Stock Markets Risk Analysis performance by Self Organized Maps AI techniques: Comparison and Implementation of three SOM methods

Gissela E. Pilliza[1,2](✉), Luis Zhinin-Vera[1,2], Rafael Valencia-Ramos[1,2], and Ronny Velasteguí[1]

[1] School of Mathematical and Computational Sciences, Yachay Tech University, 100650, Urcuqui, Ecuador.
[2] MIND Research Group - Model Intelligent Networks Development
mind-researchgroup.com

Abstract. Despite the exponential increase in the use of AI tools, the financial field has become a target just in the latest years. The stock markets meant a decisive factor for economic growth as it works as a management mechanism for money generated by the industrial force of the countries. In order to obtain the improved algorithm, this work focus on establishing the best SOM architecture for stock market treatment in an initial step. Therefore, after the literature review, the data extraction was performed using Yahoo Finance open source to get the historical data of the selected financial index. The ISOM SP40 proposed in this work uses an adequate combination of hexagonal SOM architecture and neighbor function based on Manhattan distance. Moreover, two SOM methods more denominated SOM IBEX35 and SOM NYSE were tested by the same conditions for compare, and determinate the best scenario for SP Latin America 40 data set. Thus the risk investment was analyzed with density correlations of profit, industrial area, and geography detected with an 80% of success rate using the top 9 companies in the stock index, also it was verified in a time-frequency analysis developed here with the top 6 companies reference companies from 2014-2019. The training time in the proposed ISOM SP40 method also improves two decimal places in comparison with the other tested techniques. In this sense, there is appropriated to establish that the improved algorithm was found, and it succeeds in the adaptation to SP Latin America 40 index data set.

Keywords: Self Organized Maps · Stock Market · Stock Index · S&P Latin America 40 · IBEX35 · NYSE · NASDAQ · Investment Risk.

1 Introduction

Risk analysis for stock exchange markets investment, in any region, means a relevant issue when huge amounts of money are circulating and producing around the world directly affected by economic, social, and even political events [4]. The stock exchange market or bursal field is a financial mechanism that allows to the brokers and trades the exchange and negotiation of different financial instruments just as bonds, titles, stocks, among others. Thus, the risk analysis for this purpose is conceptualized as the process in which the investors evaluates probabilistic the incidence of negative episodes on the transactional movements of capitals to avoid significant losses and perform the purchases-sells at the right time for the company [14]. Most of those analyses have been treated by traditional statistical approaches [15].

2 Related Works

Within the AI increasing area, several sub-branches have been born, being applied almost in any field. Thus, the complex modeling of the behavior of the markets makes necessary the use of complex and integral forecasting and prediction tools as artificial intelligence tools. This section of the current work analyzes some of the existing methods applied in the finances field to find the best adaptation and make the final architecture selection for the optimization and comparison if this study. Despite there are several methods implemented before, just the methods which have shown results kindred to the project objectives and good approximations were reviewed and listed below:

2.1 Bayesian Network and ANN Hybrid Method

The work developed on Bayesian Network and ANN fits in the category of AI-ANN-machine learning (ML) method, which uses a Back Propagation Learning (BPL) algorithm and Directed Acyclic Graph (DAG) structure for analyzing the bank liquidity risk [13].

When the methods approximate risk function have some troubles estimating the distribution function, but as the BN has good results clustering qualitatively, it achieves good approximations. On the other hand, as the input data were statically selected, they do not count with a dynamic setting that limits the realistic behavior closer to the bank liquidity risk [13].

2.2 Recurrent neural networks (RNNs) with transfer learning and Long-term memory

The method applied by Kraus and Feuerriegel in 2017 is an AI-ML-ANN method, which uses specific hierarchical structures and a large number of hidden layers to support financial decision [7]. It takes as an input financial disclosure documents containing sequences of words which allow its superficial analysis. After the process, the predictor target is the obtaining of the return or the tendency of the price change [7]. The main finding appears when long short-term-memory (LSTM) is applied. The approximation of the trend shows a better accuracy achieving 5.6% points according to figure. On the other hand, the method sub-utilizes the deep learning benefits as it "computes word tuples instead of processing the raw text" and at the same time, the used RNNs fail to improve the efficiency against the classic DL techniques.

