DOI: https://doi.org/10.37075/EA.2021.1.07 Economic Alternatives, 2021, Issue 1, pp. 113-133

Portfolio Optimization Using an Alternative Approach. Towards Data Mining Techniques Way

Valentin Burcă[*]

Abstract

In the light of the current budget constraints, the investors face a challenge when building their stock portfolios that should lead to a minimized risk for an expected level of return. Mathematical tools have become essential for portfolio theory formulation in the last decades. In this article, our main objective is to illustrate the utility of some data mining tools and techniques, with a focus on principal components analysis and cluster analysis. The case study reveals a comparative empirical results analysis of the classical Markowitz portfolio optimization model and a combined data mining techniques model. The results show how useful data mining techniques can be for the finance area, with positive implications for investors' strategy design and implementation. Our study reveals that stocks selection requires the use of modern techniques that take in account the multidimensional perspective of investment decision. Henceforth, we propose that a debate should be launched concerning the design of stock markets design, which generally focus on simple design oriented to the stocks liquidity. In order to help investors, those indices should combine multiple dimensions of stocks definition, as the return, the risk and the liquidity of a stock are at least of the same importance from an investment decision perspective.

Keywords: principal components analysis; portfolio optimization; PVMA model; return; beta; BET indice.

JEL: C80, G11, G32.

Introduction

Investment decision optimization represents, in the actual configuration of global economy, an essential part of the decision-making process, with focus on the aim of taking advantage of the leverage capital markets offer to the investors. The strong capital markets, becoming more and more connected worldwide, translate nowadays into a core source of financing for the companies. Moreover, the different rationale of capital markets' actors, clearly explained in terms of the traditional economic theories and the recent behavioural finance theories, create greater opportunities for investors to run different practices of hedging in order to minimize the negative effects of the risk and uncertainty of the economic environment. All of these translate into the main objective of a risk manager, that imply risk minimization on a certain level of return of the stocks traded, so

[*] Faculty of Economics and Business Administration, West University of Timisoara, Romania

that the company does not only have access to funds covering their financing needs, but also benefits from the market arbitrage and makes profit.

The Romanian capital market, currently classified as a frontier market, is like the Estonian capital market, the Croatian capital market, the Slovenian capital market, or the Serbian capital market. However, the trend of the Romanian capital market is positive, towards the status of secondary emerging capital market in the next years, under the condition that the liquidity of the capital market will grow to meet the minimum requirements set by MSCI[1] (FTSE Russel, 2016).

Currently, the reference for the Romanian capital market evolution is the BET index, calculated from 19 Sept 1997, as a weighted price index reflecting the first 10 stocks with the highest free-float capitalization, traded on the regular level of BVB, with the exception of the financial investment companies. The index structure is adjusted in case there is significant influence of some corporate events, or on a regular quarterly basis.

Company capitalization is calculated as a product of: (i) the price $p_{i,t}$, (ii) the number of shares $q_{i,t}$, (iii) the free-float factor Ff_i (free tradable shares percentage from the total number of shares), (iv) the representative factor R_i (resuming a company to have more than 20% of BET structure) and (v) the price correction factor $c_{i,t}$.

$$BET_t = BET_{t-1} \cdot \frac{\sum_{i=1}^{n} p_{i,t} \cdot q_{i,t} \cdot Ff_i \cdot R_i \cdot c_{i,t}}{\sum_{i=1}^{n} p_{i,t-1} \cdot q_{i,t} \cdot Ff_i \cdot R_i \cdot c_{i,t-1}} \quad (1)$$

These are other BET index forms, focusing on different investors' objectives analysis, as the standard BET index does not reflect on a fair scale all BSE[2] industries, such as the institutional investors sector. Even if the correction factors included in the index try to solve the issues generated by the outliers, the actual construction methodology is not able to give a proper reply even regarding the liquidity criteria. But, the BET[3] index ignores the value relevance of stock returns, or risk returns, considering that a more liquid stock means that an expected average investor return is achieved at a lower stock risk level. But the bid-ask spread, showing the relation between the offer and the demand for a stock, can be less relevant than perceived by the investors, if we take into account the impact of finance behavioural theories. Those theories reject the efficient capital market hypothesis, replacing it with some more recent market hypothesis, like the fractal market hypothesis (Peters, 1994) or adaptive market hypothesis (Lo, 2004), which focus more on investors' behavioural patterns, on different time horizons and in relation with the high volatility of the capital markets.

In order to speed up market liquidity, we would propose to review the value relevance of the actual BSE index, seen as the main reference to describe the Romanian capital market. As an alternative, we look for an index that does not start only from the most liquid stocks, but from the stocks that would configure an optimized stocks portfolio as well, combining risk metrics with return metrics and liquidity measures, for stocks selection. This reference portfolio could represent a dynamic tool used by institutional investors, who have to update it, at least on a quarterly basis.

We would propose to review the results of a portfolio optimization model based on data mining techniques, combined with the PVMA[4]

[1] MSCI is a leading provider of critical decision support tools and services for the global investment community; the indices developed consider a set of criteria to assess maturity and the level of capitalization of capital markets;

[2] BSE – Bucharest Stock Exchange

[3] BET – name of Bucharest Stock Exchange index;

[4] PVMA - Portfolio with absolute minimum variance;

portfolio optimization model (portfolio with absolute minimum variation). Our algorithm consists of a running Principal Components Analysis (PCA) for stocks selection, while the Markowitz portfolio optimization model is used as a measure of efficiency in terms of risk and return. This way, we propose to ensure a proper stocks portfolio diversification, using as sequential criteria the return characteristic, the risk characteristic and the liquidity characteristic of the BSE traded stocks. In the end, we will determine the stocks portfolio structure, using the Markowitz portfolio diversification model that assumes risk portfolio minimization. Afterwards, we will compare each portfolio structure analysed, based on risk and return metrics.

The goal of the paper is to propose an alternative approach for constructing capital market indices, which can incorporate better both measures of return and risk of traded stocks on respective capital markets. This way, the current methodology that reduces to a list of most liquid stocks, will be extended with attention to the stocks with the highest ratio between risk and return. Applying classical portfolio optimization models to establish weights of each stock included in the final selection, using the Principal Components Analysis, leads to a more representative capital market index. This alternative index better reflects the ratio between the return of stocks considered in portfolio construction and the risk investors are willing to take. The liquidity of stocks is related more to the demand for specific stocks. However, as Lo (2004) highlighted, the current design of capital markets and the actual set-up of investors' mindset deviate from the rationality assumed by the efficient capital markets hypothesis. Therefore, an alternative index is needed. Such an index reflects better the measures of diversity on risk and return of stocks, based on historical data. On the other side, this way of selecting stocks lowers the effect of the short-term trends in the market, highly influenced by investors' psychological behaviour.

