ISSN 1870-4069

A Novel Machine Learning Method Applied to the Forecast of a Stock Market Index

Enrique González-Nuñez, Luis A. Trejo

Tecnologico de Monterrey,
Escuela de Ingeniería y Ciencias,
Mexico

A00457801@itesm.mx, ltrejo@tec.mx

Abstract. The present paper recalls the aim of the ongoing work of applying the Artificial Organic Networks (AON) machine learning framework, to develop a new algorithm capable of generating a prediction for a stock market index, based on the Index Tracking Problem (ITP); thus, a conception of a new AON arrangement is needed. Pursuing this main goal, a first approach toward the definition of a new topology is presented, stating some general ideas along with the following discussion. Finally, we offered some preliminary results related to these main notions, considering the employment of a multiple non-linear regressive (MNLR) model, to build an AON structure; the relative error obtained along the experiments was of the order of $1x10^{-2}$.

Keywords: Machine learning, nature-inspired, metaheuristic, stock market index, forecast.

1 Introduction

The stock market prediction or Index Tracking Problem (ITP), is a complex process affected by dynamic and non-linear factors across time, and yet an important research topic in financial area with big challenges [2–4, 6–8, 21–23, 25]. The ITP is a trading strategy based on the buy-and-hold of assets, that uses an index tracker to reproduce the performance of a stock market index or any other security found in the capital markets; the behavior of the performance is reproduced by developing models capable of yielding a forecast.

1.1 Related Work

Previously, in [5] has been reported the aim of applying the Artificial Organic Networks (AON) metaheuristic machine learning framework, to develop a new efficient algorithm, able of generating a short-term market trend forecast; the complexity of the problem was narrowed down by two main constraints:

1 The forecast would be done using the historic prices of the concerning index rate and at least two additional macroeconomic variables (MEVs).
2 The MEVs would be selected based on their correlated coefficient (CC) to the index being analyzed.

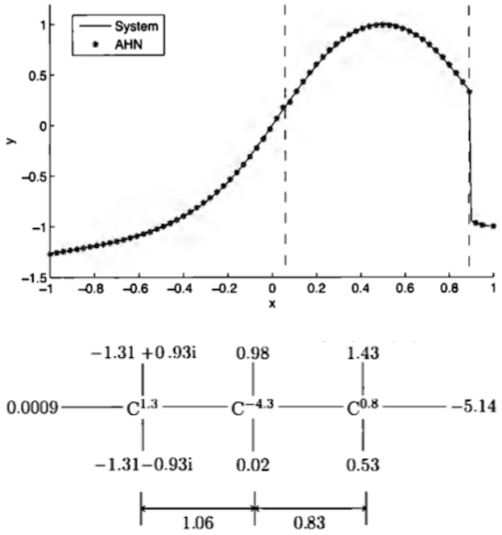

Fig. 1. Example [19, 20] of the AHN process as a learning method, where an AON structure is built through the algorithm identified as f. Along the process, a structure of an organic compound is produced by segmenting a dataset (target function f) and fitting a second-degree or third-degree polynomic term for each section.

2 Methodology

Following the notions and main characteristics of AON as a machine learning class [19–20], a new algorithm is to be defined. Within its characteristics, AON requires to use of a topological configuration for its implementation; one main limitation of this method is that it only has been implemented through one existing topology, Artificial Hydrocarbon Networks (AHN), which have shown improvements in predictive power and interpretability in contrast with other well-known machine learning models, comparatively to neural networks and random forest, but has the disadvantage of being very time-consuming and is not able to deal big data since the model uses stochastic gradient descent (SGD).

AHN as a chemically bio-inspired algorithm performs an optimization of a cost-energy function in two levels to build organic compound structures:

1 It uses least-squares regression (LSR) to define the structure of molecules.
2 It uses SGD to optimize the position of the molecules in the feature space.

