Information Sciences xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

A Morphological-Rank-Linear evolutionary method for stock market prediction

Ricardo de A. Araújo [a,*], Tiago A.E. Ferreira [b]

[a] Information Technology Department, $[gm]^2$ Intelligent Systems, Campinas, SP, Brazil
[b] Statistics and Informatics Department, Rural Federal University of Pernambuco, Recife, PE, Brazil

ARTICLE INFO

Article history:
Received 11 July 2008
Received in revised form 2 April 2009
Accepted 5 July 2009
Available online xxxx

Keywords:
Morphological-Rank-Linear filters
Genetic algorithms
Intelligent hybrid models
Financial time series
Stock market prediction

ABSTRACT

This work presents an evolutionary morphological-rank-linear approach in order to overcome the random walk dilemma for financial time series forecasting. The proposed Evolutionary Morphological-Rank-Linear Forecasting (EMRLF) method consists of an intelligent hybrid model composed of a Morphological-Rank-Linear (MRL) filter combined with a Modified Genetic Algorithm (MGA), which performs an evolutionary search for the minimum number of relevant time lags capable of a fine tuned characterization of the time series, as well as for the initial (sub-optimal) parameters of the MRL filter. Then, each individual of the MGA population is improved using the Least Mean Squares (LMS) algorithm to further adjust the parameters of the MRL filter, supplied by the MGA. After built the prediction model, the proposed method performs a behavioral statistical test with a phase fix procedure to adjust time phase distortions that can appear in the modeling of financial time series. An experimental analysis is conducted with the method using four real world stock market time series according to a group of performance metrics and the results are compared to both MultiLayer Perceptron (MLP) networks and a more advanced, previously introduced, Time-delay Added Evolutionary Forecasting (TAEF) method.

1. Introduction

Financial time series forecasting is considered a rather difficult problem, due to the many complex features frequently present in time series, such as irregularities, volatility, trends and noise. For such, a widely number of linear and nonlinear statistical models have been proposed in order to predict future tendencies of financial phenomena based on present and past historical data [4,32,26,31,33,8].

Alternatively, approaches based on Artificial Neural Networks (ANNs) have been successfully proposed for nonlinear modeling of time series in the last two decades ([25,7,42,18,38,5,15,30,41,12]). However, in order to define a solution to a given problem, ANNs require the setting up of a series of system parameters, some of them are not always easy to determine. The ANN topology, the number of processing units, the algorithm for ANN training (and its corresponding variables) are just some of the parameters that require definition [9]. In this context, hybrid intelligent approaches have produced interesting results [19,22,1,3,9].

However, a dilemma arises from all these models regarding financial time series, known as random walk dilemma [38], where the predictions generated by such models show a characteristic one step delay regarding original time series data. This behavior has been seen as a dilemma regarding the financial time series representation, where it has been posed that the series follow a random walk like model and cannot, therefore, be predicted [20].

[*] Corresponding author.
E-mail addresses: ricardo@gm2.com.br (R. d. A. Araújo), tiago@deinfo.ufrpe.br (T.A.E. Ferreira).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.07.007

In this context, this work presents an Evolutionary Morphological-Rank-Linear Approach to overcome the random walk dilemma. The proposed Evolutionary Morphological-Rank-Linear Forecasting (EMRLF) method is inspired on Takens theorem [39] and consists of an intelligent hybrid model composed of a Morphological-Rank-Linear (MRL) [27] with a Modified Genetic Algorithm (MGA) [19], which searches for the particular time lags capable to optimally characterize the time series and estimates the initial (sub-optimal) parameters of the MRL filter (mixing parameter (λ), rank (r), linear FIR filter ($\underline{b}$) and the MR filter ($\underline{a}$) coefficients). Then, each individual of the MGA population is improved by the Least Mean Squares (LMS) algorithm to further adjust the MRL filter parameters supplied by the MGA. After model training, the EMRLF method chooses the most fitted forecasting model, and performs a behavioral statistical test [9] in the attempt to adjust time phase distortions observed in financial time series.

An experimental analysis is conducted with the proposed method using four real world stock market time series (Directv Group Inc Stock Prices, Microsoft Corporation Stock Prices, Petrobras Company Stock Prices and Yahoo Inc Stock Prices), employing five well-known performance metrics to assess the performance of the method. The results achieved by the EMRLF method have shown a much better performance when compared to MultiLayer Perceptron (MLP) networks, and a better performance when compared to a previous hybrid model, named the Time-delay Added Evolutionary Forecasting (TAEF) method [9].

This paper is organized as follows. Section 2 presents the fundamentals of time series forecasting, the MGA description and the concepts of the MRL filter. Section 3 describes the proposed EMRLF method. Section 4 shows the performance measures used to compare assertiveness of all methods. In Section 5, simulations and experimental results are described with the EMRLF method, MLP networks and the TAEF algorithm [9] for two relevant stock market time series: the Coca-Cola Company and Microsoft Corporation Stock Prices. Section 6 presents, to conclude, the final remarks of this work.

2. Fundamentals

2.1. The time series prediction problem

A time series is a sequence of observations about a given phenomenon, where it is observed in discrete or continuous space. In this work all time series will be considered time discrete and equidistant.

Usually, a time series can be defined by

$$X_t = \{x_t \in \mathbb{R} | t = 1, 2, \dots, N\}, \tag{1}$$

where t is the temporal index and N is the number of observations. The term X_t will be seen as a set of temporal observations of a given phenomenon, orderly sequenced and equally spaced.

The aim of prediction techniques applied to a given time series (X_t) is to provide a mechanism that allows, with certain accuracy, the prediction of the future values of X_t, given by $X_{t+k},\ k = 1, 2, \dots$, where k represents the prediction horizon. These prediction techniques will try to identify certain regular patterns present in the data set, creating a model capable of generating the next temporal patterns, where, in this context, a most relevant factor for an accurate prediction performance is the correct choice of the past window, or the time lags, considered for the representation of a given time series.

Box and Jenkins [4] shown that when there is a clear linear relationship among the historical data of a given time series, the functions of auto-correlation and partial auto-correlation are capable of identifying the relevant time lags to represent a time series, and such procedure is usually applied in linear models. However, when it uses a real world time series, or more specifically, a complex time series with all their dependencies on exogenous and uncontrollable variables, the relationship that involves the time series historical data is generally nonlinear, which makes the Box and Jenkins' analysis procedure of the time lags only a crude estimate.

In mathematical sense, such relationship involving time series historical data defines a d-dimensional phase space, where d is the minimum dimension capable of representing such relationship. Therefore, a d-dimensional phase space can be built so that it is possible to unfold its corresponding time series. Takens [39] proved that if d is sufficiently large, such phase space is homeomorphic to the phase space that generated the time series. Takens' Theorem [39] is the theoretical justification that it is possible to build a state space using the correct time lags, and if this space is correctly rebuilt, Takens' Theorem [39] also guarantees that the dynamics of this space is topologically identical to the dynamics of the real system state space.