Our results show that the BET index, describing the BSE capital market, is not relevant for making an investment decision, as it incorporates just indirectly information about the risk and return of stocks. The proposed alternative index methodology better incorporates the information about risk and return of stocks in the portfolio selection decision, which is more representative of the capital market and from a historical point of view. The results show a positive return of the optimal portfolio of stocks selected using PCA analysis, compared with the return of the portfolio of stocks considered on BET index construction. This high discrepancy on portfolio return is obtained by having an insignificant increase in portfolio risk.

We underline the fact that our intention is not to reject traditional portfolio optimization models, but just to give an alternative tool for investors, especially in terms of data reduction. The essence of the article is to analyse an optimal combination of data mining techniques with the traditional optimization models, concentrating on simulation considerations of the proposed methodology. Thereby, we can apply the PCA model as many times as needed, without using another filtering method. We can consider, as well, a methodology considering limited PCA analysis rounds, but with a final filtering procedure, which can be cluster analysis, or any other classification model.

Review of related studies

Data mining applications in the decision-making process have grown exponentially in the last years, especially because of the statistical characteristics of those techniques that offer a wider trust in the research results.

There is a visible trend of extending the use of data mining models and techniques, even more in the finance area, where the technical analysis of the stocks is completed by extremely useful techniques such as:

- The principal components analysis, used in portfolio selection of the stocks;
- classification-trees, met in credit risk assessment models;
- neural networks analysis used in case of stock prediction models;
- cluster analysis widely used in building groups of stocks.

In the end, data mining techniques represent a sum-up of multivariate statistical analysis models, artificial intelligence (built on heuristics, rather than on statistics) and database systems. Those techniques can provide, as an outcome, some predictive models or descriptive models, extremely relevant to simplifying the decision-making process (Gorunescu, 2011).

The data mining area seems to provide a core set of tools and techniques that are used by an increasing number of researchers in all areas of research, especially because of the abilities of those techniques to solve problems of clustering, classification, function approximation and optimization, working with big data. In the area of economic research this trend is visible as well. As Liao et. al. (2012) underlined, data mining techniques become an alternative to traditional research techniques, in a wide range of research areas, including in the area of social sciences. The same trend is observed in the area of decision-making, where data mining techniques bring value added to the optimization of the management decision. The increasing focus of companies on adopting business intelligence models has led to the need of intelligent use of data available that allow managers to identify patterns in the behavior of organizations' processes (Maheshwari, 2015). Software solutions provided through applications of business intelligence have at their basis the use of different data mining techniques. This way, management will acquire a better understanding of the business model and firm interaction with its economic environment. Under those circumstances, we have underlined below some of the main directions of use of data mining in other economic research areas:

- Amani & Fadlalla (2017) have underlined the usefulness in the accounting and auditing research area. They have summarized the literature review, emphasizing the importance of data mining techniques that bring benefits to the areas of assurance and compliance, including fraud detection, business health and forensic accounting;
- Jayasree et. al. (2013), or Hassani et. al. (2018) show that the use of data mining techniques in the area of banking proves to be opportune especially in the area of commercial and consumer product marketing, credit risk analysis and credit scoring, security and fraud detection, customer segmentation, or customer relationship management;
- Bi & Cochran (2014) have made a survey of the literature review that reveal an overall framework of the use of data mining techniques to support data acquisition, storage, and analytics in data management systems in modern manufacturing;
- Silwattananusarn & Tuamsuk (2012) show the role of data mining techniques for processing and synthetizing big data in the area of knowledge management; on the other hand, Schuh et. al. (2019) have shown that data mining techniques are essential to the management of production complexity, by proving relevant results in the area of product integration,

standardization and modulization, process planning, production lead times and cycle times prediction and optimization, or in the area of value stream complexity;

- Plotnikova et. al. (2020) emphasize a general trend of adapting data mining techniques to support knowledge management from the perspective of knowledge acquisition and use in management decision; facilitate a better awareness of the business model context; highlight the need to increase the degree of formalization of the processes, essential to integrating data mining solutions with key organizational processes and frameworks.

Principal component analysis started to be widely used also in portfolio analysis, in combination with other data mining techniques, such as cluster analysis, VaR analysis, outlier analysis, extreme value analysis etc. These combinations reflect, in fact, a multiple steps problem solving methodology.

We recall the study of Alexander (2008) that illustrates an example of using PCA analysis when studying 30 DJIA[5] stocks returns from 31 December 2004 to 26 April 2006, obtaining four components that each reflect a possible portfolio stock selection. The problem with this model is the hypothesis that tests equal distribution of resources to buy stocks selected in each portfolio variant, which is not a practical approach as the selected stocks are characterized by different returns, while investors' main purpose is to maximize the profit generated by a portfolio. That is why stocks portfolio selection should not be limited only to one-step problem solving, as the first principal component analysis results do not eliminate drastically the issue of collinearity if we talk about a wide range of variables considered in the first stage of analysis. At least a two-steps problem solving methodology is recommended for a proper portfolio stocks selection model. The methodology can be translated into a multi-stages PCA analysis. Alternatively, we can perform a PCA analysis followed by a VaR analysis, or a cluster analysis together with a subsequent outlier analysis. The combinations can be various, but researchers /practitioners have to keep the model simple and reliable so that the output can be relevant for the decision-making process.

Yang (2015) decided to analyze ASX200 index value from the perspective of a potential reduction of the index composition. In order to reduce the composition of the index, he has proceeded to multiple-steps PCA analysis, extracting from the composition in each step the stocks highly correlated with components having an eigenvalue less than 1, according to the KMO[6] measure. This way, they have reduced significantly the side effects of the collinearity issue between stocks risk/return metric.

In the end, the principal components analysis is nothing than a mathematical optimization problem that maximizes the variation between the groups of a multivariate analyzed sample. Yang's option to eliminate the problem mentioned by Alexander (2008) is to determine the weights of each stock starting from principal components loadings for each stock. More precisely, in case of positive loading, the weight of a stock is determined by dividing the factor loading by the sum of all factors loading in case they are positive. In case they are negative, the absolute value of the factor loadings is used. The main idea is that the signs of those factor loadings represent actually the strategy that

[5] DJIA – Dow Jones Industrial Average

[6] Kaiser-Meyer-Olkin criteria, used to identify most relevant factors obtained running a principal components analysis;

PCA would provide in case of stock, namely, for positive factor loadings the investor will adopt a long position, while for a negative factor loading, the investor's strategy would be to adopt a short position.

Another approach widely used in the literature is the three-steps portfolio stocks selection. This approach consists of a first stage when running a PCA analysis, followed by a cluster analysis that determines families with similar stocks, from which are selected only some of the stocks based on a final set of criteria, like the case of stock VaR minimization (Fulga and Dedu, 2012).

Similar was the case of the Cardoso (2015) study, who has created 4 clusters, based on the first 2 components that resulted from PCA run on the entire stocks sample considered in the study. After all, this study was built in order to determine the model of stocks selection that would maximize the leverage an investor will get in case he analyzes three different emerging markets, with different transactions characteristics and different volatility on the market. Thus, for the final selection of a triple-pair of stocks, the co-integration approach was used.

