New Approach to Portfolio Creation Using the Minimum Spanning Tree Theory and Its Robust Evaluation

DOI: 10.12776/QIP.V24I2.1450

Jakub Danko, Vincent Šoltés, Tomáš Bindzár

Received: 2020-04-20 Accepted: 2020-05-25 Published: 2020-07-31

ABSTRACT

Purpose: The aim of this paper is to describe another possibility of portfolio creation using the minimum spanning tree method. The research contributes to the existing body of knowledge with using and subsequently developing a new approach based on graph theory, which is suitable for an individual investor who wants to create an investment portfolio.

Methodology/Approach: The analyzed data is divided into two (disjoint) sets – a training and a testing set. Portfolio comparisons were carried out during the test period, which always followed immediately after the training period and had a length of one year. For the sake of objectivity of the comparison, all proposed portfolios always consist of ten shares of equal weight.

Findings: Based on the results from the analysis, we can see that our proposed method offers (on average) the best appreciation of the invested resources and also the least risky investment in terms of relative variability, what could be considered as very attractive from an individual investor's point of view.

Research Limitation/implication: In our paper, we did not consider any fees related to the purchase and holding of financial instruments in the portfolio. For periods with extreme market returns (sharp increase or decrease), the use of Pearson's correlation coefficient is not appropriate.

Originality/Value of paper: The main practical benefit of the research is that it presents and offers an interesting and practical investment strategy for an individual investor who wants to take an active approach to investment.

Category: Research paper

Keywords: portfolio creation; S&P 500; minimum spanning tree; graph theory; optimization

1 INTRODUCTION

In the last decade, investors in most developed countries have been facing an environment of persistently low-interest rates. As a result of the financial and economic crisis from the turn of the years 2008 and 2009, several central banks have sought to foster economic growth and stabilize their economies by using and applying an expansionary monetary policy. In the context of financial markets and investment, this aspect results in a low-interest-rate environment. Investors in financial markets are currently facing the challenge of how to optimally appreciate their investment in terms of investment risk. The range of investment instruments available for investors is wide, as is the range of viable investment strategies and approaches.

In this article, we present the possibility of creating an equity investment portfolio using the minimum spanning tree method. Using this approach, during the training period we focus on selecting specific stocks that are part of the S&P 500 stock index. When calculating the minimum spanning tree, we used adjusted correlation coefficients as a metric to determine the distance in these structures. As a result, equities that have a low correlation of returns are close to each other in these structures. This may be interesting in the context of portfolio risk diversification. To increase the robustness of the results, we use 10,000 simulations, and for each of the 10,000 training sets, we estimate the minimum spanning tree to build four portfolios in different ways. To ensure the objectivity of the comparison, all portfolios always consist of ten equities of equal weight. Subsequently, we compare the performance of these portfolios over a test period of one year. The goal is to prove that the portfolio we had proposed, marked as portfolio A, has the best performance compared to the other alternative portfolios marked as B, C, and D. Portfolios are compared mainly with regards to average annual return and their relative variability.

2 LITERATURE REVIEW

From the time since Markowitz introduced his Modern Portfolio Theory (1952;1959), there has been a lot of literature focused on new approaches to the portfolio creation process. In our paper, we use the graph theory approach for building the investment portfolio. We focus on the minimum spanning tree method (MST). One of the first authors to introduce this approach was Mantegna (1999). From this time, there were several authors dealing with this topic – see e.g. (Onnela et al., 2003a; 2003b; Bonanno et al., 2004).

Tola et al. (2008) show that clustering algorithms are suitable for improving portfolio reliability concerning the ratio of predicted and realized risk. Naylor, Rose and Moyle (2007) use two techniques – ultrametric hierarchical tree and minimum spanning tree for extraction of a topological influence map in the market of major currencies. Birch, Pantelous and Soramäki (2015) use three methods, including MST, and compare them for filtering information from the

DAX index. There have been many more authors dealing with this topic in recent times. Wang, Xie and Chen (2017) study the US stock market from a network perspective. They use the minimum spanning tree method and also planar maximally filtered graph method (PFG) and construct MST and PFG networks in the US market at different time scales. Danko and Šoltés (2018) use graph characteristics as a stock-picking tool and propose a portfolio while minimizing its standard deviation.

Over time, two basic approaches to investment have been formed - active and passive. There is a persisting debate as to which is preferable and there are several key studies in this field.