2.1 Conception of a New Topology

Since AHN is the only existing arrangement for the AON framework, the postulation of the new algorithm adept to perform the stated objectives, is going to be based on the conception of a different topology to provide better capability to deal with big data and reduce time consumption, to avoid losing predictive power as one of the two main characteristics of the original AHN topology mentioned above.

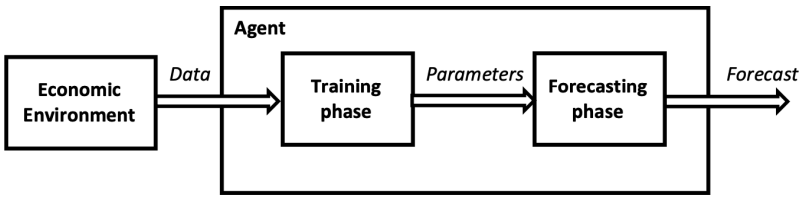

Fig. 2. Diagram of the proposed agent organized in two phases; this new topology would be capable of performing a forecast for a stock market index, using at least two additional MEVs chosen based on their CC to the index being analyzed. Along the process, a dataset will be received and segmented by the training phase, for each section a curve would be fitted using MNLR to compute the structure of an organic compound. The parameters output of the training phase are the values to build the AON structure that models a given system (function f).

Because the research presented here is still ongoing, only a general approach to a future topology is presented as a follow-up of the work introduced in [5], being aware that further experiments are being conducted to formalize the details of the new algorithm.

In the simplest notion, as explained in the literature [19–20], the algorithm that builds the structure of an AON as a learning method is identified as f, and through the AHN topology produces a structure of an organic compound by segmenting the dataset of the information received and fitting a second-degree or third-degree polynomic term for each section utilizing LSR (see Figure 1).

Consequently, the general approach of a different topology is mainly driven -up to now- by the next main ideas:

1 To define a new AON arrangement based on a functional group different from the hydrocarbons.
2 The new organic structure will fit the information of each segment by using a multiple non-linear regressive (MNLR) model [5], using three variables: the historic prices of the index, and a minimum of two additional MEVs chosen depending on their CC to the index rate.
3 As a significant feature to characterize a new AON arrangement, it will substitute the type of curve that is fitted when a polynomic term is computed for each segment of the dataset that is being modeled.

3 Exploratory Analysis & Preliminary Results

In this work, the main concepts of AON are recovered to postulate a new topology through the implementation of an agent [5] that receives external data from the economic environment; the agent will be formed in two phases (see Figure 2):

1 The training phase: will find the parameters to build the structure of an AON.
2 The forecasting phase: will estimate a prediction of a stock market, using the AON structure.

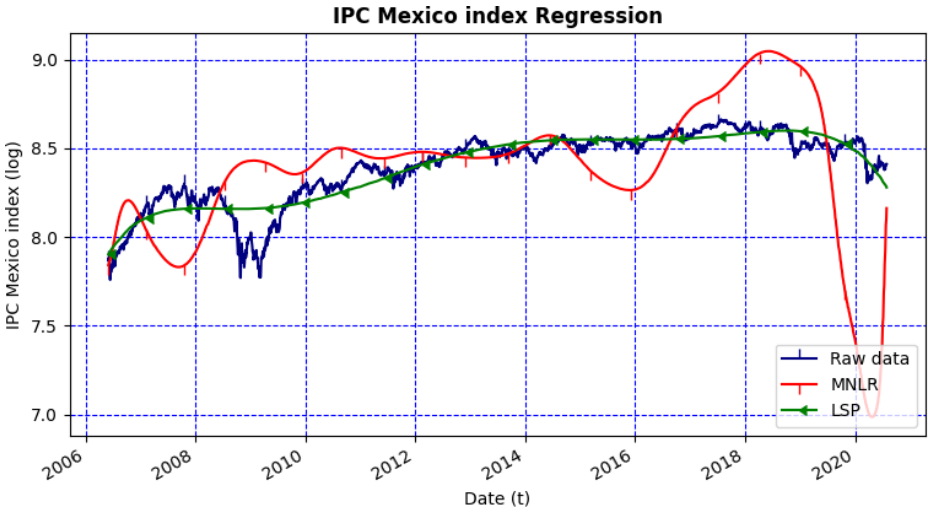

Fig. 3. Curves for the IPC Mexico using the original data (blue line), a MNLR model that replaces the second-degree term with a SIN term (red line), and the LSP model (green line).