A simpler model for stocks selection is the one proposed by Marvin (2015) that starts from the S&P 500 composition. He runs a cluster analysis based on a simple measure that considers assets rotation and the economic return of a company. In the next step, he chooses from each family the stock with the highest value of the Sharpe ratio.

Additionally, we recall the approach to selecting stocks, or a portfolio composition determination, illustrated by Craighead and Klemesrud (2002), who determine five portfolio strategies, starting from defining two clusters based on the outlier measure using the Kalman Filter/Smoother model.

PCA analysis, or cluster analysis, can be used as core tools in studying the so-called market contagion phenomenon, in the light of more and more worldwide-connected capital markets (Nobi and Lee, 2016). A similar use can be assigned for those data mining techniques, when analyzing industry effects or country effects on a portfolio analysis.

We shall not forget the main use of principal component analysis in the accounting and finance area, as PCA is widely used when regression models are built starting from a simplified list of measures that characterize a company (EPS, leverage, EVA, net profit etc.), or a country (financial stability, capital market capitalization, economic growth, financial education, investor protection legal framework etc.). All those factors can influence either a company's traded shares at a microeconomic level, or the market index at a macroeconomic level, the PCA mission being the reduction of the list of those factors, resulting in a linear function that determines a final score extremely useful for a company/country dynamic ranking. This way, with this score for each stock/index, we can run either a cluster analysis further, or different other classification/ranking techniques that would result in a short list of stocks to be used for an optimal portfolio construction.

Methodology research

The only studies analyzing the stocks selection on Bucharest Stock Exchange level, using data mining techniques, are from Fulga et. al. (2009), Fulga and Dedu (2012). In Fulga et. al. (2009), the selection of the stocks is done on the basis of selecting a stock from each cluster determined by the factors that result from a PCA run in the first step. In Fulga and Dedu (2012) the selection model is optimized by considering as final selection criteria of a stock from each cluster the stock with the minimal VaR value.

The methodology we propose, consist of three steps of problem solving:

- in the first step we will run a PCA, in order to reduce the initial sample considered in our analysis, starting from the stocks' weekly returns analysis; this way we will eliminate the stocks which are collinear in terms of returns; this way we aim for a portfolio diversification, not only for a portfolio return maximization;
- in the second step we will run an additional PCA, this time based on stocks' weekly beta analysis; this way, we will reduce the multicollinearity on stocks' volatility determined by the market BET index; under those circumstances, we can follow the core mean-risk framework, but in an original approach;
- in the third step, we have run a cluster analysis, considering the following set of quantitative measures characterizing the remaining stocks: (i) weekly traded volumes, creating this way a connection with the actual main criteria on BET index construction, (ii) weekly stock returns, (iii) weekly stocks market beta. In the end, from each cluster only the first stocks with the first 5 highest value of the following index trying to provide an aggregate final stock index were chosen

$$Index_i = Volume_{week_i} \cdot \frac{Return_{week_i}}{\left|Beta_{week_i}\right|} \quad (2)$$

We considered the absolute value of $Beta_{week_i}$ as it can take negative values as well. However, this measure is used as a filter that considers stocks liquidity as well. The nature of the stocks' risk is already included in the PCA analysis. Our model is aimed to maximize the return, in terms of a minimum accepted risk portfolio, which can be negative. But a negative return, no matter what the nature of the risk is, will surely deteriorate the overall portfolio return, the reason why we have considered the real value of $Return_{week_i}$, so that we can extract the negative values from the potential portfolio. We have not considered the *VaR* methodology, as we consider a complete year returns data analysis, while such a methodology is useful especially for short-term analysis.

Principal components analysis

Through the Principal components analysis (PCA) mathematical model we have translated the vector representing initial portfolio of n stocks (dimensions), into a new portfolio defined by m dimensions, the result being a set of k principal components.

Each principal component (dimension) is described by a linear function, given by the expression below:

$$w_i = \alpha_1^{(i)} \cdot x_1 + \alpha_2^{(i)} \cdot x_2 + \cdots + \alpha_n^{(i)} \cdot x_n$$

where $i = \overline{1,m}$. If we denote $w = \begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{pmatrix}$, being the vector of the m principal components, the matrix of each dimension coefficients

$$A = \begin{pmatrix} \alpha_1^{(1)} & \alpha_2^{(1)} & \alpha_3^{(1)} & \alpha_4^{(1)} \\ \alpha_1^{(2)} & \alpha_2^{(2)} & \alpha_3^{(2)} & \alpha_4^{(2)} \\ \alpha_1^{(3)} & \alpha_2^{(3)} & \alpha_3^{(3)} & \alpha_4^{(3)} \\ \alpha_1^{(4)} & \alpha_2^{(4)} & \alpha_3^{(4)} & \alpha_4^{(4)} \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}$$

the initial stocks portfolio vector, leading to $w = A^t \cdot x$.

The scope of PCA analysis is eventually to maximize:

$$Var(w) = \alpha^t \cdot \sum \cdot \alpha \quad (3)$$

with the constraint that $\alpha^t \cdot \alpha = 1$, which would lead, by applying Lagrange multiplier method, to first condition $\sum \cdot \alpha = \lambda \cdot \alpha$, that means the solution is represented by the covariance matrix (Σ) eigenvectors $(\lambda \cdot \alpha)$, with the variance of a principal component $Var(w_i) = \lambda_i$. In order to maximize $Var(w)$ it is necessary that we choose first the highest eigenvalue (noted by λ_i) and determine its corresponding eigenvector. Nevertheless, determining eigenvalues reduces to solve the system below:

$$\begin{cases} (\sum - \lambda_i \cdot I) \cdot \alpha^{(i)} = 0 \\ \left(\alpha^{(i)}\right)^t \cdot \alpha^{(i)} = 1 \end{cases} \quad (4)$$

The process continues with finding the next eigenvalue and its corresponding eigenvector, for all n initial components. In the end, we will realize that it can be rearranged the principal components descending $(\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n)$ as $Var(\lambda_1) \geq Var(\lambda_2) \geq Var(\lambda_3) \geq \cdots \geq Var(\lambda_n)$, reflecting each component contribution to the total variance of the covariance matrix $(Var(w) = tr(\sum) = \sum_{i=1}^{n} \lambda_i)$.

In our study, the PCA analysis will be used in order to select the stocks with the highest correlation (lowest factor loading) in return. We have considered a maximum level of factor loading as a criterion of elimination the value of *0.50*. Once the first round of PCA is performed, we will continue with a second round PCA, either on the same metric, or by changing it, in order to get a combination between the return of stocks and the associated risk. The change of the metric, between the two rounds of PCA, is reflected on the results provided in the sensitivity of results section.