The main problem in reconstructing the original state space is naturally the correct choice of the variable d, or more specifically, the correct choice of the important time lags necessary for the characterization of the system dynamics. Many proposed methods can be found in the literature for the definition of the lags [36,28,40]. Such methods are usually based on measures of conditional probabilities, which consider,

$$X_t = f(x_{t-1}, x_{t-2}, \dots, x_{t-d}) + r_t, \tag{2}$$

where $f(x_{t-1}, x_{t-2}, \dots, x_{t-d})$ is a possible mapping of the pasts values to the facts of the future (where x_{t-1} is the lag 1, x_{t-2} is the lag 2, $\dots, x_{t-d}$ is the lag d) and r_t is a noise term.

However, in general, these tests found in the literature are based on the primary dependence among the variables and do not consider any possible induced dependencies. For example, if

$$f(x_{t-1}) = f(f(x_{t-2})), \tag{3}$$

it is said that x_{t-1} is the primary dependence, and the dependence induced on x_{t-2} is not considered (any variable that is not a primary dependence is denoted as irrelevant).

The method proposed in this paper, conversely, does not make any prior assumption about the dependencies between the variables. In other words, it does not discard any possible correlation that can exist among the time series parameters, even higher order correlations, since it carries out an iterative automatic search for solving the problem of finding the relevant time lags.

2.2. The random walk dilemma

A naive prediction strategy is to define the last observation of a time series as the best prediction of its next future value ($X_{t+1} = X_t$). This kind of model is known as the Random Walk (RW) model [23], which is defined by

$$X_t = X_{t-1} + r_t, \tag{4}$$

or

$$\mathtt{D}X_t = X_t - X_{t-1} = r_t, \tag{5}$$

where X_t is the current observation, X_{t-1} is the immediate observation before X_t, and r_t is a noise term with a gaussian distribution of zero mean and standard deviation σ ($r_t \approx N(0, \sigma)$). In other words, the rate of time series change ($\mathtt{D}X_t$) is a white noise.

The model above clearly implies that, as the information set consists of past time series data, the future data are unpredictable. On the average, the value X_t is indeed the best prediction of value X_{t-1}. This behavior is common in the finance market and in the economic theory and it is so-called random walk dilemma or random walk hypothesis [23].

The computational cost for time series forecasting using the random walk dilemma is extremely low. Therefore, any other prediction method more costly than a random walk model should have a very superior performance than a random walk model, otherwise its use is not interesting in the practice.

However, if the time series phenomenon is driven by law with strong similarity to a random walk model, any model applied to this time series phenomenon will tend to have the same performance than a random walk model.

Assuming that an accurate prediction model is used to build an estimated value of X_t, denoted by $\widehat{X_t}$, the expected value ($E[\cdot]$) of the difference between $\widehat{X_t}$ and X_t must tends to zero,

$$E\left[\widehat{X_t} - X_t\right] \to 0. \tag{6}$$

If the time series generator phenomenon is supposed to have a strong random walk linear component and a very weak nonlinear component (denoted by $g(t)$), and assuming that $E[r_t] = 0$ and $E[r_t r_k] = 0 (\forall k \neq t)$, the expected value of the difference between $\widehat{X_t}$ and X_t will be

$$\begin{aligned}
&E\left[\widehat{X_t} - (X_{t-1} + g(t) + r_t)\right] \to 0\\
&E\left[\widehat{X_t}\right] - E[X_{t-1}] - E[g(t)] - E[r_t] \to 0\\
&E\left[\widehat{X_t}\right] - E[X_{t-1}] - E[g(t)] \to 0\\
&E\left[\widehat{X_t}\right] \to E[X_{t-1}] + E[g(t)].
\end{aligned}$$

But $E[X_{t-1}] \gg E[g(t)]$, then $E[X_{t-1}] + E[g(t)] \simeq E[X_{t-1}]$ and

$$E\left[\widehat{X_t}\right] \to E[X_{t-1}]. \tag{7}$$

Therefore, in these conditions, to escape of random walk dilemma is a hard task. Indications of this behavior (strong linear random walk component and a weak nonlinear component) can be observed from time series lagplot graphics. For example, lagplot graphics where strong linear structures are dominant with respect to nonlinear structures [17], generally observed in the financial and economical time series.

2.3. Filtering systems

A powerful class of nonlinear systems that can successfully solve nonlinear problems comes from Mathematical Morphology (MM) [21,37]. Morphological filters are nonlinear signal transformations that locally modify the geometrical features of signals [21], and are related to the basic operations of set theory and integral geometry.

These kind of filters employ specific sequences of neighborhood transformations in order to measure useful geometric signal features [14]. In this context, Pessoa and Maragos [27] proposed a new filter, referred to as Morphological-Rank-Linear (MRL) filter, which consists of a linear combination between a Morphological-Rank (MR) filter [35,34] and a linear Finite Impulse Response (FIR) filter [27].

Please cite this article in press as: R. d. A. Araújo, T.A.E. Ferreira, A Morphological-Rank-Linear evolutionary method for stock market prediction, Inform. Sci. (2009), doi:10.1016/j.ins.2009.07.007

2.3.1. Morphological-Rank-Linear filter preliminaries

Definition 1. *Rank Function*: the rth rank function of the vector $\underline{t} = (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n$ is the rth element of the vector $\underline{t}$ sorted in decreasing order ($t_{(1)} \geqslant t_{(2)} \geqslant \ldots \geqslant t_{(n)}$). It is denoted by [27]

$$\mathscr{R}_r(\underline{t}) = t_{(r)}, \quad r = 1, 2, \ldots, n. \tag{8}$$

For example, given the vector $\underline{t} = (3, 0, 5, 7, 2, 1, 3)$, its 4th rank function is $\mathscr{R}_4(\underline{t}) = 3$.

Definition 2. *Unit Sample Function*: the unit sample function is given by [27]

$$q(v) = \begin{cases} 1, & \text{if } v = 0, \\ 0, & \text{otherwise}. \end{cases} \tag{9}$$

where $v \in \mathbb{R}$.

Applying the unit sample function to a vector $\underline{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$, yields a vector unit sample function ($Q(\underline{v})$), given by [27]

$$Q(\underline{v}) = [q(v_1), q(v_2), \ldots, q(v_n)]. \tag{10}$$

Definition 3. *Rank Indicator Vector* : the rth rank indicator vector $\underline{c}$ of $\underline{t}$ is given by [27]

$$\underline{c}(\underline{t}, r) = \frac{Q((z \cdot \underline{1}) - \underline{t})}{Q((z \cdot \underline{1}) - \underline{t}) \cdot \underline{1}'}, \tag{11}$$

where $z = \mathscr{R}_r(\underline{t})$, $\underline{1} = (1, 1, \ldots, 1)$ and the symbol $'$ denotes transposition.

For example, given the vector $\underline{t} = (3, 0, 5, 7, 2, 1, 3)$, its 4th rank indicator function is $\underline{c}(\underline{t}, 4) = \frac{1}{2}(1, 0, 0, 0, 0, 0, 1)$.

Definition 4. *Smoothed Rank Function*: the smoothed rth rank function is given by [27]

$$\mathscr{R}_{r,\sigma}(\underline{t}) = \underline{c}_\sigma(\underline{t}, r) \cdot \underline{t}', \tag{12}$$

with

$$\underline{c}_\sigma(\underline{t}, r) = \frac{Q_\sigma((z \cdot \underline{1}) - \underline{t})}{Q_\sigma((z \cdot \underline{1}) - \underline{t}) \cdot \underline{1}'}, \tag{13}$$

where c_σ is an approximation for the rank function $\underline{c}$ and $Q_\sigma(\underline{v})$=$[q_\sigma(v_1), q_\sigma(v_2), \ldots, q_\sigma(v_n)]$ is a smoothed impulse function (where $q_\sigma(v)$ is like $sech^2(v/\sigma)$ or $exp\left[-\frac{1}{2}(v/\sigma)^2\right]$) and $\sigma \geqslant 0$ is a scale parameter.

