including SSEC index of China, Bovespa index of Brazil, Dow Jones index of US and Nikkei 225 index of Japan are used to evaluate the proposed method. Moreover, this study compares the proposed method with Wavelet-SVR, Wavelet-MARS, single ARIMA, single SVR and single ANFIS models using forecasting error as a criterion. The empirical results show that the proposed model can produce lower forecasting error and outperform other five competing models. Moreover, the model construction time for Wavelet-MARS-SVR is 2/3 shorter than Wavelet-SVR's. According to the results, it can be concluded that the proposed method can effectively select the important wavelet sub-series and improve the forecasting performance of SVR. Moreover, through the proposed approach, the data of which sessions among past stock market prices exerted significant impacts on the construction of the forecasting model can be successfully identified. Future research can aim at combining other forecasting tools, like neural networks and grey system theory, in evaluating the ability of the proposed forecasting scheme.

Acknowledgements

This work is partially supported by the Chien Hsin University of Science and Technology (formerly Ching Yun University), Grant No. 99-IE-002-CM. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

Appendix A

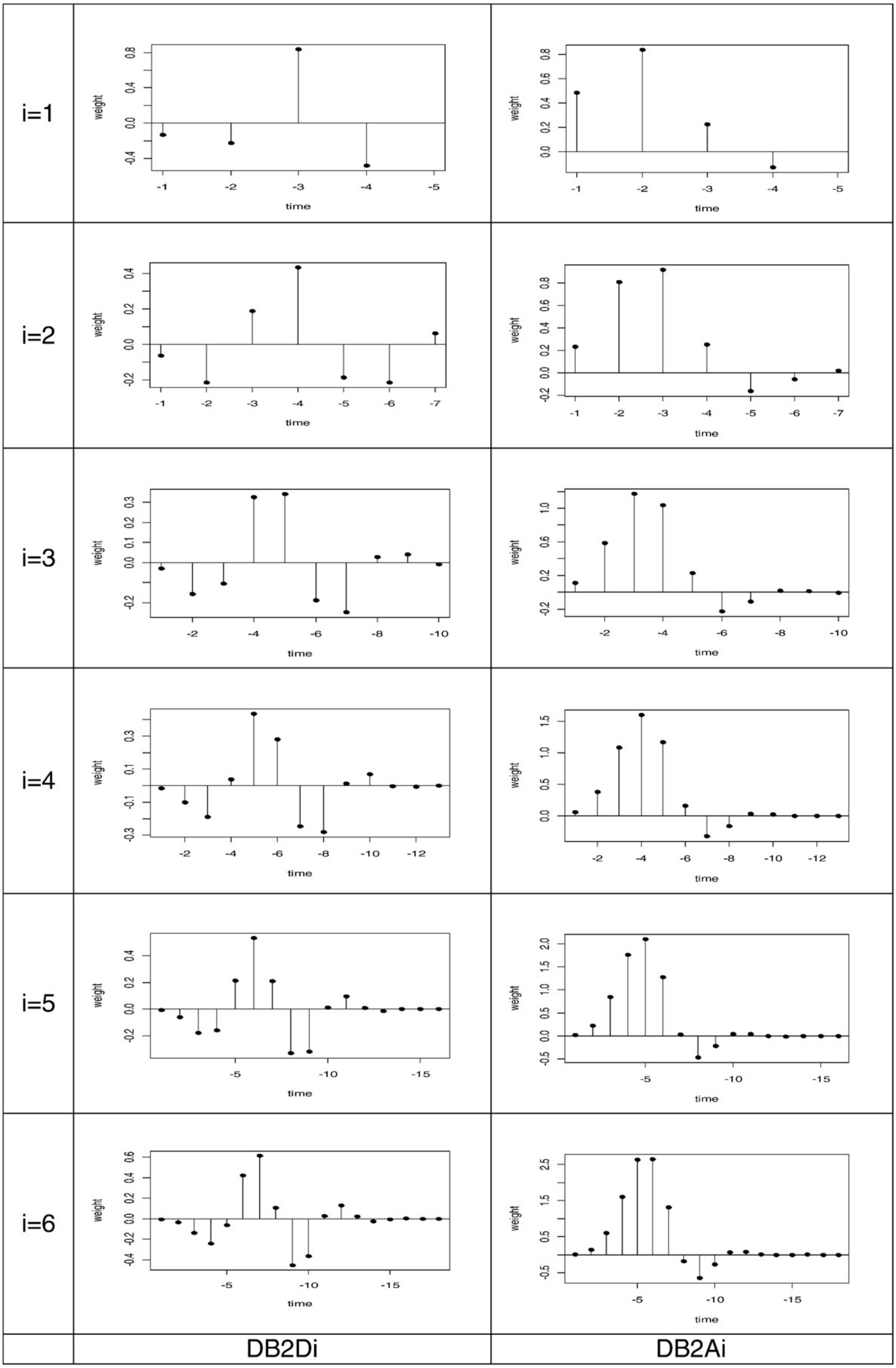


Fig. A1. The weights of each period of DB2.