Portfolio diversification model

The weekly return for a stock was calculated as average, omitting any dividend payment, considering its traditional defining relation $r_i = \frac{P_1 - P_0}{P_0}$, where P_1 is the price on day t_1 and P_0 is the price on day t_0. The weekly risk of a stock, reflected by its market beta, is determined using the relation below:

$$b_i = \frac{\sum_{t=1}^{n}\left(r_{i_t} - \bar{r}_i\right)\cdot\left(r_{m_t} - \bar{r}_m\right)}{\sum_{i=1}^{n}\left(r_{m_t} - \bar{r}_m\right)^2} \quad (5)$$

where r_{i_t} and $\bar{r}_i$ are the daily return and the weekly return of stock i at time t; while r_{m_t} and $\bar{r}_m$ are the daily return and the weekly return of a stock at time t.

The portfolio return for n stocks is determined with the relation below:

$$r_p = \sum_{i=1}^{n} x_i \cdot r_i \quad (6)$$

The portfolio risk is given by the expression below:

$$\sigma_p^2 = \sum_i \sum_j x_i \cdot x_j \cdot \sigma_{ij} \quad (7)$$

where σ_{ij} is the covariance between the evolution of return of stock i and the evolution of the return of stock j.

The composition of the optimal portfolio we want to determine will be done applying the minimization risk portfolio diversification model (PVMA model), expressed by the following optimization program:

$$\begin{cases} Min\ \sigma_p^2 = \sum_i \sum_j x_i \cdot x_j \cdot \sigma_{ij} \\ \sum_i x_i = 1 \end{cases} \quad (8)$$

where σ_p is portfolio risk, $\mathbf{x_i}$ is the weight of stock i in the optimal portfolio and σ_{ij} is the covariance between stock i return risk evolution and stock j risk evolution.

The Lagrange function to be minimized is described by the relation $L = Min\left(\frac{1}{2} \cdot \sum_i \sum_j x_i \cdot x_j \cdot \sigma_{ij} + \mu \cdot (\sum_i x_i - 1)\right)$. The solution of the problem follows the system of equations below:

$$\begin{cases} \frac{\partial L}{\partial x_i} = \sum_j x_j \cdot \sigma_{ij} + \mu \cdot 1 = 0 \\ \frac{\partial L}{\partial \mu} = \sum_j x_i = 1 \end{cases} \quad (9)$$

Data collection

12,348 daily prices for the first Bucharest Stock Exchange most liquid *69* stocks considered in the study were collected, throughout the entire year *2016*, as illustrated in **Table 1**. There is no difference based on the activity area. Part of the stocks were eliminated from the beginning because of the small number of transactions throughout the year, as shown in **Table 1**, only *45* stocks remaining for our analysis.

Our analysis will try to underline the marginal effects of considering additional steps in the stocks selection procedure. Consequently, we will compare the return

and the risk associated with each portfolio obtained in case of the following scenarios:

- optimization performed on a sample of stocks, considering only a preliminary filter based on the stock index we've defined previously;
- optimization performed on a smaller sample of stocks, considering two steps, namely: one PCA round and a round of filtering top *10* stocks, based on the stock index;
- optimization performed on an even smaller sample of stocks, considering three steps, namely two PCA rounds and a round of filtering top *10* stocks, based on the stock index.

Table 1. *Sample selection*

Stocks included on our selection analysis					Stocks taken out from selection analysis, because of few transactions			
ALR	BCM	CMF	ELGS	PEI	ARM	ARTE	BRM	CBC
ALT	BIO	CMP	ELJ	PPL	CMF	ECT	MECF	PPL
ALU	BRD	CNTE	ELMA	PREH	RMAH	SNO	TUFE	UZT
ARM	BRK	COTE	EPT	PTR	ART	BCM	CAOR	PEI
ARS	BRM	COTR	FP	RMAH	COTR	ELJ	PEI	PREH
ART	BVB	EBS	IMP	ROCE	RTRA	STZ	UAM	VEZY
ARTE	CAOR	ECT	MECF	RPH				
ATB	CBC	EFO	OIL	RRC				
BCC	CEON	EL	OLT	RTRA				
SIF3	SNG	SNP	STIB	SCD				
SIF4	SNN	SOCP	STZ	SIF1				
SIF5	SNO	SPCU	TBM	SIF2				

Through sensitivity analysis different combinations of PCA metrics are performed, in order to understand the impact of return metrics considered in the analysis. This way, we want to emphasize the marginal effect of each PCA analysis and of the index selection filtering procedure. We want to understand as well, if portfolio diversification can be improved, either considering only return metric, or combining the return metric with the risk metric. Our analysis is limited to the stocks with positive returns, as no investor will look for stocks with average negative returns.

For statistics, the software used is SPSS. 20.

Results and discussions

Assuming the efficient capital markets hypothesis, we conclude that the investors' main objective is to maximize a stocks portfolio return at a minimal overall risk. Nevertheless, stocks liquidity will be considered as well as filtering criteria, even if the Romanian capital market is classified as a frontier and standalone market, meaning there is a lower international capital market integration.

We have focused our aim especially on illustrating the usefulness of data mining techniques for stocks portfolio management, especially in case of stocks selection. In **Table 2** we summarize the lists of stocks taken out of our initial portfolio composition, following each step described in the section on the methodology research of the paper.

Table 2. *Stocks portfolio composition*

Stocks selected after negative returns elimination				**Stocks selected after first principal components analysis** *- stocks' return* correlation avoidance			**Stocks selected after second principal components analysis** *- stocks' risk* correlation avoidance		
ALR	BRK	EFO	OIL	ALR	ELGS	SIF4	ALR	ELGS	SIF4
ALU	BVB	ELGS	OLT	BRD	FP	SIF5	BRD	FP	SIF5
ARS	CEON	ELMA	PTR	BVB	PTR	TEL	BVB	PTR	TEL
BCC	CMP	EPT	ROCE	CEON	RRC	TGN	CEON	RRC	TGN
BRD	CNTE	FP	RPH	CMP	SIF1	TLV	CMP	SIF1	TLV
SPCU	TBM	TEL	TGN	EFO	VNC		EFO	VNC	
VNC	RRC	SIF1	SIF4						
SIF5	SOCP	TLV							
Index analysis				ALR	BVB	CEON	CMP	ELGS	VNC
				PTR	RRC	SIF1	SIF4		

Source: portfolio selection, based on statistics performed with SPSS 20.0

At the first stage, we eliminate stocks with small trading volumes, as they are not of high interest for investors, resulting in the elimination of *24* stocks. Afterwards, we have eliminated the stocks with *negative returns*, as they are not attractive for investors at all, resulting in the elimination of *14* stocks. Starting from a sample of *45* stocks, we have reached a portfolio that consists of *31* stocks. As a third step in our approach, we perform the first round of PCA analysis, based on stock returns, generating a decrease of our portfolio composition to only *17* stocks. The next step is to combine the measure of stock returns with the measure of stock risk, by performing an additional PCA analysis, this time considering the maximization of variance of stocks' beta measure, which does not affect our portfolio composition.