Term $\underline{c}_\sigma$ is an approximation for the rank indicator vector $\underline{v}$. Using ideas of fuzzy set theory, $\underline{c}_\sigma$ can also be interpreted as a membership function vector [27]. For example, if the vector $\underline{t} = (3, 0, 5, 7, 2, 1, 3)$, $q_\sigma(v) = sech^2\left(\frac{v}{\sigma}\right)$ and $\sigma = 0.5$ then its smoothed 4th rank indicator function is

$$\underline{c}_\sigma(\underline{t}, 4) = \frac{1}{2}(0.9646, 0, 0.0013, 0, 0.0682, 0.0013, 0.9646),$$

where $\underline{c}(\underline{t}, 4) = \frac{1}{2}(1, 0, 0, 0, 0, 0, 1)$.

2.3.2. Morphological-Rank-Linear (mrl) filter

The MRL filter [27] is a linear combination between a Morphological-Rank (MR) filter [35,34] and a linear Finite Impulse Response (FIR) filter [27].

Definition 5. *MRL Filter* [27]: Let $\underline{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ represent the input signal inside an n-point moving window and let y be the output from the filter. Then, the MRL filter is defined as the shift-invariant system whose local signal transformation rule $\underline{x} \to y$ is given by [27]

$$y = \lambda\alpha + (1 - \lambda)\beta, \tag{14}$$

with

$$\alpha = \mathscr{R}_r(\underline{x} + \underline{a}) = \mathscr{R}_r(x_1 + a_1, x_2 + a_2, \cdots, x_n + a_n), \tag{15}$$

and

$$\beta = \underline{x} \cdot \underline{b}' = x_1 b_1 + x_2 b_2 + \cdots + x_n b_n, \tag{16}$$

where $\lambda \in \mathbb{R}$, $\underline{a}$ and $\underline{b} \in \mathbb{R}^n$. Terms $\underline{a} = (a_1, a_2, \ldots, a_n)$ and $\underline{b} = (b_1, b_2, \ldots, b_n)$ represent the coefficients of the MR filter and the coefficients of the linear FIR filter, respectively. Term $\underline{a}$ is usually referred to "structuring element" because for $r = 1$ or $r = n$

the rank filter becomes the morphological dilation and erosion by a structuring function equal to $\pm\underline{a}$ within its support [27]. The structure of the MRL filter is illustrated in Fig. 1.

2.3.3. Morphological-Rank-Linear (mrl) filter adaptive design

Pessoa and Maragos [27] presented an adaptive design of MRL filters based on the LMS algorithm [35,34], the "rank indicator vector" [27] and "smoothed impulses" [27] for overcoming the problem of nondifferentiability of rank operations.

Pessoa and Maragos [27] have shown that the main goal of the MRL filter is to specify a set of parameters $(\underline{a}, \underline{b}, r, \lambda)$ according to some design requirements. However, instead of using the integer rank parameter r directly in the MRL filter definition Eqs. (14)–(16), they argued that it is possible to work with a real variable ρ implicitly defined through the following rescaling [27]:

$$r = \text{round}\left(n - \frac{n-1}{exp(-\rho)}\right), \tag{17}$$

where $\rho \in \mathbb{R}$, n is the dimension of the input signal vector $\underline{x}$ inside the moving window and round$(\cdot)$ denotes the usual symmetrical rounding operation. In this way, the weight vector to be used in the filter design task is defined by [27]

$$\underline{w} \equiv (\underline{a}, \underline{b}, \rho, \lambda). \tag{18}$$

The framework of the MRL filter adaptive design is viewed as a learning process where the filter parameters are iteratively adjusted. The usual approach to adaptively adjust the vector $\underline{w}$, and therefore design the filter, is to define a cost function $J(\underline{w})$, estimate its gradient $\nabla J(\underline{w})$, and update the vector $\underline{w}$ by the iterative formula

$$\underline{w}(i+1) = \underline{w}(i) - \mu_0 \nabla J(\underline{w}), \tag{19}$$

where $\mu_0 > 0$ (usually called step size) and $i \in \{1, 2, \ldots\}$. The term μ_0 is responsible for regulating the tradeoff between stability and speed of convergence of the iterative procedure. The iteration of Eq. (19) starts with an initial guess $\underline{w}(0)$ and stops when some desired condition is reached. This approach is known as the method of gradient steepest descent [27].

The cost function J must reflect the solution quality achieved by the parameters configuration of the system. A cost function J, for example, can be any error function, such as

$$J[\underline{w}(i)] = \frac{1}{M} \sum_{k=i-M+1}^{i} e^2(k), \tag{20}$$

where $M \in \{1, 2, \ldots\}$ is a memory parameter and $e(k)$ is the instantaneous error, given by

$$e(k) = d(k) - y(k), \tag{21}$$

where $d(k)$ and $y(k)$ are the desired output signal and the actual filter output for the training sample k, respectively. The memory parameter M controls the smoothness of the updating process. If we are processing noiseless signals, $M = 1$ is recommended [27]. However, when we use $M > 1$, the updating process tends to reduce the noise influence of noisy signals during the training [27].

Hence, the resulting adaptation algorithm is given by [27]

$$\underline{w}(i+1) = \underline{w}(i) + \frac{\mu}{M} \sum_{k=i-M+1}^{i} e^2(k) \frac{\partial y(k)}{\partial \underline{w}}, \tag{22}$$

where $\mu = 2\mu_0$ and $i \in \{1, 2, \ldots\}$. From Eqs. (14), (15), (16) and (18), term $\frac{\partial y(k)}{\partial \underline{w}}$ [27] may be calculated as

$$\frac{\partial y}{\partial \underline{w}} = \left(\frac{\partial y}{\partial \underline{a}}, \frac{\partial y}{\partial \underline{b}}, \frac{\partial y}{\partial \rho}, \frac{\partial y}{\partial \lambda}\right) \tag{23}$$

with

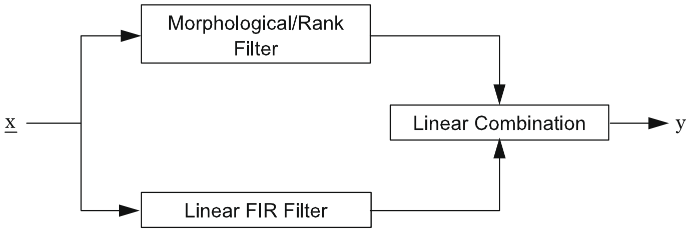

Fig. 1. Structure of the MRL filter.

$$\frac{\partial y}{\partial \underline{a}} = \lambda \frac{\partial \alpha}{\partial \underline{a}}, \tag{24}$$

$$\frac{\partial y}{\partial \underline{b}} = (1-\lambda)\underline{x}, \tag{25}$$

$$\frac{\partial y}{\partial \rho} = \lambda \frac{\partial \alpha}{\partial \rho}, \tag{26}$$

$$\frac{\partial y}{\partial \lambda} = (\alpha - \beta), \tag{27}$$

where

$$\frac{\partial \alpha}{\partial \underline{a}} = \underline{c} = \frac{Q((\alpha \cdot \underline{1}) - \underline{x} - \underline{a})}{Q((\alpha \cdot \underline{1}) - \underline{x} - \underline{a}) \cdot \underline{1}'}, \tag{28}$$

$$\frac{\partial \alpha}{\partial \rho} = 1 - \frac{1}{n} Q((\alpha \cdot \underline{1}) - \underline{x} - \underline{a}) \cdot \underline{1}', \tag{29}$$

where n is the dimension of $\underline{x}$ and $\alpha = \mathcal{R}_r(\underline{x} + \underline{a})$.