Sorensen, Miller and Samak (1998) compare active and passive forms of investment by studying stock-picking skills and comparing them to index investing in various market conditions. They conclude that both strategies have justification: an active approach to the portfolio could be more effective when there are tough (bearish) periods on the market, but in bullish periods a passive investment strategy or indexing could outperform the active approach.

In his work, Blitz (2014) points out the problems of passive investment. He claims that passive investing is efficient only if there are a lot of active investors because they are inevitable for efficient capital markets. As an alternative to passive investing, he offers a factor investing approach.

Fahling, Steurer and Sauer (2019) examined the sample of 194 actively managed funds in German equity markets and compared the results with passive investing. The arithmetic average annual return of the sample is better than the benchmark. The results show that active funds perform slightly better in terms of risk-adjusted performance. In our paper, we focus on a method that we can classify as an active form of investment.

3 METHODOLOGY

In our research, we employ the closing prices of shares forming the S&P 500 index in the period from 2009-01-02 to 2019-10-18. In this period, we have complete data of 450 financial instruments for the period of 2,718 trading days (over 10 years).

From the daily closing prices of the stocks forming the analyzed index, we calculated the daily returns in the standard way:

$$r_i = \frac{P_i - P_{i-1}}{P_{i-1}} [\times 100\%], \quad (1)$$

where p_i is the closing price at the time i, $p_{(i-1)}$ is the closing price on the previous trading day (at the time $i - 1$) and r_i represents the daily rate of return at the time i. Thus, every single stock from the analyzed index is represented by a daily return vector.

The idea of the analysis is that we take any time interval of at least two years (assuming that one business year has approximately 253 business days) and divide it into two parts: a training and a testing set. The testing set is still 253 trading days (one year) long and the training set can be any length, at least 253 trading days and a maximum of 2,465 days (2,465 = 2,718 – 253), with the training set, immediately preceding the testing set. Simply put, we create a portfolio based on data from different time periods (at least for the previous 253 days – maximum for the last 9 years or 2,465 days) and we still test portfolio performance for the year after we acquired this data. Simplification is shown in Figure 1.

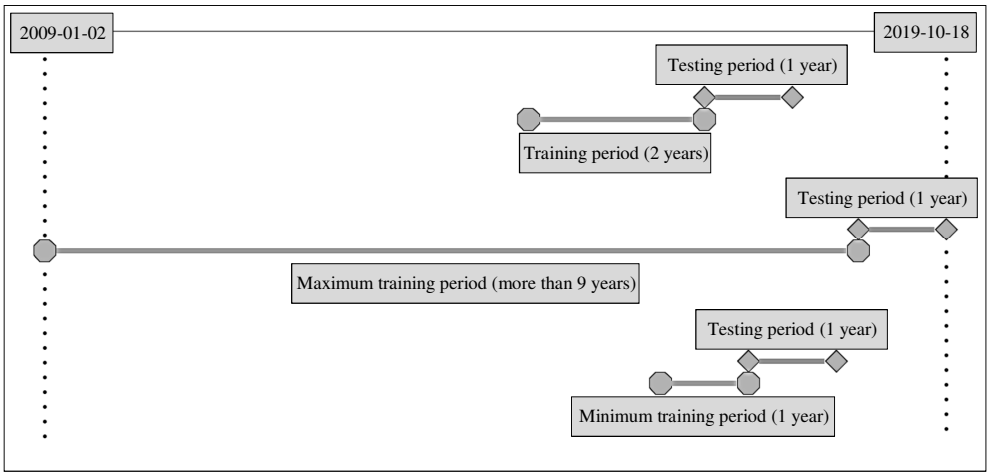

Figure 1 – Timeline

Because we want the results to be unaffected by the length of the training set, we perform 10,000 simulations and in each simulation, we select a training set of any length according to the abovementioned conditions. Thereafter we pair each training set with a testing set which immediately follows. On each training set, we calculate the correlation matrix of daily returns according to the following formula (2).

$$\rho_{i,j} = \frac{cov_{i,j}}{\sigma_i . \sigma_j}, \tag{2}$$

where $\rho_{i,j}$ is the correlation coefficient between returns of stocks i and j, $cov_{i,j}$ is the covariance between these returns and σ_i is the standard deviation of the stock's i return.