As a final step, we keep top ten stocks that remained on our portfolio, based on the stocks index value. This way, we ensure diversification of our portfolio, by reducing the collinearity between different measures of stocks performance, such as the liquidity, return and the risk metric.

The first PCA analysis is performed by *maximizing the return variance* of the sample of *45* stocks, in order to get a diversified portfolio with high stock returns. In **Table 3**, we provide results on total variance decomposition per relevance of each of the *31* stocks included in our initial portfolio. The *11* principal components, having an eigenvalue greater than *1*, account for *72.02%* of the total variation of stock returns, which fairly represents the variation of initial stocks selection, from the perspective of a reduced stock portfolio. Selected stocks were the ones that had resulted in component loadings higher than *0.50* in absolute value, as the main criteria of selection is that stock returns should be highly correlated with the principal components identified, as highlighted in **Table 5**.

Table 3. *Total variance explained for first round of PCA*

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,078	16,380	16,380	5,078	16,380	16,380	3,430	11,066	11,066
2	2,801	9,036	25,417	2,801	9,036	25,417	2,961	9,552	20,618
3	2,339	7,544	32,961	2,339	7,544	32,961	2,153	6,945	27,563
4	1,940	6,258	39,219	1,940	6,258	39,219	2,126	6,859	34,422
5	1,828	5,897	45,116	1,828	5,897	45,116	1,866	6,018	40,440
6	1,683	5,428	50,544	1,683	5,428	50,544	1,823	5,881	46,321
7	1,523	4,913	55,457	1,523	4,913	55,457	1,760	5,678	52,000
8	1,460	4,711	60,168	1,460	4,711	60,168	1,660	5,354	57,354
9	1,292	4,167	64,335	1,292	4,167	64,335	1,547	4,991	62,345
10	1,240	4,001	68,336	1,240	4,001	68,336	1,502	4,844	67,189
11	1,141	3,682	72,018	1,141	3,682	72,018	1,497	4,829	72,018

Source: statistics performed with SPSS 20.0

Analyzing factor loadings in **Table 3**, corresponding to stocks for the first component identified, explaining about *16.38%* of the total returns variation, we can see that most of them are stocks raised by companies from the financial services sector (BRD, FP, SIF1, SIF4, SIF5, BVB), having the highest factor loadings. The higher diversity in terms of risk and return for the financial sector stocks is obvious, as they are subject to capital market hedging practices. On the other hand, in the production sector, the main purpose for public listing is more related to the acquisition of funding sources at lower costs.

The second PCA analysis is performed *maximizing the risk variance* of the remaining stocks sample, so that the portfolio with diversified stocks returns can be balanced in terms of risk basis as well. In **Table 4**, we provide results on total variance decomposition per relevance of each of the *17* stocks included in our initial portfolio. This time, the analysis led to no other stock proposed for elimination, as no stock has been described by an absolute value of the loading factor less than *0.50*. According to the results from **Table 6**, we could affirm that our portfolio is balanced from a risk-return variance perspective, as there is only one stock that does have a factor loading lower than *0.50*. Accordingly, BRD must be eliminated from our portfolio composition.

Table 4. *Total variance explained for second round of PCA*

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,806	16,507	16,507	2,806	16,507	16,507	2,110	12,413	12,413
2	2,188	12,871	29,378	2,188	12,871	29,378	2,018	11,869	24,282
3	1,833	10,784	40,162	1,833	10,784	40,162	1,844	10,849	35,131
4	1,589	9,346	49,509	1,589	9,346	49,509	1,779	10,465	45,596
5	1,524	8,966	58,474	1,524	8,966	58,474	1,754	10,319	55,915
6	1,200	7,058	65,532	1,200	7,058	65,532	1,496	8,797	64,712
7	1,103	6,488	72,020	1,103	6,488	72,020	1,242	7,308	72,020

Source: statistics performed with SPSS 20.0

As a final step, we filter the portfolio, by keeping only the first *10* stocks in terms of the value of the corresponding index calculated. The threshold considered in this analysis is debatable, but our main goal is to reflect the impact of data-mining techniques on stocks selections, rather than finding an optimal threshold based on an index analysis. This marginal analysis is reflected in **Table 5**.

The portfolios considered are determined in the different steps, using the PVMA portfolio optimization model, as described in the methodology research section. We observe in **Table 5** that the optimal portfolio selection obtained by applying just the index analysis, ensures the investor a portfolio return of *0.275%*, while the initial portfolio has a return of *0.283%*, meaning a difference of *-0.008%*. This approach leads to a lower portfolio return and a much higher portfolio risk rate, because of the collinearity between stock returns and especially stock risk rates.

Table 5. *Selection of top 10 stocks, based on index value*

Symbol	Volumes (log.)	Return	Risk (beta)	Index
CMP	18,87	0,00351	0,07429	,8915
VNC	11,83	0,00367	0,07540	,5762
RRC	15,83	0,00263	-0,07991	,5218
ELGS	13,95	0,00206	-0,05554	,5170
CEON	15,96	0,00580	-0,18019	,5136
PTR	5,99	0,00460	-0,05569	,4955
ALR	10,94	0,00785	0,37140	,2310
SIF1	14,64	0,00137	0,13025	,1537
BVB	8,04	0,00053	-0,03306	,1301
SIF4	15,12	0,00024	0,03146	,1166

Source: calculation performed with Microsoft Excel

Table 6. *Component matrix with most relevant dimensions*

Results from first round of PCA - based on return metric -											
Symbol	Component										
	1	2	3	4	5	6	7	8	9	10	11
ALR	*0,632*					0,463					
ALU		-0,425		0,306	0,413		-0,346				
ARS	-0,343		0,408								
BCC	0,420	-0,401				0,392		-0,328			
BRD	*0,634*						0,451	0,344			
BRK								-0,301	0,440	-0,309	
BVB	*0,608*						0,339				
CEON		*0,523*					-0,331			0,316	
CMP			*0,716*								-0,300
CNTE				-0,337				0,312	0,405		
EFO	*0,524*		-0,315				-0,396				
ELGS				-0,454	*0,616*						
ELMA							0,369	0,342			-0,453
EPT		0,443	-0,342			0,470		0,362			
FP	*0,668*	0,433									
OIL			0,469					0,362			0,362
OLT				0,343		0,370					
PTR		*0,551*	0,334	0,324						0,308	
ROCE		0,433		0,324					-0,319		
RPH				-0,402		-0,454				0,340	
RRC		0,303	0,401	*0,572*							
SIF1	*0,803*										
SIF4	*0,741*										
SIF5	*0,816*										
SOCP			0,444				0,332	-0,495			
SPCU	0,403				0,382					0,446	
TBM				0,428		-0,367	0,378				0,302
TEL		*0,591*	-0,361	-0,403							
TGN	*0,574*	0,349						-0,369			
TLV	0,318	0,384							*0,550*		
VNC					*0,688*						

