Artificial Intelligence Models for Financial Time Series

Alina Barbulescu
Transilvania University of Braşov, Romania
alina.barbulescu@unitbv.ro
Cristian Stefan Dumitriu
SC Utilnavorep SA, Constanța, Romania
cris.dum.stef@gmail.com

Abstract

Modeling and predicting the evolution of financial series has become an essential research domain for scientists and practitioners in the field of economics or finance. In this context, the purpose of this article is to determine two artificial intelligence alternative models for NYSE monthly series recorded for 53 years and to compare their performances.

Key words: Time series, GEP, SVR, MSE, MAE, MAPE.
J.E.L. classification: C32, C58

1. Introduction

Modeling the financial market is closely linked to the assumption of Adaptive Market Hypotheses (Lo, 2004) that represent a harmonization of economic theories constructed on the Efficient Market Hypothesis with behavioral economics. Given the highly nonlinear dynamics of the stock market and the stochastic behavior of some elements, modeling such series by classical methods is inappropriate. Therefore, alternative methods have been proposed for solving such problems, most of them being recently based on artificial intelligence methods.

In this context, this article proposes a Gene Expression Modeling and a Support Vector Regression models for the New York Stock Exchange (NYSE) weekly close data series recorded during the period 27/12/1965-4/12/2018 (2764 values) and to compare the results.

2. Theoretical background

Using techniques of artificial intelligence, researchers conducted forecasting studies in various financial areas, such as money exchange markets (Álvarez-Díaz, 2010; Sermpinis *et al.*, 2012; Vasilakis *et al,* 2013), macro-economic time series (Thakur *et al*, 2008; Postolache and Ariton, 2013), financial time series (Bărbulescu, 2018; Bărbulescu and Băutu 2012; Karatahansopoulos *et al*, 2014; Simian *et al*, 2020; Dragomir, 2017; Tache, 2009, Tache et al., 2010), forecasting of mutual funds (Tsai *et al*, 2011; Chen *et al*, 2014). Most studies show that the objectives of time series analysis are the interpretation of past series fluctuations and the determination of a behavioral pattern, followed by the prediction of future behavior based on the found pattern (Barbulescu and Postolache, 2021; Martinez Alvares, 2010). Many scientists investigated the variations in structure and behavior of the complex adaptive systems, one typical example being the stock markets. The evolution of the stock indices is a clear and complete image of the global equity market and the real economy. The stock indices that are subject to modeling present similarities: are price-weighted indexes and their components are periodically revised.

John Holland's (1992) theory demonstrated how the evolutionary process can be assigned to artificial systems. It was shown that any adaptation issue can be defined in genetic terms and can often be explained by what the scientific literature calls the "genetic algorithm". These algorithms are search methods built on the natural selection mechanics on which the fittest individual survives.

Koza defined the Genetic Programming paradigm as a generalization of Holland's Genetic Algorithms (Langdon and Poli, 2002). Some advantages of genetic algorithms are presented by (Arifovic and Gençay, 2000).

The applications of genetic algorithms to economic modeling consist primarily of an investigation of the systems' behavior by computer simulations. Several alternatives of Genetic Programming have been subsequently developed and the different models used to encode the solutions prompt the distinctions among them. In this article, the Gene Expression Programming algorithm (GEP) (Ferreira, 2001) is utilized.

3. Research methodology

3.1. GEP

GEP is an automatic programming algorithm that relies on the natural selection principle. The idea is to represent the solutions of the study problem as individuals whose evolution is assured utilizing genetic operators. The GEP individuals are composed of genes with the same length, which code expressions that are generally nonlinear. A gene has two parts - a head and a tail. The number of genes should be set by the user. The methods used in GEP for selecting the individuals are the proportionate roulette-wheel scheme and simple elitism. A measure of performance is utilized for evaluating the individuals in each generation, the result being a number of fitness values equal to the number of individuals. The individuals with the best fitness are selected and participate in the replication process. The mutation, transposition, and crossover are operations utilized for performing the evolution.