Correlations of daily returns are mostly positive, with negative correlations of daily returns in these periods of at least one year being very rare. If there are any, they are statistically insignificant and negatively linearly dependent (independent) with values close to zero. Nevertheless, these negative values must be taken into account, and the way we do this is outlined below. We decided not to transform these correlation coefficients into a distance matrix, as is common in

similar analyzes (Mantegna, 1999), but to calculate the minimum spanning tree based on correlation coefficients. Based on distance matrices and according to Mantegna, due to the character of the minimum spanning tree, they form a structure in which the most similar objects are closest to each other. Using the correlation coefficients, the least similar objects will be close to each other in these structures (those stocks that have a very low mutual correlation of returns).

Due to the fact that we need to know the distances between objects when estimating the minimum spanning tree, if we consider the correlation coefficients to be this metric, it is necessary to slightly adjust them to avoid negative values. The idea is to have the most dissimilar shares as close as possible. Taking into account that the correlation coefficient takes values from -1 (absolutely negative linear dependence) to 1 (absolutely positive linear dependence), we logically want lower correlation coefficient values to represent smaller distances and higher correlation coefficients greater distances, taking into account non-negativity. The simplest solution seems to be to adjust the correlation coefficients by adding the unit to them. Thus we get a kind of "pseudo-correlation" coefficient (distance measure), which takes values from 0 (absolutely negative linear dependence) to 2 (absolutely positive linear dependence). This fulfills the condition of non-negativity of the distance of objects, which is required for the calculation of the minimum spanning tree.

$$d(i,j) = \rho_{i,j} + 1 \tag{3}$$

In this way, we estimate the minimum spanning tree on each of the 10,000 training sets and build the portfolios in 4 different ways, denoted as A, B, C, and D in the work. To ensure a fair comparison of the results, we create the portfolios still out of ten shares with the same weights in the portfolio. With such a large number of analyzed stocks (450), the resulting structures that emerge still profile one large-scale stock (a kind of center from which the edges associate with other stocks). An example is shown in Figure 2 – *Stock Structure*, where we divided the stocks into three groups with the following color differentiation:

- The black color represents the center of the graph, i.e. (that is) stock with a maximum degree.
- The grey color shows all stocks that represent the neighbors of the graph center, i.e. we can get to the center of the graph through one edge.
- The white color represents all other stocks, i.e. we can get to the center of the graph through at least two edges.

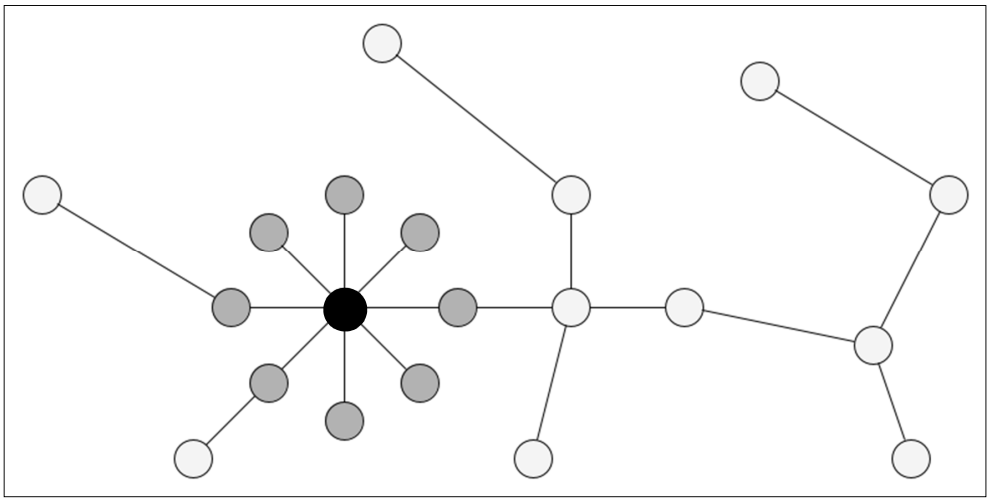

Figure 2 – Stock Structures

Type A Portfolios

We build the type A portfolios in the following way:

- We identify the center of the graph (the stock with maximum degree) that represents the first stock in the portfolio.
- Of all the shares that are incidental (adjacent) to this stock, we select nine closest to the center. This means that these ten stocks are a combination of shares that are the least similar based on the correlation of their daily returns.
- In this way, we will get ten stocks that create the type A portfolios, and each of those stocks has the same weight.