In Python, some preliminary experiments have been carried out [5] and still are being performed, to evaluate the viability and effectiveness of conceptualizing a new AON arrangement by applying an MNLR method to build an organic compound structure, while substituting the type of curve that is fitted when a polynomic term is computed for each segment along the process. These backtest experiments have been done using the SciPy [24], and Scikit-learn [9] libraries.

The tests have been carried out using existing data from Mexico comprehending the period from 1/6/2006 to 30/7/2020, the cognition behind the size of data selected is to assure that at least one short economic cycle is used [5].

The dataset includes the following variables: the daily reported (labor days) IPC Mexico stock market index, the quarterly reported gross domestic product (GDP), the daily reported (labor days) MXN-USD foreign exchange rate (FX), the monthly reported consumer price index (CPI), the monthly risk-free rate (RFR), the monthly unemployment rate (UR), the monthly reported current account to GDP rate (BOP), and the monthly reported Investment rate (GFCF). The IPC data was retrieved from Yahoo Finance, the FX was acquired from the USA's Federal Reserve Board, and the rest of the variables were obtained from the OECD.

It is relevant to explain that -up to now- the experiments have been done randomly splitting the values of the data into subsets of 80% for training and the rest for testing, as commonly happens in machine learning, since the approach is to find good solutions by reducing the computing time, in contrast of the classic statistical model like ARIMA where the data is split based on the DateTime [1, 25].

Despite being an unusual practice for time series, good performance has been obtained previously with this approach [5]. Also, it would be noticed that the tests include an approximation using a least-squares polynomial regression (LSP) based only on the historic data of the IPC Mexico, this was for baseline purposes.

Table 1. Mean, SD, MAD, Max, Min, and Range of the relative error from the MNLR model using a SIN term.

Relative Error (MNLR)					
Mean	**SD**	**MAD**	**Max**	**Min**	**Range**
0.029072	0.031475	0.021225	0.167772	0.000013	0.167758

Table 2. RSS, SSR, TSS, and R-square of the MNLR model using a SIN term. It can be observed that SSR is larger than SSE, following one of the criteria of a good regression model.

Relative Error (MNLR)			
RSS	**SSR**	**TSS**	**R-square**
367.413282	418.723554	786.136836	0.532634

For each experimentation, a MNLR model of the IPC was computed employing the previously defined seven MEVs (FX, GDP, CPI, RFR, UR, BOP, and GFCF); the data was preprocessed in three steps:

1 The MEVs were treated as "continuous signals", so for each input, an independent approximation was done.
2 The data was standardized by removing the mean.
3 The dimensionality of the data was reduced using principal component analysis (PCA), this was done using three components.

Furthermore, is important to remark here that being an initial approach, the dataset was not -yet- segmented as the AON framework would formally do. The experiments considered most relevant will be described now in the subsequent sections.

3.1 Experiment 1: Establishing a MNLR Model Using SIN Terms

Exploring the possibility of changing the type of curve that is fitted by a molecule for each segment, when an AON structure is computed, the MNLR model for the IPC Mexico was estimated by replacing the second-degree polynomic terms with a sinusoidal (SIN) term; Figure 3 shows the graph of the obtained curve using this approach. As expected, in the graph discrepancies can be appreciated across the time (t) between the original curve of the IPC Mexico and the computed MNLR model.