It is important to mention that the unit sample function Q is frequently replaced by smoothed impulses Q_σ, in which case an appropriate smoothing parameter σ should be selected (which will affect only the gradient estimation step in the design procedure [27]).

3. The proposed method

The proposed Evolutionary Morphological-Rank-Linear Forecasting (EMRLF) method consists of an intelligent hybrid model, which uses a Modified Genetic Algorithm (MGA) [19] to adjust the initial MRL filter parameters and then it uses the LMS algorithm to further improve the parameters supplied by the MGA. The advantage of those models is that not only they have linear and nonlinear components, but are quite attractive due to their simpler computational complexity when compared to other approaches such the model introduced by Araújo et al. [1,2], other ANN-GA models [9] and other statistical models [32,26,31,33].

The EMRLF method is based on the definition of the two main elements necessary for building an accurate forecasting system according to Ferreira [9]: (a) the minimum number of time lags adequate for representing the time series, and (b) the model structure capable of representing such underlying information for the purpose of prediction. It is important to consider the minimum number of time lags because the larger the number of lags, the larger the cost associated with the model training.

Following this principle, the EMRLF model uses the MGA [19] to adjust the MRL filter and the LMS algorithm [27] to train it. The purpose of using the MGA [19] is to identify the following important parameters: (1) the minimum number of time lags and their corresponding specific positions to represent the time series (initially, a maximum number of lags (*MaxLags*) is defined and then the MGA can choose any value in the interval $[1, MaxLags]$ for each individual of the population), and (2) the initial (sub-optimal) parameters of the MRL filter (mixing parameter (λ), the rank (r), the linear Finite Impulse Response (FIR) filter ($\underline{b}$) and the Morphological-Rank (MR) filter ($\underline{a}$) coefficients). The LMS algorithm [27] is then used to train each individual of the MGA population, since it has proved to be effective in speeding up the training process while limiting its computational complexity.

The MGA used here is based on the work of Leung et al. [19]. The MGA is a second version of the Simple Genetic Algorithm (SGA) [10,16,24] that was modified to improve search convergence. The SGA was firstly studied, and, then, was modified to accelerate its convergence through the use of modified crossover and mutation operators (described later). The algorithm is described in Fig. 2.

According to Fig. 2, the MGA procedure consists of selecting a parent pair of chromosomes and then performing crossover and mutation operators (generating the offspring chromosomes–the new population) until the termination condition is reached; then the best individual in the population is selected as a solution to the problem.

The crossover operator is used for exchanging information from two parents (vectors $\underline{p}_1$ and $\underline{p}_2$) obtained in the selection process by a roulette wheel approach [19]. The recombination process to generate the offsprings (vectors $\underline{C}_1, \underline{C}_2, \underline{C}_3$ and $\underline{C}_4$) is done by four crossover operators, which are defined by the following equations [19]:

$$\underline{C}_1 = \frac{\underline{p}_1 + \underline{p}_2}{2}, \tag{30}$$

$$\underline{C}_2 = \underline{p}_{max}(1-w) + max(\underline{p}_1, \underline{p}_2)w, \tag{31}$$

$$\underline{C}_3 = \underline{p}_{min}(1-w) + min(\underline{p}_1, \underline{p}_2)w, \tag{32}$$

$$\underline{C}_4 = \frac{(\underline{p}_{max} + \underline{p}_{min})(1-w) + (\underline{p}_1 + \underline{p}_2)w}{2}, \tag{33}$$

where $w \in [0, 1]$ denotes the crossover weight (the closer w is to 1, the greater is the direct contribution from parents), $max(\underline{p}_1, \underline{p}_2)$ and $min(\underline{p}_1, \underline{p}_2)$ denotes the vector whose elements are the maximum and the minimum, respectively, between

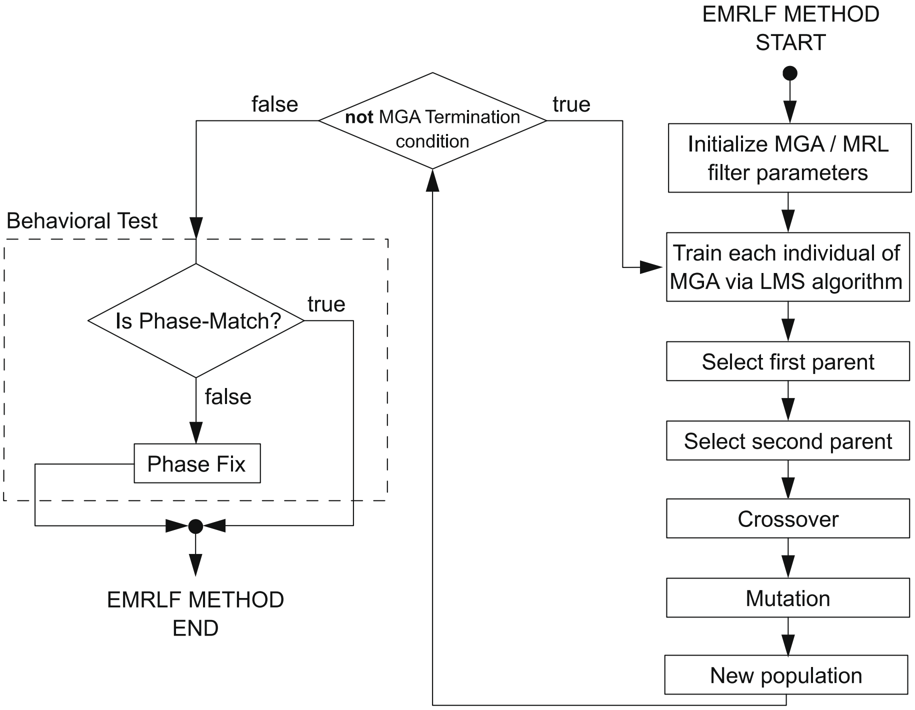

Fig. 3. The proposed method.

In order to provide a more robust forecasting model, a multi-objective fitness function is defined, resulting from a combination of five well-known performance measures: the Prediction Of Change In Direction (POCID), the Mean Square Error (MSE), the Mean Absolute Percentage Error (MAPE), the Normalized Mean Square Error (NMSE) and the Average Relative Variance (ARV). Those measures will be formally defined in Section 4. The function is given by:

$$\text{Fitness function} = \frac{\text{POCID}}{1 + \text{MSE} + \text{MAPE} + \text{NMSE} + \text{ARV}}. \tag{35}$$

It can be noticed that there are linear and nonlinear metrics in the function and each one of them can contribute differently to the evolution process. Eq. (35) was built, accordingly, to have all information necessary to estimate the time series generator phenomenon and was tested empirically.

