Anais do IX Congresso Brasileiro de Redes Neurais / Inteligência Computacional (IX CBRN)
Ouro Preto 25–28 de Outubro de 2009

A New Evolutionary Morphological-Rank-Linear Approach for Time Series Prediction

Ricardo de A. Araújo[1]
[1]Information Technology Department, [gm]^2 Intelligent Systems, Campinas, SP, Brazil
ricardo@gm2.com.br

Abstract – This work presents a new evolutionary morphological-rank-linear approach in order to overcome the random walk dilemma for time series prediction. The proposed approach, referred to as 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. 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 two 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.

Keywords: Morphological-Rank-Linear Filters, Genetic Algorithms, Intelligent Hybrid Models, Financial Time Series, Stock Market Prediction.

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 [1]. Approaches based on Artificial Neural Networks (ANNs) have been successfully proposed for nonlinear modeling of time series in the last two decades [2]. In this context, hybrid intelligent approaches have produced interesting results [3].

However, a dilemma arises from all these models regarding financial time series, known as random walk dilemma, 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 [4].

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 [5] and consists of an intelligent hybrid model composed of a Morphological-Rank-Linear (MRL) [6] with a Modified Genetic Algorithm (MGA) [2], 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. 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 [3] in the attempt to adjust time phase distortions observed in financial time series.

An experimental analysis is conducted with the proposed method using two real world stock market time series, 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 [3].

2 Fundamentals

2.1 The Time Series Prediction Problem

A time series is an observation sequence of a given variable in a time period. This variable is observed in discrete or continuous time points, usually time equidistant. The analysis of this temporal behavior evolves the process or phenomenon

description that generates such observation sequence. A time series can be defined as,

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

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

The aim of predictive techniques applied to a time series X_t is to provide a mechanism that allows, with a certain accuracy, the prediction of the future values of X_t, given by $X_{t+k},\ k = 1, 2, ...$, where k represents the prediction horizon of k step ahead. Nevertheless, in order to provide proper prediction performance, the most relevant factor to guarantee prediction accuracy is the correct choice of the time lags for representing a given time series [3].

2.2 Morphological-Rank-Linear (MRL) Filter

The MRL filter [6] is a linear combination between a Morphological-Rank (MR) filter [7] and a linear Finite Impulse Response (FIR) filter [6].

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 $\underline{x} \rightarrow y$ is given by [6]

$$y = \lambda\alpha + (1 - \lambda)\beta, \tag{2}$$

with

$$\alpha = \mathcal{R}_r(\underline{x} + \underline{a}) = \mathcal{R}_r(x_1 + a_1, x_2 + a_2, \ldots, x_n + a_n), \tag{3}$$

and

$$\beta = \ \underline{x} \cdot \underline{b}' = x_1 b_1 + x_2 b_2 + \ldots + x_n b_n, \tag{4}$$

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 [6].

2.3 Morphological-Rank-Linear (MRL) Filter Adaptive Design

Pessoa and Maragos [6] 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 equations (2-4), they argued that it is possible to work with a real variable ρ implicitly defined through the following rescaling [6]

$$r = round\left(n - \frac{n-1}{exp(-\rho)}\right), \tag{5}$$

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 [6]

$$\underline{w} \equiv (\underline{a}, \underline{b}, \rho, \lambda). \tag{6}$$

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

$$\underline{w}(i+1) = \underline{w}(i) - \mu_0 \nabla J(\underline{w}), \tag{7}$$

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 Equation 7 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 [6].

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

$$J[\underline{w}(i)] = \frac{1}{M} \sum_{k=i-M+1}^{i} e^2(k), \tag{8}$$

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

$$e(k) = d(k) - y(k), \tag{9}$$

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.

