Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

www.czasopisma.uni.lodz.pl/foe/
1(333) 2018

DOI: http://dx.doi.org/10.18778/0208-6018.333.06

Jerzy Korzeniewski
University of Łódź, Faculty of Economics and Sociology, Department of Statistical Methods, jurkor@wp.pl

Efficient Stock Portfolio Construction by Means of Clustering

Abstract: When investors start to use statistical methods to optimise their stock market investment decisions, one of fundamental problems is constructing a well-diversified portfolio consisting of a moderate number of positions. Among a multitude of methods applied to the task, there is a group based on dividing all companies into a couple of homogeneous groups followed by picking out a representative from each group to create the final portfolio. The division stage does not have to coincide with the sector affiliation of companies. When the division is performed by means of clustering of companies, a vital part of the process is to establish a good number of clusters. The aim of this article is to present a novel technique of portfolio construction based on establishing a number of portfolio positions as well as choosing cluster representatives. The grouping methods used in the clustering process are the classical *k*-means and the PAM (Partitioning Around Medoids) algorithm. The technique is tested on data concerning the 85 biggest companies from the Warsaw Stock Exchange for the years 2011–2016. The results are satisfactory with respect to the overall possibility of creating a clustering-based algorithm requiring almost no intervention on the part of the investor.

Keywords: investment portfolio construction, clustering, number of clusters, Sharpe index

JEL: G20, G29, G35

1. Overview of clustering-based approaches

Cluster analysis is not applied to construct investment portfolios very often. Much more popular are classical mean-variance approaches, value at risk approaches, or even econometric models. However, some authors have tried to follow this idea successfully. The main drawback of classical approaches is their sensitivity to input data, i.e. a small change in variance estimates may cause big changes of the portfolio contents. This drawback can be seriously reduced when one tries to use cluster analysis.

A common feature of clustering-based approaches is that clustering is accompanied by other statistical or financial methods which allow to choose representatives of the clusters of stocks returned by the clustering procedure. Choosing a reasonable number of representatives is absolutely necessary as there cannot be too many positions in an investment portfolio for technical reasons. Craighead and Klemesrud (2002) investigated weekly stock return data series concerning 138 different stocks from the New York SE from the time period 1998–2002. Clustering accounted for roughly a third of the portfolio construction. Apart from clustering, the authors applied the Kalman filtering in order to remove stocks belonging to one-element clusters as well as some other stocks. This approach returned good results because one of the companies removed in this way went bankrupt within the next few months. A weak point of this construction process is associated with the step which consists in using “the investor’s experience” in order to limit the number of 30 cluster representatives to 9. This hardly seems to be a scientific approach. However, general conclusions referring to the usefulness of clustering in portfolio construction were positive. Bensmail and DeGennaro (2004) used an original modelling technique and achieved very good results in clustering companies, however, they propped their analysis with economic variables characterising companies. Marvin (2015) also investigated the NYSE stocks, however, in a completely different way. She applied clustering based on some variables of financial nature, such as revenues divided by assets or income divided by assets. Then, in order to construct a portfolio with a reasonable number of stocks, she used the Sharpe index. The results of this investigation confirm the usefulness of cluster analysis. What is worth noticing is the use of the Sharpe index, which is probably the simplest index referring the investment return rate to its risk. In some research, time series correlation measures were used. This is not clustering in the sense of classical clustering algorithms but the effect is similar, i.e. we arrive at disjoint groups of similar stocks. Ren (2005) used a threshold of 0.2 for the linear correlation coefficient based on which he qualified the stocks as “similar to one another”. This analysis is not very clear, the very grouping procedure can result in all stocks being thrown into the same cluster. Some additional thresholds are necessary. Rosén (2006) also grouped stocks based solely on their mutual correlation coefficients.

All approaches based on correlation coefficients have, as Marvin (2015) pointed out, one serious drawback. The value of correlation coefficients changes quickly with the overall market condition, probably more quickly than the distance between the two stocks in question. For example, very strongly correlated return rates time series of two stocks from the same sector, ExxonMobil and Chevron (the coefficient of linear correlation of daily returns equal to 0.85), in the period from 2000 to 2011, correspond to 2,541 days with the same sign of return rates and different signs on 589 days. What can one expect then from two stocks correlated at the level of 0.2? Cluster analysis can be successfully applied (Pasha, Leong, 2013) even to high-frequency data with the help of some econometric time series modelling.