Results from first round of PCA - based on risk metric -							
Symbol	Component						
	1	2	3	4	5	6	7
ALR					*0,545*		
BRD		-0,336	-0,487	0,449		-0,324	
BVB	*0,646*					0,453	
CEON			*0,647*	-0,301			
CMP	-0,389	*0,560*					
EFO						*-0,606*	*0,573*
ELGS		0,352	*0,571*				
FP		*-0,620*		*-0,511*			
PTR	*0,726*		0,350		0,379		
RRC	*0,604*				0,548		
SIF1	-0,383		0,403	0,384			*-0,515*
SIF4		*0,578*					
SIF5	-0,323	*-0,694*	0,435				
TEL	-0,401			*-0,545*	*0,520*		
TGN	*-0,609*	0,303				0,315	
TLV	0,431		0,338		*-0,526*		
VNC	*0,519*	0,485		-0,369		-0,311	

Source: factor loadings, based on statistics performed with SPSS 20.0

Table 7 summarizes the descriptive statistics related to each portfolio considered in our analysis. The section in this table is dedicated to the *traditional selection model* examines the portfolio of stocks remaining after eliminating the stocks with negative returns. The section dedicated to the *Data-Mining selection model* provides results describing the performance of portfolios obtained using principal components analysis (PCA).

Once the first PCA analysis is applied, the portfolio return decreases from the percentage of *0.283%* to the level of *0.194%*, especially because the stocks selected have a lower return (a mean of *0.161%* compared to *0.187%*). As expected, because of sample variance maximization, the risk is higher, as the portfolio includes a larger range of risk classes (a beta of *0.008* compared to *0.086*). However, once we select only the first *10* stocks based on the index analysis, the portfolio return increases from *0.275%* to *0.308%*, while the portfolio risk increases from the value of *0.00077%* to a value of *0.00108%*, meaning an increase of portfolio return of a higher amplitude than the portfolio risk increase.

Table 7. *Portfolio diversification (risk-return model analysis)*

Model	**Techniques used**	**Count stocks**	**Mean return**	**Std. dev. return**	**Mean beta**	**Std. dev. beta**	**Portfolio return**	**Portfolio risk**
Traditional selection model	*Initial sample of stocks*	45	0.187%	0.868%	0.086	1.280	0.283%	0.00019%
	Stocks eliminated after index analysis	14	0.111%	0.830%	0.099	1.185	0.275%	0.00077%
	Remaining stocks after index analysis	31	0.350%	0.949%	-0.011	1.314		
Data-Mining selection model	*Stocks eliminated after PCA I* (return-based)	14	0.166%	0.825%	0.008	1.060	0.194%	0.00073%
	Remaining stocks after PCA I (return-based)	17	0.161%	0.890%	0.182	1.450		
	Stocks eliminated after PCA II (risk-based)	0	0.166%	0.825%	0.008	1.060	0.194%	0.00073%
	Remaining stocks after PCA II (risk-based)	17	0.161%	0.890%	0.182	1.450		
	Stocks eliminated after index analysis	7	0.157%	0.835%	0.032	1.115	0.308%	0.00108%
	Remaining stocks after index analysis	10	0.187%	0.908%	0.240	1.572		

Source: portfolio performance, based on statistics performed with SPSS 20.0

In **Table 8** we can observe that the portfolio built using data mining techniques is higher than the portfolio prescribed by BET composition index, in terms of portfolio return, as our portfolio has a return of *0.308%*, while BET optimal portfolio structure provides a negative return. On the other hand, our portfolio has a higher risk than the one indicated by BET composition. Under those circumstances, we could affirm that only a combination of data mining techniques

with the index analysis will lead to a higher portfolio return level.

Overall, the results in **Table 8** show that portfolio diversification is essential to getting higher returns. However, this higher portfolio return is achieved, assuming a higher portfolio risk. Instead, using PCA analysis in the selection of stocks, lead to a slight decrease of portfolio risk measure, because of the reduction of collinearity on stock returns. This evolution of portfolio risk becomes even more obvious when looking at the performance of portfolio selecting the first 10 stocks selected using PCA analysis in two rounds. Through PCA analysis, we created a high gap in terms of stock returns and risk, the reason why the top ten stocks selected after the highest value of the aggregate index, lead to a better performing portfolio, achieving a portfolio return of *0.308%*, assuming a risk reflected by a portfolio beta of *0.0011%*. Thus, using PCA analysis helps we differentiate better stocks with high performance from stocks with moderate or insignificant performance. Considering the relation of our index score as a sorting criterion, we filter exactly the stocks with wide gap between stock return and the beta of respective stock. Therefore this last filtering step becomes essential, as it ensures elimination of stocks with a high correlation between their return and the risk investors assume is involved in such returns.

Table 8. *Comparative analysis on portfolio metrics*

Data-mining techniques						BET index					
Shares selected	*Weight*	*Volume* (log)	*Return*	*Beta*	*Index*	*Shares selected*	*Weight*	*Volume*	*Return*	*Beta*	*Index*
ALR	10.58%	10.94	0.785%	37.14%	0.231	BVB	15.85%	8.04	0.05%	-3.31%	0.13
BVB	13.46%	8.04	0.053%	-3.31%	0.130	TGN	11.20%	12.16	0.07%	9.48%	0.093
CEON	13.64%	15.96	0.580%	-18.02%	0.514	BRD	*-7.55%*	15.7	0.10%	19.84%	0.082
CMP	8.07%	18.87	0.351%	7.43%	0.892	TLV	14.36%	16.5	0.10%	24.30%	0.071
ELGS	14.46%	13.95	0.206%	-5.55%	0.517	FP	*-1.27%*	7.41	0.09%	15.74%	0.043
PTR	8.56%	5.99	0.460%	-5.57%	0.496	TEL	24.77%	17.43	0.00%	22.91%	0.002
RRC	*-2.17%*	15.83	0.263%	-7.99%	0.522	EL	27.95%	3.39	-0.06%	9.64%	-0.021
SIF1	6.85%	14.64	0.137%	13.02%	0.154	SNP	5.52%	2.97	-0.24%	21.49%	-0.033
SIF4	17.50%	15.12	0.024%	3.15%	0.117	SNG	1.24%	15.91	-0.11%	20.46%	-0.088
VNC	9.04%	11.83	0.367%	7.54%	0.576	SNN	7.93%	14.31	-0.18%	7.54%	-0.35
Portfolio return					0.308%						-0.023%
Portfolio risk					0.0011%						0.001%

Source: portfolio performance, based on statistics performed with SPSS 20.0

The PMVA model has been used as a benchmark to determine the optimal structure of a portfolio. If principal components analysis has been considered for stocks selection, the PVMA model is used to determine the weight each stock has to have on the final portfolio. For this purpose, we want to check if the BET index is relevant for institutional investors' strategy of diversifying their stocks portfolios, compared to the selection that resulted by performing principal components analysis.