Hence, the resulting adaptation algorithm is given by [6]

$$\underline{w}(i+1) = \underline{w}(i) + \frac{\mu}{M} \sum_{k=i-M+1}^{i} e^2(k) \frac{\partial y(k)}{\partial \underline{w}}, \tag{10}$$

where $\mu = 2\mu_0$ and $i \in \{1, 2, \ldots\}$. From Equations (2), (3), (4) and (6), term $\frac{\partial y(k)}{\partial \underline{w}}$ [6] may be calculated as

$$\frac{\partial y}{\partial \underline{w}} = \left(\frac{\partial y}{\partial \underline{a}}, \frac{\partial y}{\partial \underline{b}}, \frac{\partial y}{\partial \rho}, \frac{\partial y}{\partial \lambda} \right) \tag{11}$$

with

$$\frac{\partial y}{\partial \underline{a}} = \lambda \frac{\partial \alpha}{\partial \underline{a}}, \tag{12}$$

$$\frac{\partial y}{\partial \underline{b}} = (1 - \lambda)\underline{x}, \tag{13}$$

$$\frac{\partial y}{\partial \rho} = \lambda \frac{\partial \alpha}{\partial \rho}, \tag{14}$$

$$\frac{\partial y}{\partial \lambda} = (\alpha - \beta), \tag{15}$$

where

$$\frac{\partial \alpha}{\partial \underline{a}} = \underline{c} = \frac{Q\left((\alpha \cdot \underline{1}) \ - \underline{x} - \underline{a}\right)}{Q\left((\alpha \cdot \underline{1}) \ - \underline{x} - \underline{a}\right) \cdot \underline{1}'}, \tag{16}$$

$$\frac{\partial \alpha}{\partial \rho} = 1 - \frac{1}{n} Q\left((\alpha \cdot \underline{1}) \ - \underline{x} - \underline{a}\right) \cdot \underline{1}', \tag{17}$$

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

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) [2] 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 Ferreira [3] and other linear and nonlinear statistical models [1].

The EMRLF method is based on the definition of the two main elements necessary for building an accurate forecasting system according to Ferreira [3]: (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 [2] to adjust the MRL filter and the LMS algorithm [6] to train it. The purpose of using the MGA [2] 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 [6] 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 idea used here is to conjugate a local search method (LMS) to a global search method (MGA). While the MGA makes possible the testing of varied solutions in different areas of the solution space, the LMS acts on the initial solution to produce a fine-tuned model.

The algorithm starts with the definition of the MGA and MRL filter pre-determined parameters. Initially, a population with I individuals is generated. Each individual represents a MRL filter, where the input of the MRL filter ($\underline{x}$), as defined by selected time lags, represents the time series and the output of the MRL filter (y) represents the prediction horizon (in this case of one step ahead). At each MGA generation, all individuals of the MGA population are trained by the LMS algorithm for a period of E epochs.

In order to provide a more robust forecasting model, a multi-objective fitness function is used, resulting from a combination of five well-known performance measures, which is given by:

$$\text{Fitness} = \frac{\text{POCID}}{1 + \text{MSE} + \text{MAPE} + \text{NMSE} + \text{ARV}}. \tag{18}$$

After model training (the end of EMRLF method's iterations), the proposed method uses the phase fix procedure introduced by Ferreira et al. [3] in the TAEF method, to adjust time phase distortions observed ("out-of-phase" matching) in financial time series. Ferreira et al. [3] 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: (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 [8]: i) The Maximum number of epochs, ii) The increase in the validation data error or generalization loss (Gl) beyond 5%, and iii) The decrease in the training data error (Pt) below 10^{-6}.

Each individual of the MGA population is an MRL filter. The individuals are represented by chromosomes that have the following genes (MRL filter parameters): i) $\underline{a}$: MR filter coefficients, ii) $\underline{b}$: linear FIR filter coefficients, iii) ρ: variable used to determine the rank r, iv) λ: mixing parameter, and v) $\underline{lag}$: 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 ($lag_i \geq 0$) or not ($lag_i < 0$).

4 Simulations and Experimental Results

A set of two financial time series was used as a test bed for evaluation of the EMRLF method: 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 [8]: 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 [3], 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 [3] 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 [9] 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.

4.1 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 $\rho = 0.5321$ and $\lambda = 0.0011$ 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.

Table 1: 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

Figure 1(a) 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.

4.2 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.