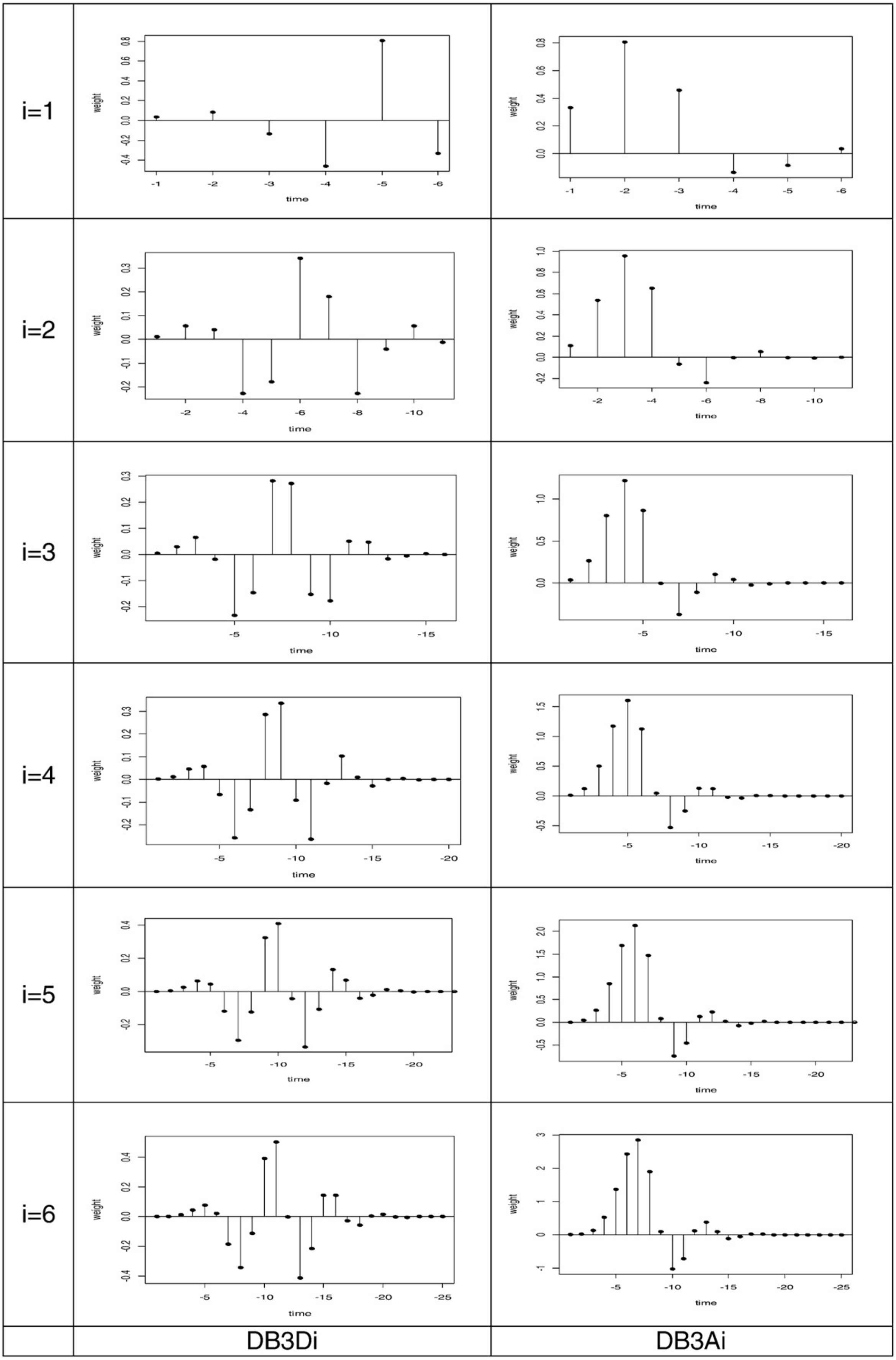


Fig. A2. The weights of each period of DB3.