The goal of time series modeling by GEP is finding a model that approximates well the recorded values, $\{x_1, \ldots, x_n\}$. This can be done in the following stages:

- Define the fitness function;
- Select the sets of terminals and functions to create chromosomes. The most used set of functions is formed by addition, multiplication, subtraction, and division;
- Select the structure of the chromosome and the number of genes;
- Select the linking function;
- Select the operators that participate in the algorithm and the corresponding rates.

Choosing the window size, m, is essential in GEP in the estimation of x_t which is given by:

$$x_t = f(x_{t-m}, x_{t-m+1}, \ldots, x_{t-1}) + \varepsilon_t,\ m+1 \le t \le n. \quad (1)$$

The fitness we are working with is defined by:

$$MSE = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2 \quad (2)$$

where x_t and $\hat{x}_t$ are the given and computed values, respectively, n is the sample volume.

The experiments performed used the DTREG software, setting the maximum number of genes in a chromosome to 6, the gene head size = 5 symbols, the population size = 1000, the maximum number of generations = 200. The default values of the GEP operator rates have been employed. The set of functions utilizes were the four basic arithmetic operations, together with sine and cosine. the selection scheme, the fitness proportionate selection, enhanced with elitist survival of the best 10% of the individuals in each generation onto the next. We performed 50 runs for each window size in the interval [1,12] and we report the best models found.

2.2. Support Vector Regression (SVR)

SVR is a supervised learning technique based on the errors' minimization principle (Sapankevych and Sankar, 2009). SVR uses a set of training data instances (x_i, y_i) for building a function f, utilized for estimating the y_i-s where only x_j –s are known. The model is firstly trained on a dataset, then evaluated on another set, completely different from the first one, called a test set. The model's accuracy is assessed by different indicators, presented in tables in the next section.

ε – SVR (Vapnik, 1995) utilizes an ε-insensitive loss function to minimize the generalized error (Basak *et al*, 2007; Smola and Scholkopf, 2004). The imposed constraints transform the problem at hand into a convex optimization one.

An equivalent form of the problem to be solved is: find the minimum of the function

$$\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*)\right\} \tag{3}$$

under the constraints:

$$\begin{cases} y_i - \langle w, x_i \rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \le \varepsilon + \xi_i^*. \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{4}$$

Utilizing the dual problem, the objective function *f* becomes:

$$f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\langle x_i, x \rangle + b, \tag{5}$$

with

$$\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0,\ \alpha_i, \alpha_i^* \in [0, C]. \tag{6}$$

The optimization constant C and the parameter ε have to be estimated.

For finding the solution to a non-linear problem, the input data is projected on a Hilbert space. This operation transforms the problem to be solved in a new problem involving the use of a kernel function (Smola and Scholkopf, 2004; Specht, 1992). For performing the forecast, the choice of kernel parameters should be provided. This selection is many times realized by hand, and then the parameters are adjusted based on the experimental results.

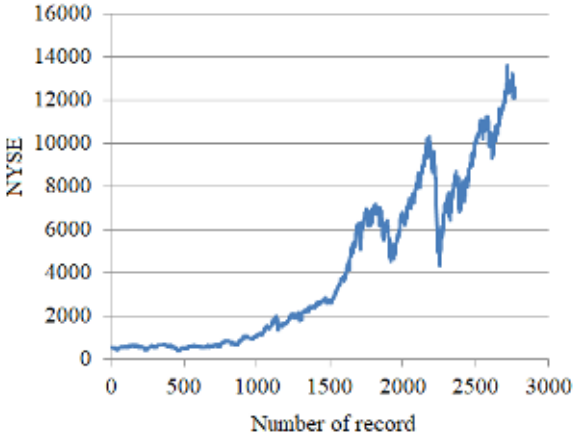
Different software can be used to perform the modeling, among which R and DTREG.