Nonetheless, as stated above, the dataset was not -yet- segmented as the AON framework would formally do, considering this is a significant factor for the discrepancies. It is considered that in the future these divergences will be reduced once the data is segmented; moreover, in the graph can be observed along (t) that in some intervals (e.g., around the years 2013-2014) the approximated model behavior is very much the same as the trend of the original data.

Despite the discrepancies, the estimation provided satisfactory results based on the the relative error ε_t [5]; in this regard, Table 1 shows the mean, the standard deviation (SD), the mean absolute deviation (MAD), the maximum (Max) value, the minimum (Min) value, and the range of the obtained relative error. Figure 4 shows the relative error of the MNLR model using a SIN term.

The performance of the model was also measured by computing the Residual Sum of Squares (RSS), the Sum of Squares Regression (SSR), the Total Sum of Squares (TSS), and the coefficient of determination R-square; the results are reported in Table 2.

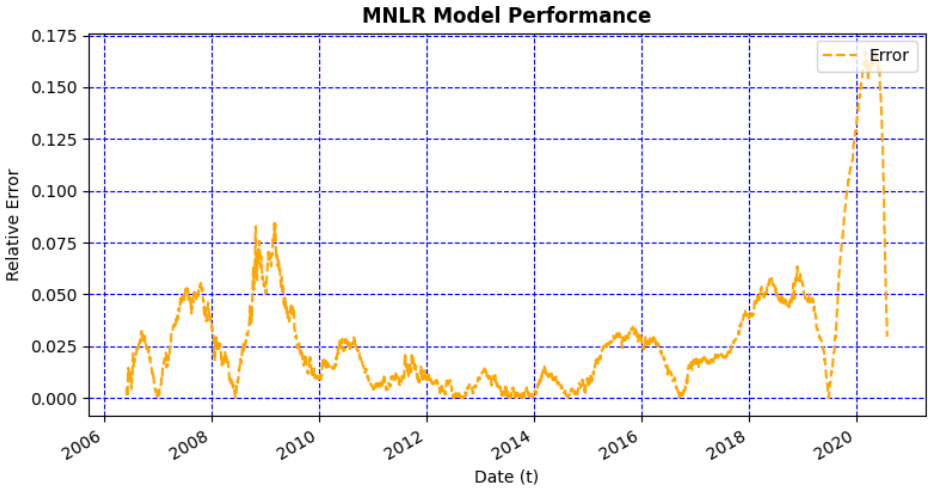

Fig. 4. Relative error for the estimated IPC using the MNLR model and replacing the second-degree term with a SIN term.

Table 3. Mean, SD, MAD, Max, Min, and Range of the relative error from the MNLR model using an EXP term.

Relative Error (MNLR)					
Mean	**SD**	**MAD**	**Max**	**Min**	**Range**
0.016755	0.014252	0.011017	0.063554	0.00002	0.063534

3.2 Experiment 2: Establishing a MNLR Model Using EXP Terms

Again, to test the possibility of changing the type of curve that is fitted by a molecule for each segment, when an AON structure is computed, in experiment two the MNLR model was assessed by replacing the second-degree polynomic terms with an exponential (EXP) term; Figure 5 shows the graph of the curve attained using this approach. Once more, in the graph are appreciated discrepancies across (t) between the original curve and the computed MNLR model.

As before, the data was not -yet- segmented as the AON framework would formally do, considering this as an important factor for the discrepancies. As well, it is considered that in the future these divergences will also be reduced once the data is segmented; likewise, in the graph can be observed along (t) that in some intervals (e.g., around the years 2012-2013, and 2020) the approximated model behavior is very much the same as the trend of the original data.

Once more, after estimating the model, the relative error was computed; however, in this case, the performance increased in comparison to the results acquired in the first experiment. Table 3 shows the mean, the SD, the MAD, the Max value, the Min Value, and the range achieved by the relative error in this case. Figure 6 shows the relative error of the MNLR model using an EXP term.