Type B Portfolios

We build the type B portfolios in the following way:

- We identify the center of the graph (the stock with maximum degree) that represents the first stock in the portfolio.
- Of all the shares that are incidental (adjacent) to this stock, we randomly select nine stocks.
- In this way, we will get ten stocks that create the type B portfolios. Each stock has the same weight.

Type C Portfolios

We build the type C portfolios in the following way:

- We identify the center of the graph (the stock with maximum degree) – this stock does not constitute an object of our interest.

- We identify all stocks that are incidental (adjacent) to this stock – these stocks do not constitute objects of our interest.
- Of all the other stocks, we randomly select ten stocks.
- In this way, we will get ten stocks that create the type C portfolios. Each stock has the same weight.

Type D Portfolios

We build the type D portfolios in the following way:

- Of all shares, we randomly choose ten shares.
- In this way, we will get ten stocks that create the type D portfolios. Again, each stock has the same weight.

From the set theory point of view, type C portfolios and type B portfolios have an empty intersection and their unification creates type D portfolios. Type A portfolios are a subset of type B portfolios and they are a subset of all possible portfolios (type D portfolios).

Assuming that using minimum spanning tree method we have identified the stocks that are, because of their greatest dissimilarity in the context of yield correlation, the most appropriate for portfolio creation, type A portfolios should perform best in terms of both real return and relative risk measured by the coefficient of variation. These should be followed by type B portfolios because they have the same principle of creation, but not under fully effective conditions (we do not select the nearest shares but randomly select the shares that are adjacent to the graph center). If we assume that the type C portfolios contain the stocks we consider to be the least appropriate, we expect these portfolios to provide us with the worst performance. We can consider type D portfolios as a benchmark, due to the fact that we build them by random stock selection. We expect that the type A portfolios we propose will perform better than portfolios B and that portfolios C will perform worse than portfolios A and B.

For each of the 10,000 simulations, we build a training set on which we estimate the minimum spanning tree, from which we create typologically 4 different portfolios. In the second part of the analysis, we consider a testing set with a length of one year (253 trading days) immediately following by the training set. In this testing set, we compare the portfolio's value at the beginning and at the end of the period to obtain the real annual return on a particular portfolio for each simulation. In this way, we obtain 10,000 real annual return values for all types of portfolios.

The risk quantification of these portfolios is a bit more complicated. We start from the covariance matrix of daily returns from the beginning to the end of the test period. The standard deviation of daily returns for this period is calculated simply by the following formula (4).

$$\sigma = \sqrt{w.\sum.w^T}, \tag{4}$$

where w represents row weight vector, $\sum$ is the covariance matrix of profitability and w^T transposed row weight vector (column weight vector).

In our case, the vector w for all portfolio types is a vector with ten equal weights (1/10). In this way, we calculate the standard deviation of daily returns, which we need to transform into a standard deviation of annual returns in order to compare it with real annual returns. We use the approximate calculation given by the formula (5).

$$\sigma_{year} = \sigma_{day} \cdot \sqrt{n_{days\ per\ year}}, \tag{5}$$

Where σ_{year} represents the standard deviation of annual returns, σ_{day} represents the standard deviation of daily returns, and $n_{days\ per\ year}$ represents the average number of days that the business year contains (we used 253 business days throughout the whole analysis).

In this way, we get the result for each simulation in the form of the composition of individual portfolios, their real annual profitability, and the approximate standard deviation of the annual profitability.

By dividing the standard deviation and the annual rate of return of a particular portfolio, we get the ratio indicator – the coefficient of variation of the annual rate of return of the portfolio, which represents a measure of relative variability according to the relation (6), which can be interpreted as inverted Sharpe ratio.

$$\frac{\sigma_P}{E(r_p)} \tag{6}$$

For portfolios of each type, we have 10,000 real annual return values and 10,000 approximate standard deviations (absolute risk) values. By averaging these values, we get the results that are presented in the next section.

4 RESULTS

In this section, we present the results we have obtained from comparing our proposed portfolio with alternative portfolios. The summary results are shown in Table 1.

Table 1 – Comparison of Portfolios

Portfolio type	A	B	C	D
Average annual rate of return	16.62%	13.95%	12.74%	13.75%
Average standard deviation	0.00192	0.00194	0.00181	0.00186
Average coefficient of variation	0.01156	0.01392	0.01422	0.01350

Type A portfolio

The type A portfolio proposed by us achieved an average annual return of 16.62% and an average standard deviation of 0.00192, which represents an average coefficient of variation of 11.56%.