2.3 Long short-term Memory Networks

Meanwhile, in 2018 Fisher and Krauss applied an AI-ML-RNN method Long short-term memory (LSTM) to predict "directional movements for the constituent stocks of the SP 500 from 1992 until 2015" [3]. As input data, it uses the monthly constituent list of Thomson Reuters enterprise from the SP 500 index in the time interval from 1989 to 2015. The purposed method can overcome the vanishing gradient problem by exploding different gradients. This paper summarizes other applications of AI technologies in several domains of business administration.

2.4 Neurocomputing

It is an AI-ML/aNN/DL method using a descriptive model for business computing has made considerable progress and offered a new opportunity for academic research and applications in many fields, especially for business activities and enterprises development. Very general and spread. This paper summarizes different applications of AI technologies in several domains of business administration. Finance, retail industry, manufacturing industry, and enterprise management are all included. In spite of all the existing challenges, we conclude that the rapid development of AI will show its significant impact on more fields worked in 2018.

2.5 Chartist Analysis

It is a method of statistical nature. An example of a chartist analysis is to establish that the market, concerning to a specific title, can be in an upward or downward trend. Thus, based on the past contribution behavior for similar situations, we can establish that to pass to specific predetermined phase.

2.6 Oscillators Analysis

It is a statistical smoothness method. The simplest example of an oscillator (and one of the most used) is the mobile average. The moving average of a period is the average of the prices of a particular title during this period. By smoothing the price curve, it is a more straightforward way to observe market trends.

2.7 Automatic Traders and Talentum

It is a method AI method which uses “automated trading systems, or automatic traders allow you to establish a series of automatic rules that dictate when to perform an operation and when to close it, so that it is executed autonomously without further human intervention. The programmed rules in these can be relatively simple, like the half-mobiles that we saw earlier, or much more complex”.

3 Self Organized Maps

In 1981 the professor Teuvo Kohonen proposed this model motivated with the idea of “abstract feature maps found in the biological central nervous systems”. This model is presented as an alternative for data sets that can not be linearly modeled, and the calibration as input works even there is a big amount of data or not. The idea of SOM is cluster and abstract the data dimensionality, to become one of the most popular unsupervised neural network techniques [6]. In general terms, a SOM is one 2D structure used as an input layer of an ANN (Fig. 1), that could have multidimensional input space. The SOM units grid is associated with the vector model in an ANN. The mapping performed in the learning process uses the topological distance between the patterns according to their similarity [1].

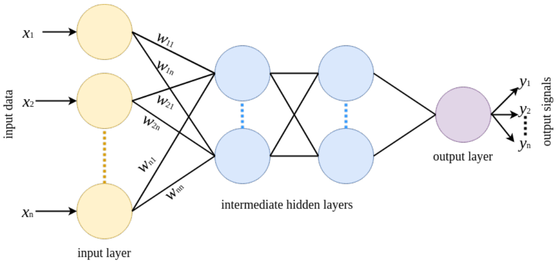


Fig. 1: Artificial Neural Network Classic Architecture *Source:[11]*

The SOM capability to detect relations in data patterns and variables of different dimensions without labeling caused that SOM is used in different areas such as chemical models[12], manufacturing process [8], biometric systems, robotics [5], and much more others [2].

SOM structure comes from the concept of ANN with an interconnection layer to layer defined by a winner model. Once the significant patterns are detected, they are located on a map according to an established geometry and distance obtained by the network. The output number and distribution depend on the number of relevant clusters catch by the network Fig.2.

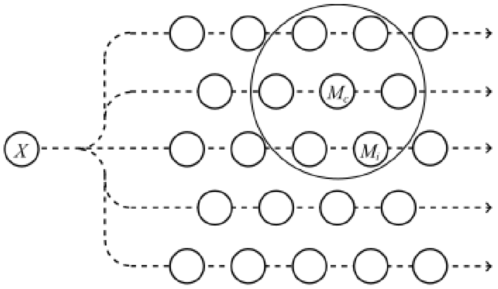


Fig. 2: Illustration of a self-organizing map. An input data item X is broadcast to a set of models Mi, of which Mc matches best with X. All models that lie in the neighborhood (larger circle) of Mc in the grid match better with X than with the rest. *Source:[6]*

4 S&P Latin America 40

In the stock markets field, there is important to emphasize that a stock market is not the same as a stock market index. While the first one is a group of companies listed in the same stock market, the second measure the performance of a group of enterprises that can be part of different stock markets. From the S&P Dow Jones series indexes arises the S&P Latin America 40, which concentrates the 70% of the capitals in Latin America.