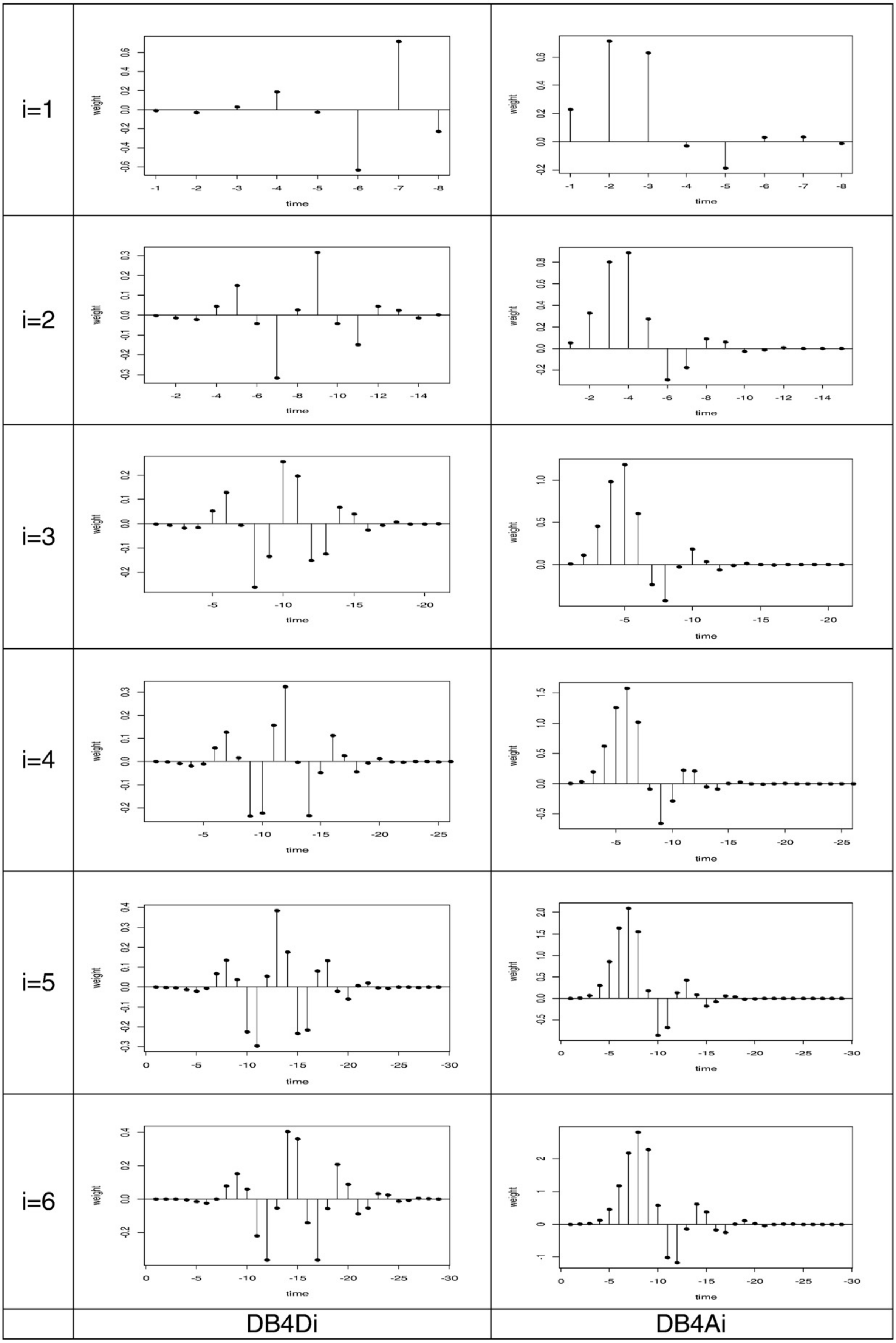


Fig. A3. The weights of each period of DB4.

References

[1] D.P. Ahalpara, A. Verma, J.C. Parikh, P.K. Panigrahi, Characterizing and modelling cyclic behaviour in non-stationary time series through multi-resolution analysis, Pramana 71 (3) (2008) 459–485.

[2] V. Alarcon-Aquino, J.A. Barria, Multiresolution FIR neural-network-based learning algorithm applied to network traffic prediction, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 36 (2) (2006) 80–92.

[3] A. Andalib, F. Atry, Multi-step ahead forecasts for electricity prices using NARX: a new approach, a critical analysis of one-step ahead forecasts, Energy Conversion and Management 50 (3) (2009) 739–747.

[4] G.S. Atsalakis, K.P. Valavanis, Surveying stock market forecasting techniques – Part II: Soft computing methods, Expert Systems with Applications 36 (3) (2009) 5932–5941.

[5] G.S. Atsalakis, K.P. Valavanis, Forecasting stock market short-term trends using a neuro-fuzzy based methodology, Expert Systems with Applications 36 (7) (2009) 10696–10707.

[6] A. Bahrammirzaee, A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems, Neural Computing & Applications 19 (8) (2010) 1165–1195.

[7] V. Bjorn, Multiresolution methods for financial time series prediction, in: Proceedings of the 1995 IEEE/IAFE on Computational Intelligence for Financial Engineering, New York, 1995, p. 97.

[8] M.A. Boyacioglu, D. Avci, An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock market return: the case of the Istanbul stock exchange, Expert Systems with Applications 37 (12) (2010) 7908–7912.

[9] P.C. Chang, C.Y. Fan, A hybrid system integrating a wavelet and TSK fuzzy rules for stock price forecasting, IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 38 (6) (2008) 802–815.

[10] J.R. Chang, L.Y. Wei, C.H. Cheng, A hybrid ANFIS model based on AR and volatility for TAIEX forecasting, Applied Soft Computing Journal 11 (1) (2011) 1388–1395.

[11] V. Cherkassky, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Networks 17 (2004) 113–126.

[12] T.M. Choi, Y. Yu, K.F. Au, A hybrid SARIMA wavelet transform method for sales forecasting, Decision Support Systems 51 (1) (2011) 130–140.

[13] W. Dai, C.J. Lu, Financial time series forecasting using a compound model based on wavelet frame and support vector regression, in: Proceedings of 2008 Fourth International Conference on Natural Computation, Jinan, China, 2008, pp. 328–332.

[14] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Pennsylvania, USA, 1992.

[15] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, Journal of Business and Economic Statistics 13 (1995) 253–263.