Subsequently, the performance of the model was also measured by computing the RSS, the SSR, the TSS, and the R-square; the results are reported in Table 4.

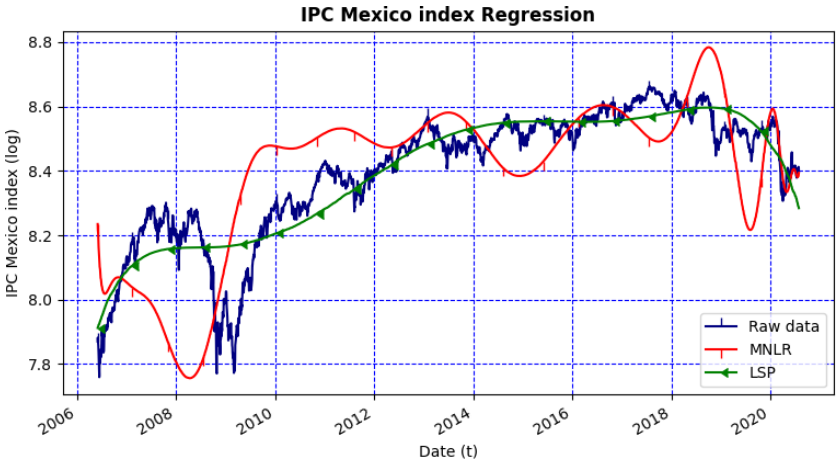

Fig. 5. Curves for the IPC Mexico using the original data (blue line), a MNLR model that replaces the second-degree term with an EXP term (red line), and the LSP model (green line).

Table 4. RSS, SSR, TSS, and R-square of the MNLR model using an EXP term. It can be observed that SSR is larger than SSE, following one of the criteria of a good regression model.

Relative Error (MNLR)			
RSS	**SSR**	**TSS**	**R-square**
93.568285	168.641547	262.209833	0.643155

4 Conclusions & Future Work

Through this paper is evoke the objective of producing a new algorithm based on the AON machine learning class, to yield a short-term stock market trend forecast, using at least two additional MEVs.

As stated, the AON compelling concepts are observed to design a new AON topology; in this respect, contemplating that the AHN topology produces a structure by segmenting the dataset, and fitting a curve to each section, the present state of the ongoing research has been focused on undertaking experiments to identify different type of math expressions for substituting the second-degree and third-degree polynomic terms employed by the AHN algorithm.

As a first approach, sinusoidal and exponential terms have been used in the MNLR model to replicate the behavior of the IPC Mexico. Through the first approximations exemplified here, disparities have been observed between the acquired results and the raw data; however, it has been remarked that the procedure applied still has not segmented the data as the AON framework do.

In addition to the last statement, the obtained results from the experiments have provided a relative error of the order of $1x10^{-2}$; regarding the last remarks, is being pondered that is plausible to define in further experiments a new AON arrangement, by using the MNLR model and substituting the type of math expression to fit different curves of the segments along (t) of the function f.