The algorithms have been run for the data obtained by taking logarithms, and the predictor variables form lag 1 to lag 5, using DTREG. The best result is reported. The search criterion was the minimization of the total error and the kernel used was RBF.

3.2. Data series

Data used for this modeling are weekly close data of the New York Stock Exchange (NYSE) recorded during the period 27/12/1965-4/12/2018. Data series is presented in Figure 1.

Figure no.1. Data series



Source: Chart built by the authors using the data from https://www.nyse.com/market-data/historical

For the study purpose, the logarithm has been applied to the series values. The new series is called lnNYSE in the following.

4. Findings

The series has been divided into two parts, in a ratio 95:5, the first one for training and the second one for test.

Running GEP algorithm with different lag variables as regressors, the best results, in terms of errors have been obtained when using GEP with lag 1 regressor. The generated expression is

$$\text{lnNYSE} = \text{lnNYSE\_Lag1}+(0.0003905/\text{lnNYSE\_Lag1})+0.0010894. \quad (7)$$

The goodness of fit indicators are presented Table 1 for both, training and validation datasets. On both sets, the MSE, MAE, and MAPE have low values, while the correlation actual - predicted values has a value close to unity, showing high performances of the algorithm.

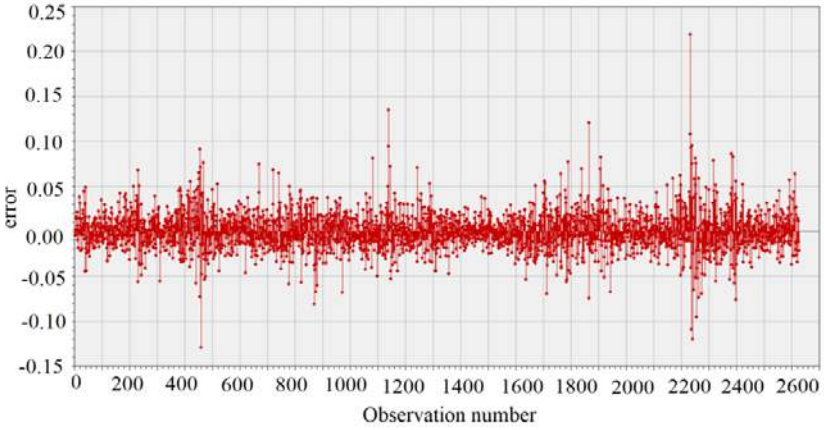
Figure 2 presents the errors in the GEP model. All but six of them are between -0.10 and 0.10, with a very low variance (row 1, in Table 1)

Table no. 1 Goodness of fit indicators in the GEP algorithm

Indicator	Training	Test
Residual (unexplained) variance after model fit	0.0013651	0.0042944
Coefficient of variation (CV)	0.0048230	0.0069890
Correlation between actual and predicted	0.9994450	0.9155540
MSE (Mean Squared Error)	0.0013651	0.0042944
MAE (Mean Absolute Error)	0.0166131	0.0551799
MAPE (Mean Absolute Percentage Error)	0.2227928	0.5859089

Source: Output of the modeling using the DTREG software

Figure no. 2. Errors in the GEP model



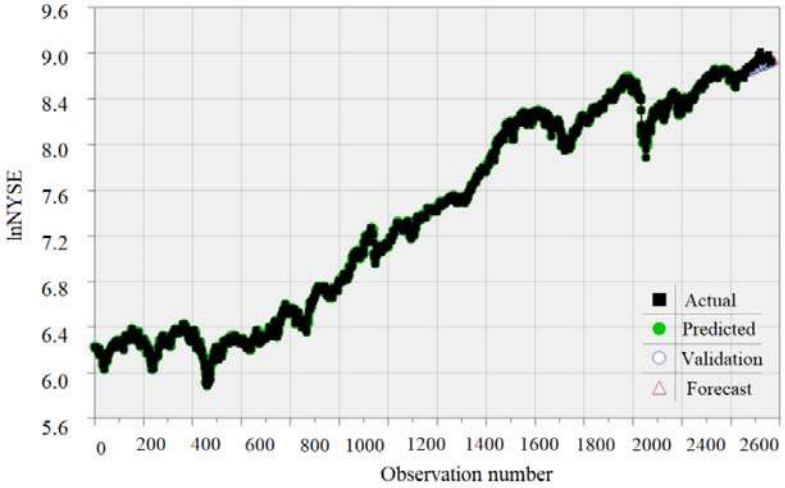
Source: Output of the modeling using the DTREG software