After model training (the end of EMRLF method's iterations), the proposed method uses the phase fix procedure introduced by Ferreira et al. [9] in the TAEF method, to adjust time phase distortions observed ("out-of-phase" matching) in financial time series. Ferreira et al. [9] have shown that the representation of some time series (natural phenomena) were developed by the model with a very close approximation between the actual and the predicted values of the series (referred to as "in-phase" matching), whereas the predictions of others (mostly financial time series) were always presented with a one step delay, regarding the original data (referred to as "out-of-phase" matching).

The EMRLF method uses the statistical test (t-test) to check if the MRL model has reached an in-phase or out-of-phase matching by conducting a comparison between the outputs of the predictive model and the actual series, making use of the validation data set. This comparison is a simple hypothesis test, where the null hypothesis is that the prediction corresponds to in-phase matching and the alternative hypothesis is that the prediction does not correspond to in-phase matching (corresponds to out-of-phase matching). If this test accepts the in-phase matching hypothesis, the elected model is ready for practical use. Otherwise, the EMRLF method performs a two step procedure to adjust the relative phase between the prediction and the actual time series (shown in Fig. 4): (i) the validation patterns are presented to the MRL filter and the outputs are re-arranged to create new input patterns (reconstructed patterns), and (ii) the reconstructed patterns are presented the same MRL filter and the output is set as the final predictive response. This procedure considers that the MRL filter does not behave like a random walk, but it shows a peculiar behavior approximated to a random walk: the $t+1$ prediction is taken as the t value (the random walk dilemma). If the MRL filter were a random walk model, the phase adjustment procedure would not be capable of correcting the time phase.

The termination conditions for the MGA are [29]:

(1) The maximum number of epochs.
(2) The increase in the validation data error or generalization loss (Gl) beyond 5%.
(3) The decrease in the training data error (Pt) below 10^{-6}.

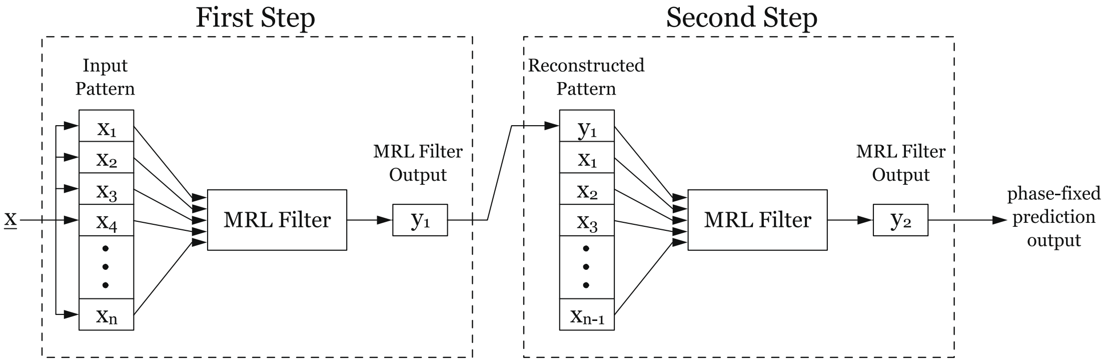

Fig. 4. Phase fix procedure.

Each individual of the MGA population is an MRL filter. The individuals are represented by chromosomes that have the following genes (MRL filter parameters):

- $\underline{a}$: MR filter coefficients,
- $\underline{b}$: linear FIR filter coefficients,
- ρ: variable used to determine the rank r,
- λ: mixing parameter,
- $\underline{lag}$: a vector having size *MaxLags*, where each position has a real-valued codification, which is used to determine whether a specific time lag will be used ($\text{lag}_i \geqslant 0$) or not ($\text{lag}_i < 0$).

4. Performance evaluation

Many performance evaluation criteria are found in literature. However, most of the existing literature on time series prediction frequently employ only one performance criterion for prediction evaluation. The most widely used performance criterion is the Mean Squared Error (MSE), given by

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}(\text{target}_j - \text{output}_j)^2, \tag{36}$$

where N is the number of patterns, target_j is the desired output for pattern j and output_j is the predicted value for pattern j.

The MSE measure may be used to drive the prediction model in the training process, but it cannot be considered alone as a conclusive measure for comparison of different prediction models [6]. For this reason, other performance criteria should be considered for allowing a more robust performance evaluation.

A measure that presents accurately identifying model deviations is the Mean Absolute Percentage Error (MAPE), given by

$$\text{MAPE} = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{\text{target}_j - \text{output}_j}{x_j}\right|, \tag{37}$$

where x_j is the time series value at point j.

The random walk dilemma can be used as a naive predictor ($X_{t+1} = X_t$), commonly applied to financial time series prediction. Thus, a way to evaluate the model regarding a random walk model is using the Normalized Mean Squared Error (NMSE) or U of Theil Statistic (THEIL) [13], which associates the model performance with a random walk model, and given by

$$\text{THEIL} = \frac{\sum_j^N(\text{target}_j - \text{output}_j)^2}{\sum_j^N(\text{target}_j - \text{target}_{j-1})^2}, \tag{38}$$

where, if the THEIL is equal to 1, the predictor has the same performance than a random model. If the THEIL is greater than 1, then the predictor has a performance worse than a random walk model, and if the THEIL is less than 1, the predictor is better than a random walk model. In the perfect model, the THEIL tends to zero.

Another interesting measure maps the accuracy in the future direction prediction of the time series or, more specifically, the ability of the method to predict if the future series value (prediction target) will increase or decrease with respect to the previous value. This metric is known as the Prediction Of Change In Direction (POCID) [9], and is given by

$$\text{POCID} = \frac{100}{N}\sum_{j=1}^{N}D_j, \tag{39}$$

where

$$D_j = \begin{cases} 1, & \text{if } (\text{target}_j - \text{target}_{j-1})(\text{output}_j - \text{output}_{j-1}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{40}$$

The last measure used associates the model performance with the mean of the time series. The measure is the Average Relative Variance (ARV), and given by

$$\text{ARV} = \frac{\sum_{j=1}^{N}(\text{target}_j - \text{output}_j)^2}{\sum_{j=1}^{N}(\text{output}_j - \overline{\text{target}})^2}, \tag{41}$$

where, $\overline{\text{target}}$ is the mean of the time series. If the ARV is equal to 1, the predictor has the same performance of the time series average prediction. If the ARV is greater than 1, then the predictor has a performance worse than the time series average prediction, and if the ARV is less than 1, the predictor is better than the time series average prediction. In the ideal model, ARV tends to zero.

5. Simulations and experimental results

A set of four financial time series was used as a test bed for evaluation of the EMRLF method: Directv Group Inc Stock Prices, Microsoft Corporation Stock Prices, Petrobras Company Stock Prices and Yahoo Inc Stock Prices. All series investigated were normalized to lie within the range $[0, 1]$ and divided into three sets according to Prechelt [29]: training set (first 50% of the points), validation set (second 25% of the points) and test set (third 25% of the points).