[16] J.H. Friedman, Multivariate adaptive regression splines (with discussion), The Annals of Statistics 19 (1991) 1–141.

[17] R. Gençay, F. Selçuk, B. Whitcher, An Introduction to Wavelets and Other Filtering Methods in Finance and Economics, Academic Press, London, 2002.

[18] Z. Gonghui, J.L. Starck, J. Campbell, F. Murtagh, The wavelet transform for filtering financial data streams, Journal of Computational Intelligence in Finance 12 (1999) 18–35.

[19] J.W. Hall, Adaptive selection of U.S. stocks with neural nets, in: G.J. Deboeck (Ed.), Trading on the Edge: Neural, Genetic and Fuzzy Systems for Chaotic Financial Markets, Willey, New York, 1994, pp. 45–65.

[20] S.C. Huang, T.K. Wu, Combining wavelet-based feature extractions with relevance vector machines for stock index forecasting, Expert Systems 25 (2) (2008) 133–149.

[21] J.S. Jang, ANFIS: Adaptive-Network-based Fuzzy Inference Systems, IEEE Transactions on Systems, Man, and Cybernetics 23 (3) (1993) 665–685.

[22] L. Khansa, D. Liginlal, Predicting stock market returns from malicious attacks: a comparative analysis of vector autoregression and time-delayed neural networks, Decision Support Systems 51 (4) (2011) 745–759.

[23] P.A.W. Lewis, J.G. Stevens, Nonlinear modeling of time series using multivariate adaptive regression splines (MARS), Journal of the American Statistical Association 86 (2009) 864–877.

[24] L. Li, Q. Li, S. Zhu, M. Ogihara, A survey on wavelet applications in data mining, SIGKDD Explorations 4 (2) (2003) 49–68.

[25] C.J. Lin, C.W. Hsu, C.C. Chang, A practical guide to support vector classification, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003.

[26] C.J. Lu, T.S. Lee, C.C. Chiu, Financial time series forecasting using independent component analysis and support vector regression, Decision Support Systems 47 (2) (2009) 115–125.

[27] C.J. Lu, T.S. Lee, C.M. Lian, Sales forecasting for computer wholesalers: a comparison of multivariate adaptive regression splines and artificial neural networks, Decision Support Systems (2012), http://dx.doi.org/10.1016/j.dss.2012.08.006.

[28] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 674–693.

[29] M. Misiti, Y. Misiti, G. Oppenheim, J.M. Poggi, MATLAB Wavelet Toolbox User's Guide, The MathWorks Inc., Massachusetts, USA, 1996.

[30] A. Ozun, A. Cifter, Modeling long-term memory effect in stock prices. A comparative analysis with GPH test and Daubechies wavelets, Studies in Economics and Finance 25 (1) (2008) 38–48.

[31] P.F. Pai, C.S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting, Omega 33 (2005) 497–505.

[32] Z. Pan, X. Wang, A stochastic nonlinear regression estimator using wavelets, Computational Economics 11 (1998) 89–102.

[33] D.B. Percival, A.T. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, Cambridge, UK, 2000.

[34] A.C. Pollock, A. Macaulay, M.E. Thomson, D. Önkal, Performance evaluation of judgmental directional exchange rate predictions, International Journal of Forecasting 21 (3) (2005) 473–489.

[35] J.B. Ramsey, Wavelets in economics and finance: past and future, Studies in Nonlinear Dynamics & Econometrics 6 (2002) 1–27.

[36] R.M. Rao, A.S. Bopardikar, Wavelet Transforms: Introduction to Theory and Applications, Addison Wesley, Boston, 1998.

[37] Y. Shahriar, W. Ilona, R. Dominik, Wavelet-based prediction of oil prices, Chaos, Solitons and Fractals 25 (2005) 265–275.

[38] T. Shin, I. Han, Optimal signal multi-resolution by genetic algorithms to support artificial neural networks for exchange-rate forecasting. Optimal signal multi-resolution by genetic algorithms to support artificial neural networks for exchange-rate forecasting, Expert Systems with Applications 18 (2002) 257–269.