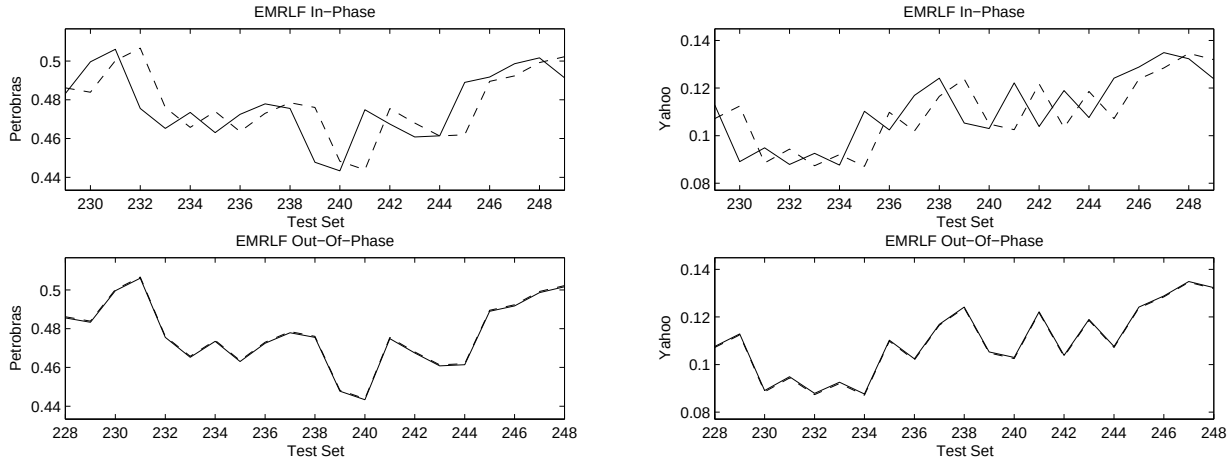
For the prediction of the Yahoo Inc Stock Prices series (with one step ahead of prediction horizon), the proposed method automatically chose the lags 2 and 9 as the relevant lags for the series representation, defined the parameters $\rho = 0.5360$ and $\lambda = 0.0019$ and classified the model as "out-of-phase"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

Table 2: 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

Figure 1(b) 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 [3], 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.



(a) (b)

Figure 1: Prediction results for the analyzed financial time series (test set): actual values (solid line) and predicted values (dashed line).

5 Conclusion

A new 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 [3] for two real world time series from the financial market with all their dependence on exogenous and uncontrollable variables (Petrobras Company Stock Prices and Yahoo Inc Stock Prices).

This five different metrics were used into a multi-objective empirical fitness function in order to improve the description of the time series phenomenon as better as possible. It was also observed that the proposed model obtained a much better performance than a random walk model [10] 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 [3], 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 TAEF model [3] and other linear and nonlinear statistical models [1].

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 [3]. 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 non-linearities. 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] G. E. P. Box, G. M. Jenkins and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, New Jersey, third edition, 1994.

[2] F. H. F. Leung, H. K. Lam, S. H. Ling and P. K. S. Tam. "Tuning of the Structure and Parameters of the Neural Network Using an Improved Genetic Algorithm". *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 79–88, January 2003.

[3] T. A. E. Ferreira, G. C. Vasconcelos and P. J. L. Adeodato. "A New Intelligent System Methodology for Time Series Forecasting with Artificial Neural Networks". In *Neural Processing Letters*, volume 28, pp. 113–129, 2008.

[4] B. G. Malkiel. *A Random Walk Down Wall Street, Completely Revised and Updated Edition*. W. W. Norton & Company, April 2003.

[5] F. Takens. "Detecting Strange Attractor in Turbulence". In *Dynamical Systems and Turbulence*, edited by A. Dold and B. Eckmann, volume 898 of *Lecture Notes in Mathematics*, pp. 366–381, New York, 1980. Springer-Verlag.

[6] L. F. C. Pessoa and P. Maragos. "MRL-Filters: A General Class of Nonlinear Systems and Their Optimal Design for Image Processing". *IEEE Transactions on Image Processing*, vol. 7, pp. 966–978, 1998.

[7] P. Salembier. "Adaptive rank order based filters". *Signal Process.*, vol. 27, no. 1, pp. 1–25, 1992.

[8] L. Prechelt. "Proben1: A set of Neural Network Benchmark Problems and Benchmarking Rules". Technical Report 21/94, 1994.

[9] M. Hagan and M. Menhaj. "Training feedforward networks with the Marquardt algorithm". *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, November 1994.

[10] T. C. Mills. *The Econometric Modelling of Financial Time Series*. Cambridge University Press, Cambridge, 2003.