For this work, the daily data from April of 2014 to April 2019 were extracted to being analyzed. All of them downloaded in *.csv* the format compatible with readers or processors of spread-sheets. Using a Self Organized Maps method, SP Latin America 40 stock exchange prices are going to be analyzed. These results are going to be compared with the other two SOM methods, which show acceptable results in the stock exchanges according to the studies [10], and [16].

5 Data Verification

To verify the modeling of **SP Latin America 40**, two additional data sets were selected due to its reliability in previous studies in terms of quantity and quality necessary to perform a study with Self Organized Maps. The **IBEX35** and **NIKKEI** indexes were selected to feed three different SOM due to they have more than 20 enterprises with historic data at least one year. Also, the three data sets are considered the most robust indexes in their corresponding region [11].

6 Proposed ISOM Approach

For the model to be implemented, it is expected that follows the hexagonal structure of the IBEX35 approach developed in previous risk analysis paper [10], but

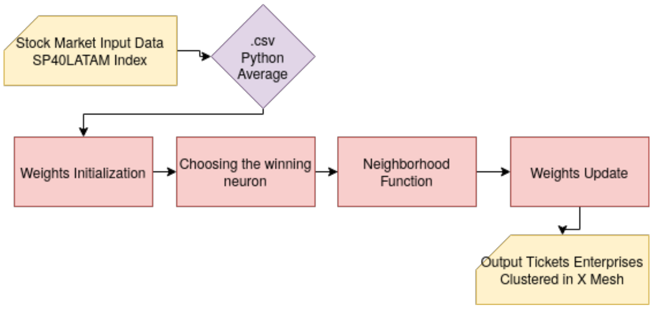


Fig. 3: Improved SOM Flow Chart*Source:[11]*

the hyper-parameters are fixed for a different test, and in the neighbor, mechanism is modified whit the Manhattan distance. Then, it would use the equations 1,2, 3, and 4, but in the weights update step according to the mathematical formula improvement, it could be applied building a topological structure representing the original surface of the system adapting the mesh organization [9]. Meanwhile, there is the automation in the initialization scheme of data the data set, through the "python prepossessing algorithm" to be entered after in Matlab, phyton was used as it shows optimal simple handling of documents preprocessing. Matlab is going to be the IDE used in the whole AI project because of the dynamic handling of the modules to build the project being compatible without the need for auxiliary algorithms with the data matrixes. In contrast with existing libraries in the traditional programming languages as C or C++ wich need of auxiliary scripts and time to implement them. The general flowchart of the proposed approach is depicted in Figure 3.

6.1 Proposed SOM Pseudo-code

Weights Initialization:

$$w(0) = random([a, b], [n, m, o]) \tag{1}$$

Choosing the winning neuron:

$$G(x(p)) = argmin_{\forall i}\{||x(p) - w_i(p)||\}\forall i = 1, 2, 3, .., n \times m \tag{2}$$

Neighborhood Function:

$$\Lambda(P_j, P_k) = Sum(Abs((P_j - P_k))) \tag{3}$$

Weights Update:

$$w_i(p + 1) = w_i(p) + \eta(p)\Lambda(P_r, P_i)(x(p) - w_i(p)) \tag{4}$$

6.2 Training Models and Resources

- The data will be prepossessed in the same way described in the above section; for the three cases, they will be introduced equally in the three algorithms IBEX35, NYSE-NASDAQ, and SPLatam40.
- Each algorithm uses its routine for training and cluster as the schemes explained before in the methodology section.
- The whole process will be run in Matlab IDE, including the prepossessing and training, and with an available in a Laptop Dell core i5 from $4th$ generation and Ubuntu operative system.

6.3 Performance Metrics

For the present work, the metrics that will allow us to measure the results in a standard manner to compare the three methods refereed before are going to be:

Topological Distance The SOM algorithms major clustering tools are based on distances, one of the most wide distances, and one that could be applied to the three selected models is the Euclidean distance. Still, as the project goal involve the modification of this function, the Manhattan distance is applied. Thus, the tautological distance metric will impact the accuracy expected in the density correlations described bellow.