[39] J.L. Starck, F. Murtagh, A. Bijaoui, Image and Data Analysis: The Multiscale Approach, Cambridge University Press, Cambridge, UK, 1998.

[40] N.R. Swanson, H. White, Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models, International Journal of Forecasting 13 (1997) 437–461.

[41] F.E.H. Tay, L.J. Cao, Application of support vector machines in financial time series forecasting, Omega 29 (2001) 309–317.

[42] F.E.H. Tay, L.J. Cao, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Transactions on Neural Networks 14 (2003) 1506–1518.

[43] D.M. Tsai, C.H. Chiang, Automatic band selection for wavelet reconstruction in the application of defect detection, Image and Vision Computing 21 (2003) 413–431.

[44] C.F. Tsai, Y.C. Hsiao, Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches, Decision Support Systems 50 (1) (2010) 258–269.

[45] M. Unser, Texture classification and segmentation using wavelet frames, IEEE Transactions on Image Processing 4 (11) (1995) 1549–1560.

[46] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 2000.

[47] W. Xiao, Q. Zhao, Q. Fei, A comparative study of data mining methods in consumer loans credit scoring management, Journal of Systems Science and Systems Engineering 15 (4) (2006) 419–435.

[48] S.A.M. Yaser, A.F. Atiya, Introduction to financial forecasting, Applied Intelligence 6 (1996) 205–213.

[49] S. Yousefi, I. Weinreich, D. Reinarz, Wavelet-based prediction of oil prices, Chaos, Solitons and Fractals 25 (2) (2005) 265–275.

[50] B.L. Zhang, R. Coggins, M.A. Jabri, D. Dersch, B. Flower, Multiresolution forecasting for futures trading using wavelet decompositions, IEEE Transactions on Neural Networks 12 (2001) 765–775.

[51] Y. Zhao, Y. Zhang, C. Qi, Prediction Model of Stock Market Returns Based on Wavelet Neural Network, in: Proceedings of 2008 Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, China, 2008, pp. 31–36.

[52] Y. Zhou, H. Leung, Predicting object-oriented software maintainability using multivariate adaptive regression splines, Journal of Systems and Software 80 (8) (2007) 1349–1361.

Ling-Jing Kao is an assistant professor in the Department of Business Management at National Taipei University of Technology, Taiwan. Her Ph.D. is from The Ohio State University in Marketing. Her research and teaching interests are in the area of application of Bayesian statistical approach and Quantitative Marketing Research. She has published articles in various journals, including Journal of the Operational Research Society, European Journal of Operational Research, and Expert System with Applications.

Chih-Chou Chiu is a professor in the Department of Business Management at National Taipei University of Technology. His Ph.D. is from Texas A&M University in Industrial Engineering. His research and teaching interests are in the area of application of artificial intelligence, Bayesian statistical approach, data mining and continuous process improvement techniques for manufacturing. He has published articles in various journals, including IIE Transactions, International Journal of Production Research, Journal of Intelligent Manufacturing, International Journal of System Science, and Quality and Reliability Engineering International.

Chi-Jie Lu is an associate professor in the Department of Industrial Management at Chien Hsin University of Science and Technology, Taiwan. He got his Ph.D. in Industrial Engineering and Management from Yuan-Ze University, Taiwan, in 2005. His research and teaching interests are in the area of data mining, time series forecasting, statistical process control, and machine vision and inspection. He has published articles in various journals, including Pattern Recognition, Decision Support Systems, Neurocomputing, Image and Vision Computing, International Journal of Production Economics, International Journal of Production Research and Computational Statistics and Data Analysis.

Chih-Hsiang Chang received his master degree from Institute of Commerce Automation and Management, National Taipei University of Technology, Taiwan, in 2010. His current research interests include machine learning, wavelet transform and stock index forecasting.