The MGA parameters used were the maximum number of GA generations, corresponding to 10^3, the crossover weight $(w = 0.9)$, the mutation probability $(p_{mut} = 0.1)$, the maximum number of lags $(MaxLags = 10)$, the maximum number of LMS training epochs $(E = 10^3)$, the MR filter coefficients and the linear FIR filter coefficients ($\underline{a}$ and $\underline{b}$, respectively), normalized in the range $[-0.5, 0.5]$, and the parameters λ and ρ, normalized in the range $[0, 1]$ and $[-MaxLags, MaxLags]$, respectively.

The simulation experiments involving the EMRLF model were conducted with and without the phase fix procedure [9], referred to as EMRLF out-of-phase model and EMRLF in-phase model, respectively. These two procedures (in-phase and out-of-phase) were used to study the possible performance improvement, in terms of fitness function, of the phase fix procedure applied to EMRLF model. For each time series, a number of ten model training repetitions were executed and the instance with the largest validation fitness function is chosen to represent the predictive model.

In order to establish a performance study, results previously published in the literature with the TAEF Method [9] on the same series and under the same conditions are employed for comparison of results. In addition, experiments with MultiLayer Perceptron (MLP) networks were used for comparison with the EMRLF method. In all experiments, ten random initializations for each model (MLP) were carried out, and the experiment with the largest validation fitness function was chosen to represent the predictive model. The Levenberg-Marquardt Algorithm [11] was employed for training the MLP network. For all the series, the best initialization was elected as the model to be beaten. The statistical behavioral test for phase fix was also applied to all the MLP models in order to guarantee a fair comparison between the models.

It is worth to mention that the results with ARIMA models were not presented in our comparative analysis since Ferreira [9] showed that MLP network obtained results better than ARIMA models, for all financial time series used in this work. In this way, it is used only MLP networks in our comparative analysis.

5.1. The Directv Group Inc Stock Prices series

The Directv Group Inc Stock Prices series corresponds to the daily records of the Directv Group Inc from March 28th 2005 to March 16th 2009, constituting a database of 1000 points.

For the prediction of the Directv Group Inc Stock Prices series (with one step ahead of prediction horizon), the proposed method automatically chose the lags $2, 4, 5$ and 9 as the relevant lags for the time series representation, defined the parameters $\rho = -0.4359$ and $\lambda = 0.0064$ and classified the model as "out-of-phase" matching. Table 1 shows the results (with respect to the test set) for all the performance measures for the MLP, TAEF and EMRLF models.

Fig. 5 shows the actual Directv Group Inc Stock Prices values (solid line) and the predicted values generated by the EMRLF model (dashed line) for the last 20 points of the test set.

5.2. The Microsoft Corporation Stock Prices series

The Microsoft Corporation Stock Prices series corresponds to the daily records of the Microsoft Corporation from March 28th 2005 to March 16th 2009, constituting a database of 1000 points.

For the prediction of the Microsoft Corporation Stock Prices series (with one step ahead of prediction horizon), the proposed method automatically chose the lags 2, 3, 4, 7, 9 and 10 as the relevant lags for the series representation, defined the

Please cite this article in press as: R. d. A. Araújo, T.A.E. Ferreira, A Morphological-Rank-Linear evolutionary method for stock market prediction, Inform. Sci. (2009), doi:10.1016/j.ins.2009.07.007

Table 1
Results for the Directv Group Inc Stock Prices series.

	MLP		TAEF		EMRLF	
	In-phase	Out-of-phase	In-phase	Out-of-phase	In-phase	Out-of-phase
MSE	2.6783e−3	2.8299e−3	2.5600e−3	9.4603e−5	2.7842e−4	6.4276e−6
MAPE	6.0643e−2	6.2741e−2	6.0547e−2	1.0220e−2	3.4999e−2	2.6190e−3
NMSE	1.0933	1.0833	1.0467	9.1081e−2	0.9994	2.4609e−3
ARV	8.9518e−3	8.4189e−2	7.6005e−2	3.6244e−3	5.6628e−3	1.9121e−4
POCID	48.40	48.38	48.59	90.79	48.80	97.38
Fitness	22.3497	21.6653	22.2297	82.1614	23.9176	96.8845

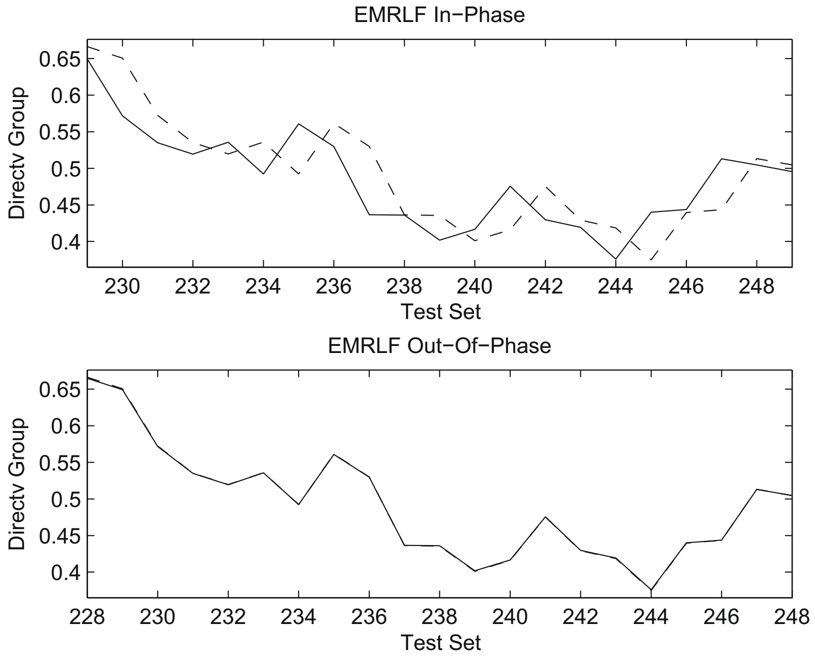

Fig. 5. Prediction results for the Directv Group Inc Stock Prices series (test set): actual values (solid line) and predicted values (dashed line).

parameters $\rho = 0.1577$ and $\lambda = 0.0028$ and classified the model as "out-of-phase" matching. Table 2 shows the results (with respect to the test set) for all the performance measures for the MLP, TAEF and EMRLF models.

Fig. 6 shows the actual Microsoft Corporation Stock Prices values (solid line) and the predicted values generated by the EMRLF model (dashed line) for the last 20 points of the test set.

5.3. The Petrobras Company Stock Prices series

The Petrobras Company Stock Prices series corresponds to the daily records of the Petrobras Company from May 02th 2005 to March 16th 2009, constituting a database of 1000 points.

For the prediction of the Petrobras Company Stock Prices series (with one step ahead of prediction horizon), the proposed method automatically chose the lags 2, 5 and 6 as the relevant lags for the series representation, defined the parameters

Table 2
Results for the Microsoft Corporation Stock Prices series.