Euclidean Distance The SOM algorithms major clustering tools are based on distances, the one of the most widely distances and one that could be applied to the three selected models is the Euclidean distance. It will be a quantitative metric and in this project will be calculated as follow [10]:

$$d(P_j, P_k) = \sqrt{(P_{j1} - P_{k1})^2 + (P_{j2} - P_{k2})^2} \tag{5}$$

$\forall k = 1, 2, 3, ..., M.$,
where:
P_j is the treated company,
P_k is the k-st company to get the distance, and
M is the total number of companies.

Density Correlations The density correlations [11] are qualitative metrics that are identified as areas of the graphics generated by the algorithms in each case, as is shown in Figure 4. Therefore, codes could be adapted to analyze the correlation areas by:

- General Company Tickets: distribution of the companies with more profit in the SOM structure.
- Geographical Distribution: distribution into the SOM of companies that belong to the same country.
- Industrial Areas: distribution of companies associated with the same business line such as banking, energy, food, chemicals, oil, and many others.

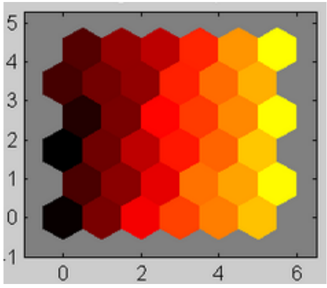


Fig. 4: SOM Density Correlation Illustration *Source:[11]*

Training Time This quantitative metric is referred to the time that the Matlab program is going to take to measure the model construction with the data set. The project will be quantified by the tic-toc function integrated into Matlab that, according to its documentation, in summary, measures the time until the program completes the last operation.

- From the three algorithms, the same result features will be extracted under the metrics founded before.
- After the three algorithms were run the Topological distance for density correlations, Training Time, will be crossed to detect the differences between the three methods.
- At the same time, the metrics will be widely analyzed to determine if the method it's efficient in the adaptation to the SP40 LATAM data set.

7 Results

The result is presented to determine the points of interest in the comparisons of ISOM SP40, SOM IBEX35, and SOM NYSE. Also, the analysis of the facts that can be deduced from the newly available information will be explored in detail. Three subsections a table is provided with three columns: the number of iterations, accuracy of the top 8 enter prices, and the execution time in seconds. The content of that columns is over understood except for the top 8 accuracies, it is about how every clustering test has its graphical result in the correspondent SOM. Thus it was affirmative if those eight companies were near or negative if they not. The Yes/No case represents a particular distribution in which the companies, even all the top 8 enterprises, were no close enough; the SOM distribution represents a relation between them in subsections. The second source is the SOM graph generated by the models and analyzed thoroughly bellow. Moreover, for the last experiment, three graphs are provided to an overall analysis trough the time [11].

Table 1: SP40 Latin America Top 9 Strongest Companies *Source:[11]*

ID	Company name	Ticker symbol	Industry	Country
31	Itaú Unibanco	NYSE: ITUB	Banking Brazil	Brazil
39	Vale	NYSE: VALE.P	Mining Brazil	Brazil
03	Banco Bradesco	NYSE: BBD	Banking Brazil	Brazil
34	Petrobras	NYSE: PBR.A	Oil Brazil	Brazil
06	Banco do Brasil	B3: BBAS3	Banking Brazil	Brazil
17	AmBev	NYSE: ABEV	Beverages Brazil	Brazil
02	América Móvil	BMV: AMX L	Telecommunications Mexico	Mexico
27	FEMSA	BMV: FEMSA UBD	Beverages Mexico	Mexico
32	Itaúsa Investimentos Itau	B3: ITSA4	Banking Brazil	Brazil