As an interesting addition to the results of the analysis from the investor's point of view, we also include a more detailed description of the type A portfolios. These portfolios represent the core of our proposed method of portfolio creation. We describe these portfolios in terms of most represented stocks in them. As a central vertex in the 10,000 portfolios, the most frequently featured stock became Newmont Corporation (NEM) – 4,475 times, followed by Weyerhaeuser (WY) – 2,744 times, Expedia (EXPE) – 635 times, Dollar Tree (DLTR) – 599 times, and Netflix (NFLX) – 566 times. Other stocks were at the center of the graph less than three hundred times.

More important than the center of the graph is the success of the shares in terms of participation in the portfolio (proximity to the center), where out of 10,000 portfolios in 5,569 of them is stock NEM (Newmont Corporation), in 4,424 portfolios is stock DLTR (Dollar Tree), in 3,814 portfolios is stock EXPE (Expedia), in 3,352 portfolios is stock PEP (PepsiCo) and in 3,137 portfolios is stock ABMD (Abiomed). The frequency of other stocks is illustrated in Figure 3 by the size of the stock-specific rectangle.

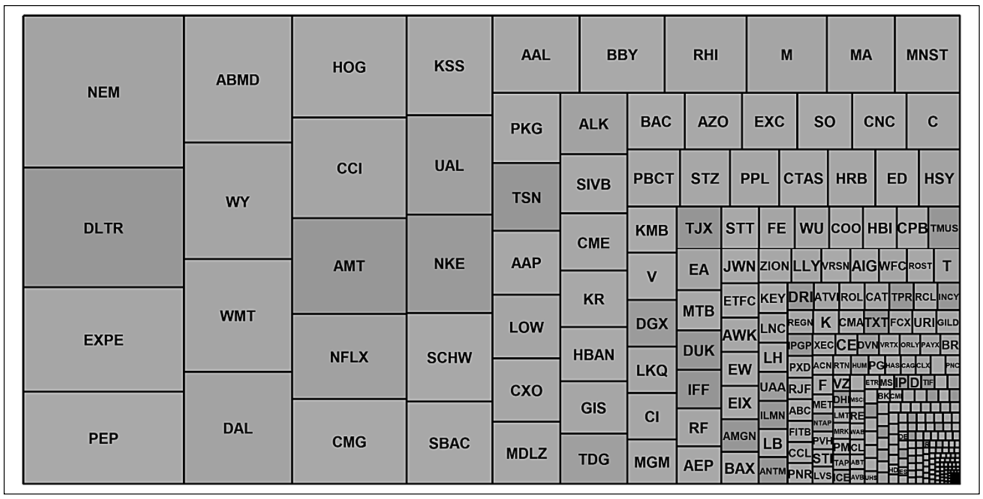

Figure 3 – Frequency in Type A Portfolio

Type B portfolio

The type B portfolio achieved an average annual return of 13.95%. The average standard deviation is 0.00194, and the average coefficient of variation is 13.92%. Compared to the portfolio proposed by us, portfolio B performed worse in all indicators.

Type C portfolio

The type C portfolio has an average annual return of 12.74%, an average standard deviation of 0.00181, and an average coefficient of variation of 14.22%. Our portfolio outperformed portfolio C in both average annual returns and the average coefficient of variation.

Type D portfolio

The type D portfolio achieved an average annual rate of return of 13.75% and an average standard deviation of 0.00186. The average coefficient of variation is 13.50%. Again, our portfolio outperformed the type D portfolio both in average annual profitability and in the average coefficient of variation.

The comparison clearly shows that in line with our assumptions, our portfolio A was able to outperform all alternative portfolios both in terms of average annual profitability and also in terms of average relative variability. From the investor's point of view, the strategy we propose brings the best appreciation of invested resources, but also the least risky investment in terms of relative variability.

5 CONCLUSION

In this work, using the minimum spanning tree method, we focused on a portfolio creation process appropriate for an individual investor. Portfolios were compiled based on data analysis from various long-term windows, while the portfolio performance evaluation was carried out based on the development for the following period of one year. The advantage of the analysis is that the results do not depend on the length of time that was chosen to build the portfolio, as we chose different time periods to increase the robustness of the results. Based on the results from the analysis, we can see that our proposed method offers (on average) the best appreciation of the invested resources and also the least risky investment in terms of relative variability – very attractive combination from an individual investor's point of view. The results we achieved were consistent with the assumptions made during the analysis. In this way, the article presents and offers an interesting and practical investment strategy for an individual investor who wants to take an active approach to investment.