Based on the determined model (trained on 95% of data and tested on 5% of data), the forecast for the next 12 months after the end of the study period is also done and the values are represented in Figure 3 together with the actual values (recorded) and predicted values (estimated by the model). There is a good concordance between the recorded and predicted values. The model's performance is emphasized by the absolute values of the difference between the recorded and predicted values as well. They are lower than 2% for most data on the test set (and much lower on the training one) showing that the algorithm learns well the data and applies what it learned on the new (test) data.

The analysis or the autocorrelation and partial autocorrelation functions of the residual show the autocorrelation absence, which means that the errors will not propagate in the model.

The same study has been performed using the SVR algorithm. We report here the best result, for which the computed minimum error was 0.000487, for the parameters ε = 0.001, C = 2928.66286, when using the five regressors (lag1 - lag5 variables). Table 2 contains the goodness of fit indicators. On the training set, the performances of GEP and SVR are comparable. On the test set, in terms of all indicators from Table no. 1, but the correlation actual - predicted values, GEP is better. MAPE is much higher for SVR, showing a higher ratio of error over the actual values. Since MAPE is not dimensional, it is a better indicator for comparing the modeling results when many algorithms are utilized. Base on it, GEP is recommended for modeling the study series.

Figure no. 3. Graphical representation of actual values (black squares), predicted values (green- there are the values estimated by the algorithm), and the forecast (red triangles) for the next 12 weeks. The blue dots (called "Validation") represent the estimated values on the test set



Source: Output of the modeling using the DTREG software

Table no. 2 Goodness of fit indicators in the SVR algorithm with 5 regressors

Indicator	**Training**	**Test**
Residual (unexplained) variance after model fit	0.0015121	0.0181238
Coefficient of variation (CV)	0.0050760	0.0143580
Correlation between actual and predicted	0.9993870	0.9325220
MSE (Mean Squared Error)	0.0015121	0.0181238
MAE (Mean Absolute Error)	0.0168513	0.1170115
MAPE (Mean Absolute Percentage Error)	0.2270184	1.2422628

Source: Output of the modeling using the DTREG software

5. Conclusions

Two artificial intelligence approaches for modeling NYSE weekly series for 53 years have been presented in this article. The study series was long, nonstationary in trend, with high variability. This is why the logarithm on the series values has been taken before modeling. Even if the best results have been obtained by GEP, the second algorithm was a good competitor. What recommends the first approach is MAPE, which was more than twice lower for GEP compared to the value obtained by using SVR. Another more important aspect, not mentioned yet, is the run time, which was few minutes for each GEP run, compared with few hours for SVR. As a future work direction, hybrid methods should be tested for increasing the obtained results quality.