Based on the results provided in **Table 8**, it is highlighted that the actual methodology of the BET index composition is not relevant

for the investors' investments diversification strategy. Considering stocks included in the BET index calculation would lead an investor to a negative return of *-0.023%* for a diversified portfolio. Instead, considering stocks selection determined by using principal components analysis, the optimal portfolio would generate a portfolio return of *0.308%*, way higher than the one generated by optimizing a portfolio with the BET index components. Consequently, it seems that liquidity is not enough for investors to balance their portfolios, in order to hedge against shocks on capital markets. Additionally, we can observe that even the risk investors should assume is oriented towards the portfolio determined using data-mining techniques.

Sensitivity analysis

It is obvious that each criterion of selection of stocks will lead to different results in terms of the sample of stocks selected. Moreover, the choice of a model used to determine the portfolio structure based on the stocks selected can lead to different results. Our main question in this study is if using data mining techniques, irrespective of the criteria of stocks selection used, will lead to higher portfolio returns.

Our approach, in the first scenario already analyzed in the previous section, was designed to reflect the mean-variance well-known approach on capital allocation models. Investors generally show a higher aversion to potential loss than aversion to the risk associated with a capital allocation decision. In order to reduce the potential loss and reduce the risk associated, the investors prefer to choose a diversification strategy. Hence, through diversification itself the portfolio risk is reduced, meaning that the focus should be on enlarging the space vectors of portfolio return maximization. This way, we first have run the first PCA analysis based on stocks returns variances. As stocks return varies based on the premium risk rate, in order to obtain diversity on stocks reaction to the systemic capital market risk, we have run an additional PCA analysis, considering this time the maximization of variance on the stocks beta index.

Table 9. *Implications of basis considered for PCA analysis*

						Steps: 1. *Index analysis*		**Steps:** 1. *PCA analysis* 2. *Index analysis*		**Steps:** 1. *PCA analysis* 2. *PCA analysis* 3. *Index analysis*	
Techniques used	Count stocks	Mean return	Std. Deviation return	Mean beta	Std. Deviation beta	Portfolio return	Portfolio risk	Portfolio return	Portfolio risk	Portfolio return	Portfolio risk
Stocks eliminated after PCA I (return metric)	14	0.161%	0.890%	0.182	1.450	-		0.194%	0.00073%	0.194%	0.00073%
Remaining stocks after PCA I	17										
Stocks eliminated after PCA II (return metric)	0	0.161%	0.890%	0.182	1.450	-		-		0.194%	0.00073%
Remaining stocks after PCA II	17										
Stocks eliminated after index analysis	7	0.187%	0.908%	0.240	1.572	0.275%	0.00077%	0.308%	0.00108%	0.308%	0.00108%
Remaining stocks after index analysis	10										
Stocks eliminated after PCA I (risk metric)	10	0.328%	0.825%	0.113	1.374	-		0.258%	0.00033%	0.258%	0.00033%
Remaining stocks after PCA I	21										
Stocks eliminated after PCA II (risk metric)	7	0.132%	0.821%	0.163	1.374	-		-		0.198%	0.00098%
Remaining stocks after PCA II	14										
Stocks eliminated after index analysis	4	0.173%	0.816%	0.224	1.428	0.275%	0.00077%	0.348%	0.00050%	0.248%	0.00122%
Remaining stocks after index analysis	10										
Stocks eliminated after PCA I (risk metric)	10	0.328%	0.825%	0.113	1.374	-		0.258%	0.00033%	0.258%	0.00033%
Remaining stocks after PCA I	21										
Stocks eliminated after PCA II (return metric)	5	0.153%	0.833%	0.096	1.400	-		-		0.232%	0.00035%
Remaining stocks after PCA II	16										
Stocks eliminated after index analysis	6	0.173%	0.845%	0.138	1.699	0.275%	0.00077%	0.348%	0.00050%	0.238%	0.00046%
Remaining stocks after index analysis	10										

Source: portfolio performance, based on statistics performed with SPSS 20.0

Using PCA analysis to get a portfolio of stocks within a diversified risk-based space of stocks, we should run the first PCA analysis based on the beta index. Moreover, if we would like to determine the optimal portfolio structure only based on portfolio return maximization with an accepted level of risk, then the second PCA analysis should be performed as well based on the beta index. This way, the investor can draw up a selection of stocks that lead to the maximum variance in the stocks beta index, while the mathematical portfolio optimization model generates the structure of the portfolio just based on maximizing the portfolio return.

The discussion can continue with the scenario where an investor wants to make a selection of stocks with the highest sample variance in portfolio returns, in order to determine a final portfolio structure only based on minimizing the risk level of the portfolio based on the selected stocks. In such a scenario, it would be recommended to perform PCA analysis in both rounds, while the mathematical portfolio optimization model should reach minimization of the portfolio risk.

In **Table 9** we have run different scenarios on the basis used for stocks metrics variance, in order to check the impact on portfolio return and risk, considering as optimal portfolio structure the same PVMA portfolio optimization used in our first scenario already analyzed. Essential from all that information is the fact that only using PCA analysis based on stocks returns in both rounds, we will get the same portfolio return and risk as in our first scenario, meaning the variance on the return of selected stocks is similar with the variance on the beta of selected stocks. Nevertheless, we want to stress the fact that the PVMA portfolio model analysis is designed to lead to portfolio risk minimization. Under those circumstances, we would appreciate that, once the risk of the portfolio through the PVMA model is minimized, we have to concentrate only on stocks returns variance maximization.

Interesting is the scenario in which we run just a PCA analysis, based on the stocks beta index and afterwards we determine the optimal portfolio structure using the PVMA model. In this case, the portfolio return is *0.348%* (compared to a level of *0.308%* as in the first scenario) and a lower portfolio risk of *0.00050%* (compared to a level of *0.00108%* as in the first scenario). Those results are justified by the fact that the PCA analysis performed based on the stocks beta index has eliminated the problem of collinearity in terms of the stocks risk level. The selection of the stock ensures the elimination of the collinearity problem, while the portfolio optimal structure just minimizes the selected space of stock risks.

Table 10. *Analysis of systemic risk on stocks market returns*

Regression Statistics	
Multiple R	28,13%
R Square	7,91%
Adjusted R Square	6,07%
Standard Error	0,0024
Observations	52

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0,0000	0,0000	4,297	0,043
Residual	50	0,0003	0,0000		
Total	51	0,0003			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0,002	0,000	5,357	0,000
Beta coefficient	-0,003	0,001	-2,073	0,043

Source: portfolio performance, based on statistics performed with SPSS 20.0

Additionally, we can observe in **Table 9**, which involves running a principal components analysis using risk metric as a first basis, will lead to an increase in the risk of portfolio considered, as the beta measure increases after the first PCA analysis from *0.00077%* to *0.00033%*. The second round of PCA analysis, using as a basis the stocks return, determines a slight increase to *0.00035%*, while the final filtering step of the top *10* stocks leads to an increased value of *0.00045%*. According **Table 9**, we observe that the stock selection obtained using risk metric as a basis for PCA analysis has a return of *0.258%*. This portfolio return is higher than the return corresponding to the portfolio obtained by using return metric as the basis for PCA analysis, as the optimization of the diversifying portfolio return leads only to a return of *0.194%*.