In all of the aforementioned investigations, the problem of determining the number of clusters, i.e. the number of positions in the portfolio, remains unsolved. This number was usually set arbitrarily.

2. New method proposal and its assessment

In order to propose a method of portfolio construction based on solid statistical foundations, we think that a proper way is to start from establishing the number of clusters or at least determining a starting point of this number. Later, this number can be modified, but our belief is that it should be given by some rule and not ad hoc adjustments. A classical way of determining the proper number of clusters consists in using for this purpose one of many indices. Unfortunately, almost all of such indices have an optimising form, i.e. they suggest the best number of clusters for a fixed grouping method. However, assuming some form of grouping method is necessary anyway because the stocks have to be divided into classes. Therefore, finding the number of clusters is equivalent to choosing a clustering method and an index of the number of clusters. As far as the clustering methods are concerned, we propose to try a couple of methods and then to adhere to the one that proves best. As far as the number of clusters is concerned, we propose to use the Caliński-Harabasz index, which proved to be one of the best in a couple of studies (cf. e.g.: Korzeniewski, 2014). This index uses the formula:

$$CH(k)=\frac{tr(\mathbf{B}_k)/(k-1)}{tr(\mathbf{W}_k)/(n-k)}, \tag{1}$$

where B_k – between group variance matrix, W_k – within group variance matrix (for the formulas, see e.g.: Gatnar, Walesiak, 2004), n – number of data set objects. The optimal number of clusters is k that maximises the value of the expression on the left hand side of formula (1).

Apart from the clustering method and the number of cluster indices, there are a few aspects which need attention if one wants to investigate the efficiency of any investment portfolio. The most important are probably the ones connected with determining the way of investing money. One should do it in such a way that the sophistication of the investing methodology should not obscure the main target, i.e. assessing properly the usefulness of the possible use of cluster analysis. In this respect, we believe that the research carried out by Craighead and Klemesrud (2002) is overburdened with the sophistication of the investing technique. The most logical seems a simple and natural choice of determining one moment of buying the chosen stocks and one of selling them. We set the length of historical observations to be worth up to 120 trading days, which roughly corresponds to half a year. We think this a reasonable choice as we will deal with daily return rates. From the economic perspective, there are little grounds for choosing a longer period. Another problem is picking representatives of the clusters. We adopt probably the simplest possible way applied by Marvin (2015), which consisted in picking from each cluster one stock with the highest Sharpe ratio. Another aspect is the final number of the portfolio positions. Should we stick to the number suggested by the Caliński-Harabasz index or should we change it? With respect to this question, we propose what follows. According to the widely accepted rules, a good investment portfolio should not contain more than a dozen positions. Therefore, we propose to keep the original index suggestion if it does not exceed 12 and to make it twice smaller otherwise. This amounts to looking for a "better half" of all cluster representatives. It can be performed in a number of ways and the ones attempted are given below. The crucial parameter deciding about the length of historical data to be considered should be chosen according to the character of data possessed.

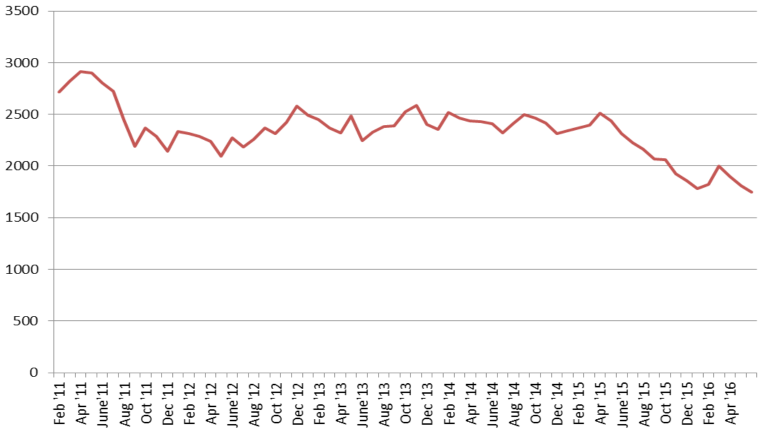

Figure 1. The values of WIG20 index of the Warsaw Stock Exchange from the beginning of 2011 to the middle of 2016