	MLP		TAEF		EMRLF	
	In-phase	Out-of-phase	In-phase	Out-of-phase	In-phase	Out-of-phase
MSE	1.2427e−3	1.2337e−3	1.8622e−3	6.4230e−4	1.2384e−3	6.9356e−7
MAPE	0.1240	0.1235	0.2617	0.1920	0.1237	6.3467e−3
NMSE	1.0069	0.9995	1.4950	0.5115	1.0034	5.5234e−4
ARV	3.0371e−2	3.0153e−2	4.5512e−2	1.5838e−3	3.0267e−2	1.7101e−5
POCID	48.19	48.19	48.19	96.97	48.19	97.59
Fitness	22.2843	22.3683	17.1786	56.8497	22.3246	96.9196

Please cite this article in press as: R. d. A. Araújo, T.A.E. Ferreira, A Morphological-Rank-Linear evolutionary method for stock market prediction, Inform. Sci. (2009), doi:10.1016/j.ins.2009.07.007

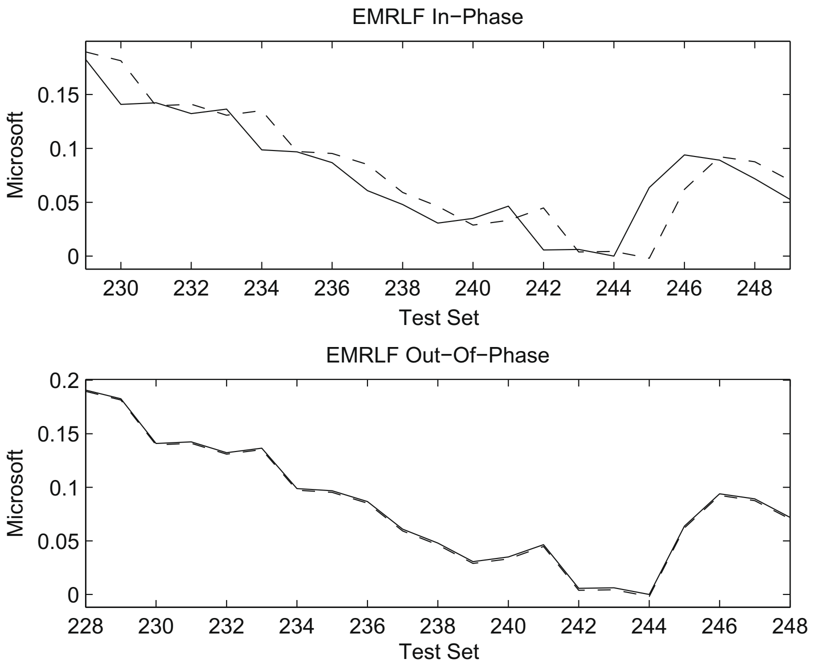

Fig. 6. Prediction results for the Microsoft Company Stock Prices series (test set): actual values (solid line) and predicted values (dashed line).

$\rho = 0.5321$ and $\lambda = 0.0011$ and classified the model as "out-of-phase" matching. Table 3 shows the results (with respect to the test set) for all the performance measures for the MLP, TAEF and EMRLF models.

Fig. 7 shows the actual Petrobras Company Stock Prices values (solid line) and the predicted values generated by the EMRLF model (dashed line) for the last 20 points of the test set.

5.4. The Yahoo Inc Stock Prices series

The Yahoo Inc Stock Prices series corresponds to the daily records of the Yahoo Inc from March 28th 2005 to March 16th 2009, constituting a database of 1000 points.

For the prediction of the Yahoo Inc Stock Prices series (with one step ahead of prediction horizon), the proposed method automatically chose the lags 2 and 9 as the relevant lags for the series representation, defined the parameters $\rho = 0.5360$ and $\lambda = 0.0019$ and classified the model as "out-of-phase" matching. Table 4 shows the results (with respect to the test set) for all the performance measures for the MLP, TAEF and EMRLF models.

Fig. 8 shows the actual Yahoo Inc Stock Prices values (solid line) and the predicted values generated by the EMRLF model (dashed line) for the last 20 points of the test set.

In general, all predictive models generated by the EMRLF have shown, using the phase fix procedure, forecasting performance much better than the MLP model and TAEF model. The EMRLF method was able to adjust the time phase distortions in all analyzed time series (the prediction generated by the out-of-phase matching hypothesis is not delayed with respect to the original data), while the MLP model was not able to adjust the time phase. This corroborates with the assumptions made by Ferreira [9], where it is discussed that the success of the phase fix procedure is strongly dependent on an accurate adjustment of the predictive model parameters and on the model itself used for forecasting.

Table 3
Results for the Petrobras Company Stock Prices series.

	MLP		TAEF		EMRLF	
	In-phase	Out-of-phase	In-phase	Out-of-phase	In-phase	Out-of-phase
MSE	6.6053e−4	5.6281e−4	6.2716e−4	1.1916e−4	5.4687e−4	7.5012e−7
MAPE	3.6130e−2	3.4104e−2	3.5429e−2	1.2451e−2	3.3463e−2	1.2638e−3
NMSE	1.2133	1.0373	1.1561	0.2188	1.0074	1.3758e−3
ARV	1.9703e−2	1.6832e−2	1.8756e−2	3.5543e−3	1.6355e−2	2.2375e−5
POCID	51.01	51.21	51.00	92.78	51.01	97.18
Fitness	22.4734	24.5165	23.0674	75.1301	24.7890	96.9219

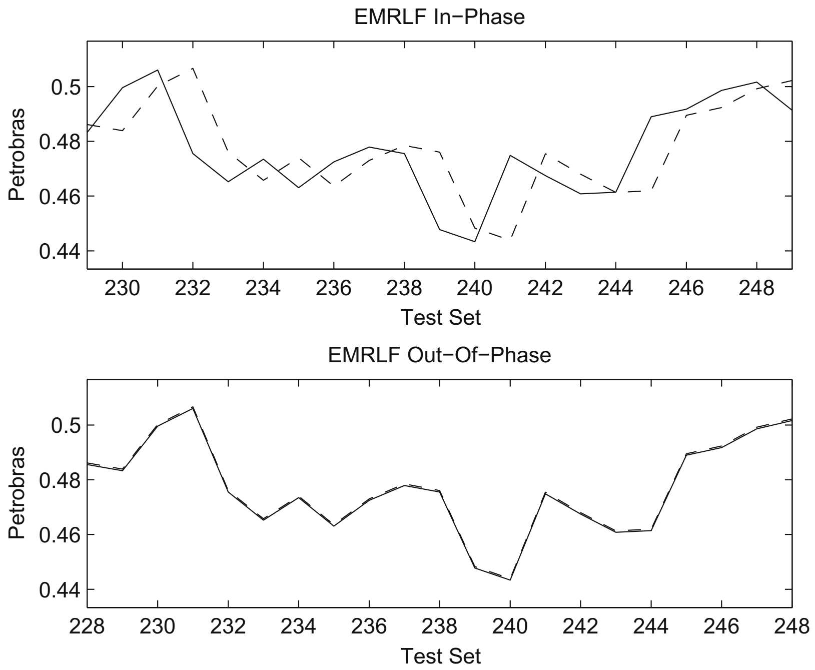

Fig. 7. Prediction results for the Petrobras Company Stock Prices series (test set): actual values (solid line) and predicted values (dashed line).

Table 4
Results for the Yahoo Inc Stock Prices series.