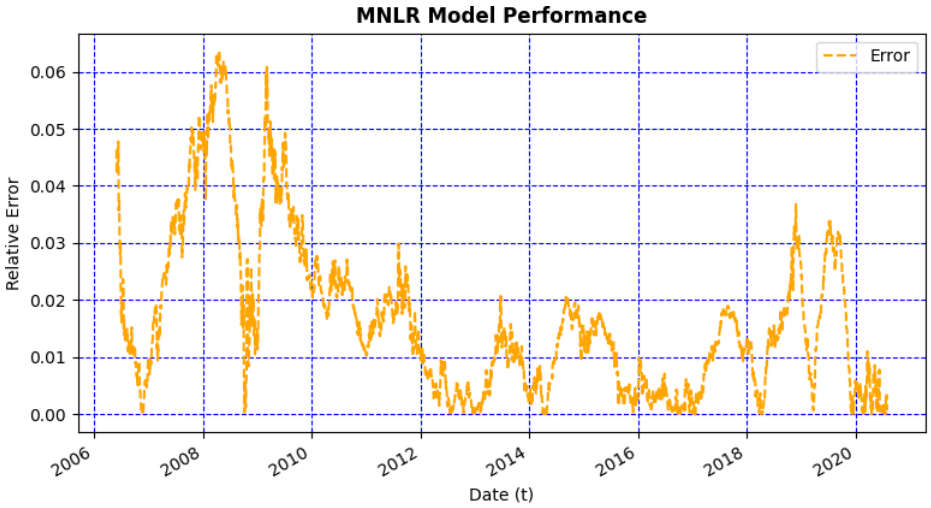

Fig. 6. Relative error for the estimated IPC using the MNLR model and replacing the second-degree term with an EXP term.

To improve the performance and reduce the discrepancies of the results obtained by now, the next tasks considered for future work include: i) segmenting the data as the AON framework formally does while producing a structure of an organic compound, ii) doing further experiments considering different kinds of math expressions that are used to characterize curves, iii) the procedure of splitting the data into train and test subsets set up on sequence (based upon the DateTime).

Aknowledgments. Supported by Tecnológico de Monterrey and CONACyT.

References

1. Aronson, L.: ARIMA Modeling and Train/Test Split. Learn (2020)
2. Chacón, H., Kesici, E., Najafirad, P.: Improving Financial Time Series Prediction Accuracy Using Ensemble Empirical Mode Decomposition and Recurrent Neural Networks. In: IEEE Access, vol. 8, pp. 117133–117145 (2020). DOI: 10.1109/ACCESS.2020.2996981.
3. Elliott, G., Granger, C., Timmermann, A.G.: Handbook of Economic Forecasting. North-Holland, vol. 1 (2013)
4. Focardi, S.M., Fabozzi, F.J.: The Mathematics of Financial Modeling and Investment Management. The Frank J Fabozzi Series, Wiley Finance (2004)
5. González, E., Trejo, L.: Artificial Organic Networks Approach Applied to the Index Tracking Problem. In: Mexican International Conference on Artificial Intelligence, pp. 23–43 (2021). DOI: 10.1007/978-3-030-89817-5 2.
6. Hou, X., Wang, K., Zhang, J., Wei, Z.: An Enriched Time-Series Forecasting Framework for Long-Short Portfolio Strategy. In: IEEE Access, vol. 8, pp. 31992–32002 (2020). DOI: 10.1109/ACCESS.2020.2973037.
7. Ordóñez, J.M.: Predicción del comportamiento de los mercados bursátiles usando redes neuronales. Technical report, Depto. Ingeniería de Sistemas y Automática, Escuela Técnica Superior de Ingeniería, Universidad de Sevilla, pp. 1–129 (2017)