Source: WGPW web site [accesed: 1.11.2016]

We will investigate our proposals based on the Warsaw Stock Exchange data. The values of the main index WIG20 of this stock exchange from February 1, 2011 to April 30, 2016 are presented in Figure 1. We believe that it is quite a long period of time (as for daily returns) and it contains varied periods of market well-being, e.g.: there is a bear market: from May 2011 – 2903 points to November 2011 – 2288 points, a bull market: from July 2012 – 2185 points to January 2013 – 2492 points, and a stable market: from April 2014 – 2439 points to September 2014 – 2463 points. This index comprises 20 biggest companies and well captures the state of the whole market. In the investigation, we take into account the daily return rates of 85 biggest companies, i.e. with the market value exceeding PLN 500 million (as of July 31, 2016). The data set is given by 85 time series, each of the length of 1,360 values. The time series are the objects and the 1,360 values are the variable values. The distances needed for the clustering are the Euclidean distances.

Now, we describe in detail the method of portfolio construction and, accordingly, the way of its assessment.

1. For the given trading days, take the daily return rates of the last 120 trading days to be treated as historical observations. Thus, for the clustering stage, there are 85 objects characterised by 120 variables.
2. Run a given clustering method over the data for the number of clusters predetermined by the Caliński-Harabasz index.
3. Out of each cluster pick out the stock with the highest Sharpe ratio as its representative to create the investment portfolio.
4. If advisable, take a half of all the portfolio positions according to the rule specified.
5. Find the mean return rate of the portfolio for a specified investment time.

For the clustering method, we chose the classical *k*-means and one more sophisticated method, i.e. the PAM algorithm (see Gatnar, Walesiak, 2004). The *k*-means was run with a random choice of starting points repeated 100 times with the result of the smallest within group variance kept as final. The Caliński-Harabasz index was attempted for the range of the number of clusters from 2 to 20.

3. Results and conclusions

The results are presented in Table 1. The investment time (20, 50, 120) is given in trading days. Success and loss are understood in the sense of return rate with respect to the market, e.g.: 16.5% denotes that the market return rate was beaten by 16.5%. Thus, the success ratio is the ratio of positive values of the constructed portfolios return rates. Mean loss and mean success are calculated as the arithmetic means of all the return rates of the companies in the portfolio (or half of them whenever specified so in the first column). The results seem to be interesting because some rules (Sharpe index usage) of clearly financial and economic nature were confirmed in spite of the fact that the portfolio construction method was largely statistical.

Table 1. The results of the investment experiment

Grouping method	Investment time	Mean loss	Mean success	Success ratio
k-means	20	–6.5%	16.5%	65%
	50	–6.9%	9.3%	62%
	120	–6.9%	5.5%	45%
k-means half max Sharpe	20	–8.0%	24.0%	62%
	50	–8.2%	11.6%	57%
	120	–7.3%	8.2%	44%
k-means half max return rate	20	–7.7%	12.3%	63%
	50	–8.4%	11.2%	59%
	120	–8.9%	8.8%	43%
k-means half min risk	20	–7.4%	13.0%	62%
	50	–8.1%	10.6%	59%
	120	–8.8%	8.2%	45%
PAM	20	–11.1%	19.6%	58%
	50	–12.6%	17.3%	59%
	120	–14.1%	14.1%	52%

Source: own investigations

The results with respect to the method of stock grouping are logical: the PAM algorithm, as more sophisticated than *k*-means, returned a much smaller number of clusters (7.5 on average) than *k*-means (17.8 on average). That is the reason why an attempt to look for the "better half" of stocks was not attempted for PAM because 3 or 4 clusters (portfolio positions) seem to be too small (risky) a number. However, the financial result was worse for the PAM algorithm, although the mean success was higher, but mean loss was also higher, with the ratio of success being considerably smaller. This is rather surprising, one can attempt to explain this phenomenon by stressing that choosing a smaller number of representatives of bigger numbers of stocks is more risky (or less efficient) than choosing more representatives of smaller clusters of stocks.

The results with respect to the length of the investment time allow to draw one clear conclusion: it is absolutely inadvisable to try to extrapolate trends from the most recent six months onto the six months which are to follow. The most efficient in this respect were short investments, shorter than one month.