7.1 Experiment

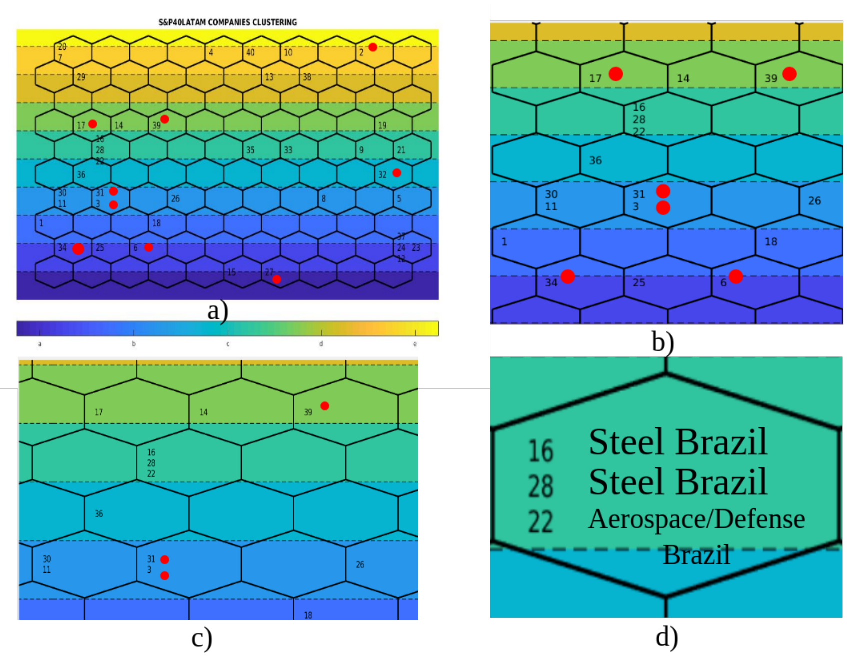
This initial experiment was developed with two objectives: first, to calibrate the parameters correctly in the proposed model, and to measure the metrics for being compared.

The first idea that can be extracted for Table 2 has a relation with the execution training time; it was found a significant improvement while the number of iterations was reduced. The referent number of iterations took form the previous studied tends to over-fitted the model. In the majority of the cases, it did no established a uniform relation of the top nine enterprises in Table 1. The best approximation was $0,579$ seconds, and it represents almost two decimal places of difference with the worst-case founded in the NYSE SOM method analyzed later. In further cases, those two decimals are essential to take advantage of the hardware resources in studies with much more amount of data.

At the same time, in Figure 5, the red points represented the top nine enterprises listed in Table 2. Here the *a*) part of the figure presents a complete overview of the SOM distribution. The verification points are well accurate clustered as at least the top 6 enterprises in part *c*) and *b*) of the graph are inside a three-range of neighborhoods. That means that from each hexagon corner, there are maxi mun three corners of separation. There, an explanation can be verified with the fourth experiment for the three essential points outside of the profit area, and it could be related to their position in the top 9 list.

On the other hand, the enterprises in the profit area that were not in the top 9 list, this time were zoom in part *d*) of the graph. Thus, verifying this behavior with the other two density correlation metrics, and there was quick to detect that they belong to the same industrial area, the metallurgic and more much consistent in the same geographical location Brazil.

In the first moment, the background colors in all graphs are expected to detect concentration areas of profit or looses, but at the end with the experimental repetitions, just the 50% of them correspond to a correct association, which that percentage it is not considered relevant. The situation for this data set and algorithm is that with a smaller size of iteration, they always find good



a) b) c) d)

Fig. 5: Clustering and Density Results for ISOM SP40 Algorithm *Source:[11]*

approximations. Still, its worst fact is that two of five in the repetitions in this experiments, the relationships are somehow ambiguous [11].

8 Conclusions

Broadly speaking, the implementation of an improved SOM algorithm for the SP40 Latin America data set was achieved successfully in terms of time execution reduction and accuracy of the density correlations. The relationships among the top nine companies of the index were corroborated in the SOM map distributions and a final time-frequency analysis [11]. Being precise in technical terms and specific metrics, there is concluded that:

- The prepossessing of the data sets were improved with practical tools as Python libraries *panda* and *numpy*, which allows the automatizing of data treatment and concatenation in different formats and compatible with multiple programming languages and development environments.
- The hexagonal architecture of the SOM have demonstrated a good performance in all the studies reviewed, specifically in the financial field for stock market prediction and clustering. Also, the algorithm implemented used this structure for being compared with the good performance of the hexagonal architectures of [10] and [16] studies showing positive results.