6. References

- Álvarez-Díaz, M., 2010.Speculative Strategies in the Foreign Exchange Market Based on Genetic Programming Predictions. *Applied Financial Economics,* 20(6), pp. 465- 76.
- Arifovic, J., Gençay, R., 2000. Statistical properties of genetic learning in a model of exchange rate. *Journal of Economic Dynamics and Control*, 24(5–7), pp. 981-1005.
- Basak, D., Pal, S., Patranabis, D.C. 2007. Support Vector Regression. *Neural Information Processing – Letters and Reviews*, 11(10), pp. 203 - 224.
- Bărbulescu, A., 2018. Do the time series statistical properties influence the goodness of fit of GRNN models? Study on financial series. *Applied Stochastic Models in Business and Industry*, 34(5), pp. 586 - 596.
- Bărbulescu, A., Băutu, E., 2012. A hybrid approach for modeling financial time series. *International Arab Journal of Information Technology*, 9(4), 327- 335.
- Bărbulescu, A., Postolache, 2021. F. New approaches for modeling the regional pollution in Europe. *Science of the Total Environment*, 753, 141993.
- Chen, W.-H., Shih, J.-Y., Wu, S., 2006. Comparison of support vector machines and backpropagation neural networks in forecasting the six major Asian stock markets. *International Journal of Electronic Finance*, 1(1), pp.49-67.
- Ferreira, C., 2001. Gene Expression Programming: A new adaptive Algorithm for Solving Adaptive. *Complex Systems*, 13(2), pp. 87-129.
- Karatahansopoulos, A., Sermpinis, G., Laws, J., Dunis, C., 2014. Modelling and Trading the Greek Stock Market with Gene Expression and Genetic Programming Algorithms. *Journal of Forecasting*, 33(8), 596-610.
- Koza J. R., 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Massachusetts: MIT Press Cambridge.
- Langdon, W. B., Poli, R., 2002. *Foundations of Genetic Programming*. Berlin, Heidelberg: Springer-Verlag.
- Lo. A. W., 2004. The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. *Journal of Portfolio Management*, 30, pp. 15-29.
- Martinez Álvares, F., 2010. Pattern sequence analysis to forecast time series, PhD diss. Universidad de Sevilla – Espana.
- Postolache, F., Ariton, V., 2013. *Economical informatics*. Galați: Ed. Danubius University.
- Sapankevych, N., Sankar, R., 2009. Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine*, 4(2), pp. 24 - 38.
- Sermpinis, G., Laws, J., Karathanasopoulos, A., Dunis, C. L., 2012. Forecasting and trading the EUR/USD exchange rate with Gene Expression and Psi Sigma Neural Networks. *Expert Systems with Applications*, 39(10), pp. 8865-77.
- Simian, D., Stoica, F., Bărbulescu, A. 2020. Automatic Optimized Support Vector Regression for Financial Data Prediction. *Neural Computing & Applications*, 2019, 32, pp. 2383-96.
- Smola, A. J., Scholkopf, B. A tutorial on support vector regression. *Statistics and Computing*, 14(3), 2004, pp. 199 - 222.
- Specht, D. F., 1991. A General Regression Neural Network. *IEEE Transactions on Neural Networks*, 2(6), pp. 568 – 76.
- Tache F. L., 2009. Advice in electronic commerce. Proceedings of 3rd IEEE Internaţional Workshop on Soft Computing Applications, Szeged (Hungary)- Arad (România), July 29- Aug. 1, 2009, pp. 111-14.
- Tache F.L. et al. 2010. Consulting in electronic commerce. *Acta Univ. Danubius, Economica,* 6(3), pp.161-69
- Thakur, G. S. M., Bhattacharyya, R., Mondal, S. S., 2016. Artificial Neural Network Based Model for Forecasting of Inflation in India. *Fuzzy Information and Engineering*, 8(1), pp. 87-100.
- Tsai, T.-J., Yang, C.-B., Peng, Y.-H., 2011. Genetic algorithms for the investment of the mutual fund with global trend indicator. *Expert Systems with Application*, 38, pp. 1697–701.
- Vapnik, V., 1995. T*he Nature of Statistical Learning Theory*. New York: Springer Verlag.
- Vasilakis, G. A., Theofilatos, K. A. Efstratios, Georgopoulos, F., Karathanasopoulos, A., Likothanassis, S. D., 2013. A Genetic Programming Approach for EUR/USD Exchange Rate Forecasting and Trading. *Computational Economics*, 42(4), pp. 415-31.