REFERENCES

Birch, J., Pantelous, A. and Soramäki, K., 2015. Analysis of Correlation Based Networks Representing DAX 30 Stock Price Returns. *Computational Economics*, [e-journal] 47(4), pp.501-525. DOI: 10.1007/s10614-015-9481-z.

Blitz, D., 2014. The dark side of passive investing. *Journal of Portfolio Management*, [e-journal] 41(1), pp.1-4. DOI: 10.3905/jpm.2014.41.1.001.

Bonanno, G., Caldarelli, G., Lillo, F., Micciche, S., Vandewalle, N. and Mantegna, R.N., 2004. Networks of equities in financial markets. *The European Physical Journal B*, [e-journal] 38(2), pp.363-371. DOI: 10.1140/epjb/e2004-00129-6.

Danko, J. and Šoltés, V., 2018. Portfolio creation using graph characteristics. *Investment Management & Financial Innovations*, [e-journal] 15(1), p.180. DOI: 10.21511/imfi.15(1).2018.16.

Fahling, E.J., Steurer, E. and Sauer, S., 2019. Active vs. Passive Funds—An Empirical Analysis of the German Equity Market. *Journal of Financial Risk Management*, [e-journal] 8(2), p.73. DOI: 10.4236/jfrm.2019.82006.

Mantegna, R.N., 1999. Hierarchical structure in financial markets. The European Physical *Journal B-Condensed Matter and Complex Systems*, [e-journal] 11(1), pp.193-197. DOI: 10.1007/s100510050929.

Markowitz, H., 1952. Portfolio Selection. *The Journal of Finance*, [e-journal] 7(1), pp.77-91. DOI: 10.2307/2975974.

Markowitz, H., 1959. *Portfolio selection: Efficient diversification of investments*. New York: John Wiley & Sons.

Naylor, M.J., Rose, L.C. and Moyle, B.J., 2007. Topology of foreign exchange markets using hierarchical structure methods. *Physica A: Statistical Mechanics and its Applications*, [e-journal] 382(1), pp.199-208. DOI: 10.1016/j.physa.2007.02.019.

Onnela, J.P., Chakraborti, A., Kaski, K., Kertesz, J. and Kanto, A., 2003a. Asset trees and asset graphs in financial markets. *Physica Scripta*, [e-journal] 2003(T106), p.48. DOI: 10.1238/Physica.Topical.106a00048.

Onnela, J.P., Chakraborti, A., Kaski, K., Kertesz, J. and Kanto, A., 2003b. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, [e-journal] 68(5), p.056110. DOI: 10.1103/PhysRevE.68.056110.

Sorensen, E.H., Miller, K.L. and Samak, V., 1998. Allocating between active and passive management. *Financial Analysts Journal*, [e-journal] 54(5), pp.18-31. DOI: 10.2469/faj.v54.n5.2209.

Tola, V., Lillo, F., Gallegati, M. and Mantegna, R.N., 2008. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, [e-journal] 32(1), pp.235-258. DOI: 10.1016/j.jedc.2007.01.034.

Wang, G.J., Xie, C. and Chen, S., 2017. Multiscale correlation networks analysis of the US stock market: a wavelet analysis. *Journal of Economic Interaction and Coordination*, [e-journal] 12(3), pp.561-594. DOI: 10.1007/s11403-016-0176-x.

ABOUT AUTHORS

Jakub Danko – (J.D.) University of Economics, Prague, Czech Republic, Faculty of Informatics and Statistics, Assist. Prof., e-mail: jakub.danko@vse.cz, Author's ORCID: 0000-0002-1790-3324.

Vincent Šoltés – (V.S.) Technical University of Košice, Košice, Slovak Republic, Faculty of Economics, Department of Finance, Prof., e-mail: vincent.soltes@tuke.sk, Author's ORCID: 0000-0002-2656-5582.

Tomáš Bindzár – (T.B.) Technical University of Košice, Košice, Slovak Republic, PhD. Student, Faculty of Economics, Department of Finance, e-mail: tomas.bindzar@student.tuke.sk, Author's ORCID: 0000-0003-1652-9737.

AUTHOR CONTRIBUTIONS

J.D. – methodology, software, visualization; V.S. – conceptualisation, supervision, validation, project administration; T.B. – writing—review and editing, formal analysis, data curation.

CONFLICTS OF INTEREST

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.