Table 2: Execution Time Comparison and Qualitative Results for ISOM SP40 Algorithm. *Source:[11]*

N. Iter.	Accuracy Top 8	Time (s)
10	Yes	5,79E-01
10	Yes/No	5,95E-01
10	Yes/No	5,77E-01
10	Yes	5,71E-01
10	Yes	5,79E-01
50	No	2,91E+00
50	Yes/No	2,94E+00
50	Yes/No	2,95E+00
50	Yes/No	2,96E+00
50	No	3,05E+00
200	No	1,23E+01
200	Yes/No	1,17E+01
200	No	1,83E+01
200	No	1,78E+01
500	No	3,25E+01
500	Yes/No	3,12E+01
500	Yes/No	3,39E+01
500	Yes/No	3,32E+01
500	No	3,14E+01

- The design of the SOM architecture comes along with the improvement of the topological distance used in the neighborhood function. This work demonstrated that besides the classic Euclidean distance, there is the Manhattan Distance, which reduces the machine operations without affecting the accuracy of the densities correlations and being adaptable to the SP40 Latin America index market.
- The performance of the three methods with the SP40 LATAM data set was compared, showing that the adequate algorithm for those companies is the ISOM SP40 proposed algorithm. This is corroborated with the metrics selected for the comparison execution time and the different density correlations. Thus, the higher accuracy corresponds to the proposed method with 80% overall in all the experiments against the 10%, and 5% of the other two methods.
- The execution time was reduced in almost two significant decimals having $5,79E-01(s)$ as the minimum time in the experiments with ten iterations and well-sorting distribution in SOM. In contrast, the IBEX35 method, which even with the ten iterations it achieves at least $6,15E+00(s)$ time of execution.
- The density correlations were pointed out by the Figure 5 with the top nine enterprises and the time-frequency analysis among the companies ITUB, VALE, BBD, PBRA, the profit analysis is done. Then, the geographic and

business sector were also verified in the amplified cell, which group SID, ERJ, and GGB three metallurgic enterprises from Brazil.

As future works, first, with the whole information generated and the interpretations of the study, these can be transformed in a more customer oriented tool with a graphical interface in which any user with non programming skills can easily set the parameters. It should be very visual and owns interactive indicators such as set combinations of the different data sets and methods of being presented with a click. If the use of this tool become real, the risk of investment would be minimized.

References

[1] Afolabi, M.O., Olude, O.: Predicting stock prices using a hybrid kohonen self organizing map (som). In: 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07). pp. 48–48 (Jan 2007)

[2] Deboeck, G., Kohonen, T.: Visual Explorations in Finance with Self-Organizing Maps (01 1998)

[3] Fischer, T., Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research **270**(2), 654 – 669 (2018)

[4] Hu, H., Tang, L., Zhang, S., Wang, H.: Predicting the direction of stock markets using optimized neural networks with google trends. Neurocomputing **285**, 188 – 195 (2018)

[5] Kohonen, T., et. al: Engineering applications of the self-organizing map. Proceedings of the IEEE **84**(10), 1358–1384 (Oct 1996)

[6] Kohonen, T.: Essentials of the self-organizing map. Neural Networks **37**, 52 – 65 (2013), twenty-fifth Anniversay Commemorative Issue

[7] Kraus, M., Feuerriegel, S.: Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems (2017)

[8] Li, Y., Pan, F.: Application of improved som neural network in manufacturing process quality control (03 2013)

[9] Oyana, T.J., et. al: A mathematical improvement of the self-organizing map algorithm. In: Mwakali, J., Taban-Wani, G. (eds.) Proceedings from the International Conference on Advances in Engineering and Technology. Oxford (2006)

[10] Pilliza, G.E., Román, O.A., Morejón, W.J., Hidalgo, S.H., Ortega-Zamorano, F.: Risk analysis of the stock market by means self-organizing maps model. In: 2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM). pp. 1–6 (Oct 2018)

[11] Pilliza, G.: Risk analysis of stocks markets by a merged unsupervised model, time evolution comparison, and optimization. (2020)

[12] Reker, D., Rodrigues, T., Schneider, P., Schneider, G.: Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. Proceedings of the National Academy of Sciences (2014)

[13] Tavana, M., Abtahi, A.R., Caprio, D.D., Poortarigh, M.: An artificial neural network and bayesian network model for liquidity risk assessment in banking. Neurocomputing **275**, 2525 – 2554 (2018)

[14] Times, T.E.: Definition of 'investment risk' (2019)

[15] Tkáč, M., Verner, R.: Artificial neural networks in business: Two decades of research. Applied Soft Computing **38**, 788 – 804 (2016)

[16] Wu, M.H.: Financial market prediction. Preprint submitted to arXiv (2015)