8. Ortiz, F., Cabrera-Llanos, A.I., López-Herrera, F.: Pronóstico de los índices accionarios DAX y S&P 500 con redes neuronales diferenciales. Contaduría y administración, vol. 58, no. 3, pp. 203–225 (2013). DOI: 10.1016/S0186-1042(13)71227-0.
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Matthieu-Perrot, E., Duchesnay: Scikit-learn: Machine learning in Python (2017)
10. Ponce, H., Acevedo, M.: Design and Equilibrium Control of a Force-Balanced One-Leg Mechanism. In: Advances in Computational Intelligence (MICAI), pp. 276–290 (2018). DOI: 10.1007/978-3-030-04497-8_2.
11. Ponce, H., Acevedo, M., Morales-Olvera, E., Martínez-Villaseñor, L., Díaz-Ramos, G., Mayorga-Acosta, C.: Modeling and Control Balance Design for a New Bio-Inspired Four-Legged Robot. In: Advances in Soft Computing (MICAI), vol. 11835, pp. 728–739 (2019). DOI: 10.1007/978-3-030-33749-0_58.
12. Ponce, H., Campos-Souza, P.V., Junio-Guimarães, A., González-Mora, G.: Stochastic Parallel Extreme Artificial Hydrocarbon Networks: An Implementation for Fast and Robust Supervised Machine Learning in High-Dimensional Data. Engineering Applications of Artificial Intelligence, vol. 89 (2020). DOI: 10.1016/j.engappai.2019.103427.
13. Ponce, H., González-Mora, G., Morales-Olvera, E., Souza, P.: Development of Fast and Reliable Nature-Inspired Computing for Supervised Learning in High-Dimensional Data. In: Nature Inspired Computing for Data Science, vol. 871, pp. 109–138 (2019). DOI: 10.1007/978-3-030-33820-6_5.
14. Ponce, H., Martinez-Villaseñor, M.L.: Interpretability of Artificial Hydrocarbon Networks for Breast Cancer Classification. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 3535–3542 (2017). DOI: 10.1109/IJCNN.2017.7966301.
15. Ponce, H., Martínez-Villaseñor, M.L.: Versatility of Artificial Hydrocarbon Networks for Supervised Learning. In: Advances in Soft Computing (MICAI), pp. 3–16 (2018). DOI: 10.1007/978-3-030-02837-4_1.
16. Ponce, H., Martínez-Villaseñor, M.L., Miralles-Pechuán, L.: A Novel Wearable Sensor-Based Human Activity Recognition Approach Using Artificial Hydrocarbon Networks. Sensors, vol. 16, no. 7 (2016). DOI: 10.3390/s16071033.
17. Ponce, H., Miralles-Pechúan, L., Martínez-Villaseñor, M.L.: Artificial Hydrocarbon Networks for Online Sales Prediction. In: Advances in Artificial Intelligence and Its Applications (MICAI), pp. 498–508 (2015). DOI: 10.1007/978-3-319-27101-9 38.
18. Ponce, P., Ponce, H., Molina, A.: Doubly Fed Induction Generator (DFIG) Wind Turbine Controlled by Artificial Organic Networks. Soft Computing, vol. 22, pp. 2867–2879 (2018). DOI: 10.1007/s00500-017-2537-3.
19. Ponce-Espinosa, H., Ponce-Cruz, P., Molina, A.: Artificial Organic Networks: Artificial Intelligence Based on Carbon Networks. Springer, vol. 521 (2014). DOI: 10.1007/978-3-319-02472-1.
20. Ponce-Espinosa, H.E.: A New Supervised Learning Algorithm Inspired on Chemical Organic Compounds. Ph.D. Thesis, Instituto Tecnológico y de Estudios Superiores de Monterrey (2013)
21. Sheta, A.F., Ahmed, S., Faris, H.: Evolving Stock Market Prediction Models Using Multi-Gene Symbolic Regression Genetic Programming. Artificial Intelligence and Machine Learning (AIML), pp. 11–20 (2015)
22. Soler-Dominguez, A., Juan, A.A., Kizys, R.: A Survey on Financial Applications of Metaheuristics. ACM Computing Surveys, vol. 50, no. 1, pp. 1–23 (2017). DOI: 10.1145/3054133.
23. Stoean, C., Paja, W., Stoean, R., Sandita, A.: Deep Architectures for Long-Term Stock Price Prediction with a Heuristic-Based Strategy for Trading Simulations. PLoS ONE, vol. 14, no. 10 (2019). DOI: 10.1371/journal.pone.0223593.

24. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van-Der–Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R., Jones, E., Kern, R., Larson, E., Carey, C.J., et al.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, vol. 17, pp. 261–272 (2020). DOI: 10.1038/s41592-019-0686-2.
25. Zheng, X., Chen, B.M.: Stock Market Modeling and Forecasting: A System Adaptation Approach. Springer, (2013). DOI: 10.1007/978-1-4471-5155-5.