Despite a higher risk for portfolio determined running PCA analysis based on return metric (*0.00073%*), compared to the risk for portfolio obtained running PCA analysis based on risk metric (0.00033%), this relation is confirmed even after running two rounds of PCA analysis. Even the lower $R^2=0.281$ in the regression model that analyzes the association between stock returns and corresponding beta coefficient could explain those results, showing that the Romanian capital market provides investors significant space for diversification. As observed in **Table 10**, the association between stocks return and the marginal effect of systemic risk on each stock market return is weak, reflecting a negative relation represented by the regression coefficient of *-0.003*. It seems that systemic risk generates a decrease on stock returns. However, our results in **Table 9** show higher stocks returns, which highlight the fact that the risk of diversification is much higher than the systemic risk on the capital market.

Conclusions

Optimizing portfolio decision involves an optimal capital allocation into a portfolio of financial instruments, assets, or development

projects that ensure high returns to the company at the lowest possible risk level. Indeed, there is a strong positive correlation between the level of the return and the level of the risk. That is why we need to decide first what would be the final aim of our investments strategy, either to reduce the uncertainty level of the investment decision so that we can reach at least a minimum level of investments return, or to maximize the returns of the investments, but on a restricted level of environment uncertainty.

In this study, we have chosen to build a stock-portfolio with the final aim of minimizing the risk portfolio. We found that using data mining techniques clearly provides the investor real support for simplifying the investment decision model. In terms of return and risk of the final portfolio considered, we have shown there is an improvement. In addition, the investment decision model becomes much simpler, because of the option of adjusting it on a regular basis (month, quarter, year etc). Moreover, this model can be transformed into an automated platform where various simulations can be done so that the investor can figure out our uncertainty impact on the investment decision.

The second aim of the study is to show the investors' need of a market reference index that could reflect a dynamic optimal portfolio, extremely useful in the investment decision. Thus, we have observed that the BET index, describing our capital market, is not relevant for an investment decision, maybe only for a macroeconomic purpose. Nevertheless, on a microeconomic level, investors need something more. Using mathematical models and data mining tools and techniques, they can build a list of the most profitable stocks, or less risky stocks, or even a combination of those criteria or others. Multivariate calculus offered through different data mining techniques (principal components analysis in this study) offer a wide range of possibilities for investors/rating companies to develop models that could reflect the market volatility in a simpler framework analysis.

However, this kind of model has limitations. First, it is important that investors understand if they want their portfolios to cover different industries obligatorily, as such a portfolio index cannot be set up to consider such information in its construction. Secondly, it is preferable that the timeframe, considered in building such a portfolio index, should be delimited by specific events that can significantly change the capital markets status. For instance, once the companies disclose financial reports, significant changes on the bid-ask spread for the stocks listed by the corresponding company appear on the market. Another example would be the entrance of a new big player in the market, that can rebalance the stocks' bid-ask spread into other direction. The third reason is the efficiency of the capital market, which we assumed to be valid throughout our study.

Overall, this study can be seen as a first step towards the academic efforts that should be performed in the direction of conceiving a reference index for the investors that has to give relevant clues about the list of the most preferred stocks traded on a capital market. Behavioural finance theory is a solid ground to call into question those multivariate quantitative models. However, the implementation on software of such models can give investors at least a starting point in their technical analysis and smooth interpretations.

References

Alexander, C. (2008). Market risk analysis. Practical financial econometrics, John Wiley & Sons, England

Amani F.A., Fadlalla A.M. (2017), Data mining applications in accounting: A review of the

literature and organizing framework, *International Journal of Accounting Information Systems*, vol. 24, p. 32-58

Bi Z., Cochran D. (2014), Big data analytics with applications, *Journal of Management Analytics*, vol. 1, issue 4, p. 249-265

Cardoso, R.G. (2015). Pair trading: Clustering based on principal component analysis, working paper available at https://run.unl.pt/bitstream/10362/15355/1/Cardoso_2015.pdf

Craighead, S., Klemesrud, B. (2002). Stock selection based on cluster and outlier analysis, working paper available at https://www3.nd.edu/~mtns/papers/110_3.pdf

Fulga, C., Dedu, S. (2012). Mean-risk portfolio optimization with prior PCA-based stock selection, International Workshop "Stochastic Programming for Implementation and Advanced Applications", July 3–6, 2012, Neringa, Lithuania

Fulga, C., Dedu, S., Serban, F. (2009). Portfolio optimization with prior stock selection, *Economic Computation and Economic Cybernetic Studies and Research*, vol. 43, issue 4, pp. 157-171

Gorunescu, F. (2011). Data Mining. Concepts, Models and Techniques, Springer, Berlin

Hassani H., Huang X., Silva E.S. (2018), Digitalization and Big Data Mining in Banking, *Big Data and Cognitive Computing*, vol. 18, issue 2, p. 1-18

Jayasree V., Balan R.V.S. (2013), A review on data mining in banking sector, *American Journal of Applied Sciences*, vol. 10, issue 10, p. 1160-1165

Lemieux, V., Rahmdel, P.S., Walker, R., Wong, B.L.W., Flood, M.D. (2015). Clustering techniques and their effect on portfolio formation and risk analysis, working paper available at https://www.researchgate.net/publication/286056111_Clustering_Techniques_And_their_Effect_on_Portfolio_Formation_and_Risk_Analysis

Liao S.H., Chu P.H., Hsiao P.H. (2012), Data mining techniques and applications – A decade review from 2000 to 2011, *Expert Systems with Applications*, vol. 39, p. 11303-11311

Maheshwari A.K. (2015), Business Intelligence and Data Mining, Business Expert Press, New York

Marvin, K. (2015). Creating diversified portfolios using cluster analysis, working paper available at: https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf

Nobi, A., Lee, J. W. (2016). State and group dynamics of world stock market by principal component analysis, *Physica A: Statistical Mechanics and its Applications*, vol. 450, pp. 85-94

Plotnikova V., Dumas M., Milani F. (2020), Adaptations of data mining methodologies: a systematic literature review, *Peer J Computer Science,* 6:e267

Schuh G., Reinhart G., Prote J.P. Sauermann F., Horsthofer J., Oppolzer F., Knoll D. (2019), Data mining definitions and applications for the management of production complexity, 52nd CIRP Conference on Manufacturing Systems, Procedia CIRP 81 (2019), p. 874-879

Silwattananusarn T., Tuamsuk K. (2012), Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012, *International Journal of Data Mining & Knowledge Management Process*, vol. 2, issue 5, p. 13-24

Yang, L. (2015). An application of principal component analysis to stock portfolio management, master thesis available at https://ir.canterbury.ac.nz/handle/10092/10293.