	MLP		TAEF		EMRLF	
	In-phase	Out-of-phase	In-phase	Out-of-phase	In-phase	Out-of-phase
MSE	4.2254e−4	4.1062e−4	5.6927e−4	1.4742e−4	4.1062e−4	5.5631e−7
MAPE	0.1703	0.1690e−2	0.1836	0.1097	0.1404	4.3174e−3
NMSE	1.2157	1.1809	1.3968	0.3596	1.0109	1.3646e−3
ARV	1.2800e−2	1.2740e−2	1.7662e−2	4.5684e−3	1.2740e−2	1.7239e−5
POCID	41.76	41.77	41.57	97.02	41.76	97.02
Fitness	17.4056	19.0232	15.9969	65.8202	19.2936	96.4701

6. Conclusion

An evolutionary morphological-rank-linear approach was presented in order to overcome the random walk dilemma for financial time series forecasting. The experimental results used five different metrics for model evaluation, Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Normalized Mean Square Error (NMSE), Prediction Of Change In Direction (POCID) and Average Relative Variance (ARV), demonstrating a consistent much better performance of the proposed model when compared to the MLP model and TAEF model [9] for four real world time series from the financial market with all their dependence on exogenous and uncontrollable variables (Directv Group Inc Stock Prices, Microsoft Corporation Stock Prices, Petrobras Company Stock Prices and Yahoo Inc Stock Prices).

This five different metrics were used to build a multi-objective empirical fitness function in order to improve the description of the time series phenomenon as better as possible. The five different evaluations measures used to compose this fitness function can have different contributions to final prediction result, where a more sophisticated analysis will must be done to determine the optimal combination of such metrics.

It was also observed that the proposed model obtained a much better performance than a random walk model [23] for the financial time series analyzed, overcoming the random walk dilemma. The EMRLF model was able to correct the one-step delay distortion using the phase fix procedure [9], while MLP networks alone were not capable of performing the correction although exactly the same procedure was applied to all the models. A feasible explanation for such phenomenon is that the phase fix procedure will depend on the complexity of the predictive model and on its ability to accurately define the best parameters to represent the time series.

Also, one of the main advantages of the EMRLF model (apart from its predictive performance when compared to all analyzed models) is that not only they have linear and nonlinear components, but they are quite attractive due to their simpler computational complexity when compared to other approaches such as [1,2], other MLP-GA models [9] and other statistical models [32,26,31,33].

Finally, the results showed that the phase fix procedure was able to correct more efficiently the prediction phase of the EMRLF model when compared to TAEF model [9]. Further studies are being developed to better formalize and explain the properties of the EMRLF model and to determine possible limitations of the method with other financial time series with components such as trends, seasonalities, impulses, steps and other nonlinearities. Also, further studies, in terms of risk and financial return, are being developed in order to determine the additional economical benefits, for an investor, with the use of the EMRLF method.

References

[1] R.A. Araújo, F. Madeiro, R.P. Sousa, L.F.C. Pessoa, T.A.E. Ferreira, An evolutionary morphological approach for financial time series forecasting. In: Proceedings of the IEEE Congress on Evolutionary Computation, 2006, Vancouver, Canada.

[2] R.A. Araújo, R.P. Sousa, T.A.E. Ferreira, An intelligent hybrid approach for designing increasing translation invariant morphological operators for time series forecasting, Part II of Lecture Notes in Computer Science, vol. 4492, Springer-Verlag, 2007, pp. 602–611. ISNN (2).

[3] R.A. Araújo, G.C. Vasconcelos, T.A.E. Ferreira, 2007b. Hybrid differential evolutionary system for financial time series forecasting. In: Proceedings of the IEEE Congress on Evolutionary Computation, Singapore.

[20] B.G. Malkiel, A Random Walk Down Wall Street, Completely Revised and Updated ed., W.W. Norton & Company, 2003. April.
[21] P. Maragos, A representation theory for morphological image and signal processing, IEEE Transaction on Pattern Analysis and Machine Intelligence 11 (1989) 586–599.
[22] M. Matilla-Garcfa, C. Argnello, A hybrid approach based on neural networks and genetic algorithms to the study of profitability in the spanish stock market, Applied Economics Letters 12 (5) (2005) 303–308. April.
[23] T.C. Mills, The Econometric Modelling of Financial Time Series, Cambridge University Press, Cambridge, 2003.
[24] M. Mitchell, A Introduction to Genetic Algorithms, MIT Press, Canbridge, 1999.
[25] T.C. Myhre, Financial forecasting at Martin Marietta Energy Systems Inc., The Journal of Business Forecasting Methods and Systems 11 (1) (1992) 28–30. April.
[26] T. Ozaki, Nonlinear Time Series Models and Dynamical Systems, HandBook of Statistics, vol. 5, North-Holland, Amsterdam, 1985.
[27] L.F.C. Pessoa, P. Maragos, MRL-filters: A general class of nonlinear systems and their optimal design for image processing, IEEE Transactions on Image Processing 7 (1998) 966–978.
[28] H. Pi, C. Peterson, Finding the embedding dimension and variable dependences in time series, Neural Computation 6 (1994) 509–520.
[29] L. Prechelt. Proben1: A set of neural network benchmark problems and benchmarking rules. Tech. Rep. 21/94, 1994. URL http://www.citeseer.ist.psu.edu/prechelt94proben.html.
[30] A. Preminger, R. Franck, Forecasting exchange rates: A robust regression approach, International Journal of Forecasting 23 (1) (2007) 71–84. January–March.
[31] M.B. Priestley, Non-Linear and Non-stationary Time Series Analysis, Academic Press, 1988.
[32] T.S. Rao, M.M. Gabr, Introduction to Bispectral Analysis and Bilinear Time Series Models, Lecture Notes in Statistics, vol. 24, Springer, Berlin, 1984.
[33] D.E. Rumelhart, J.L. McCleland, Parallel Distributed Processing Explorations in the Microstructure of Cognition, vols. 1 and 2, MIT Press, 1987.
[34] P. Salembier, Adaptive rank order based filters, Signal Process 27 (1) (1992) 1–25.
[35] P. Salembier, Structuring element adaptation for morphological filters, Journal for Visual Communication and Image Representation 3 (2) (1992) 115–136. June.
[36] R. Savit, M. Green, Time series and dependent variables, Physica D 50 (1991) 95–116.
[37] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, London, 1982.
[38] R. Sitte, J. Sitte, Neural networks approach to the random walk dilemma of financial time series, Applied Intelligence 16 (3) (2002) 163–171. May.
[39] F. Takens, Detecting strange attractor in turbulence, in: A. Dold, B. Eckmann (Eds.), Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol. 898, Springer-Verlag, New York, 1980, pp. 366–381.
[40] N. Tanaka, H. Okamoto, M. Naito, Estimating the active dimension of the dynamics in a time series based on a information criterion, Physica D 158 (2001) 19–31.
[41] G.P. Zhang, D. Kline, Quarterly time-series forecasting with neural networks, IEEE Transactions on Neural Networks 18 (6) (2007) 1800–1814. November.
[42] G. Zhang, B.E. Patuwo, M.Y. Hu, Forecasting with artificial neural networks: The state of the art, International Journal of Forecasting 14 (1998) 35–62.

Please cite this article in press as: R. d. A. Araújo, T.A.E. Ferreira, A Morphological-Rank-Linear evolutionary method for stock market prediction, Inform. Sci. (2009), doi:10.1016/j.ins.2009.07.007