The results with respect to the method of picking out the "better" half of the portfolio are quite surprising. The intuitively good methods, i.e. the half with the highest return rate of the half with the smallest risk, turned out to be worse than the half with the highest Sharpe ratio. This result is similar to Marvin's (2015) results, and the reasonability of the Sharpe measure from the 1960s was confirmed once again.

We think there is no need for comparative analysis with regards to other portfolio construction methods for a couple of reasons. There are thousands of methods applied to the task and, among them, there are no widely accepted reference points. There are probably as many portfolios as there are investors and some of the portfolios might be doing better based on the particular data analysed. However, the results achieved by the best portfolio (24% above the market with 62% of success), we believe, speak for themselves.

We believe that the way of exploiting cluster analysis presented in this article can be upgraded by following the current portfolio behaviour and reacting "on-line". The reason for this belief is the fact that the Sharpe index values show very strong day-to-day autocorrelation which may be used for reading signals for the nearest future.

References

Bensmail H., DeGennaro R. (2004), *Analyzing Imputed Financial Data: A New Approach to Cluster Analysis*, FRB of Atlanta Working Paper no. 2004–20, Atlanta, https://www.econstor.eu/bitstream/10419/100973/1/wp2004–20.pdf [accesed: 1.08.2015].

Craighead S., Klemesrud B. (2002), *Stock Selection Based on Cluster and Outlier Analysis*, Fifteenth International Symposium on Mathematical Theory of Networks and Systems, University of Notre Dame, Notre Dame, Indiana, https://www.researchgate.net/publication/272175812_Stock_Selection_Based_on_Cluster_and_Outlier_Analysis [accesed: 1.08.2015].

Gatnar E., Walesiak M. (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

Korzeniewski J. (2014), *Indeks wyboru liczby skupień w zbiorze danych*, "Przegląd Statystyczny", vol. 61, no. 2, pp. 169–180.

Marvin K. (2015), *Creating Diversified Portfolios Using Cluster Analysis*, unpublished research, pp. 1–15, https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf [accesed: 1.08.2015].

Pasha S., Leong P. (2013), *Cluster Analysis of High-Dimensional High-Frequency Financial Time Series*, IEEE Conference on Computational Intelligence for Financial Engineering & Economics, Piscataway, http://ieeexplore.ieee.org/document/6611700/ [accesed: 1.08.2015].

Ren Z. (2005), *Portfolio Construction Using Clustering Methods*, Thesis at the Worcester Polytechnic Institute, Worcester, https://web.wpi.edu/Pubs/ETD/Available/etd–042605–092010/unrestricted/ZhiweiRen.pdf [accesed: 1.08.2015].

Rosén F. (2006), *Correlation Based Clustering of the Stockholm Stock Exchange*, Master's Thesis, School of Business, Stockholm University, Stockholm, http://www.diva-portal.org/smash/get/diva2:196577/FULLTEXT01.pdf [accesed: 1.08.2015].

Konstrukcja efektywnego portfela przy użyciu metod analizy skupień

Streszczenie: Stosując metody statystyczne do optymalizacji swoich decyzji inwestycyjnych, inwestorzy stają przed bardzo istotnym problemem skonstruowania dobrze zdywersyfikowanego portfela inwestycyjnego składającego się z niewielkiej liczby pozycji. Wśród wielu metod stosowanych do konstrukcji takiego portfela są metody wykorzystujące grupowanie wszystkich spółek w homogeniczne grupy spółek, po którym to etapie następuje wybieranie reprezentanta każdej grupy w celu utworzenia ostatecznej postaci portfela. Etap grupowania nie musi pokrywać się z przynależnością sektorową spółek. Grupowanie może być wykonywane za pomocą metod analizy skupień i w tym procesie bardzo istotne jest ustalanie właściwej liczby skupień. Celem niniejszego artykułu jest zaproponowanie nowej techniki konstrukcji portfela inwestycyjnego, odnoszącej się zarówno do ustalenia liczby pozycji w portfelu, jak również do wyboru reprezentantów skupień. Stosowane metody grupowania spółek to klasyczna metoda *k*-średnich oraz algorytm PAM (*Partitioning Around Medoids*). Technika jest testowana na danych 85 największych spółek giełdowych z parkietu warszawskiego z lat 2011–2016. Wyniki są bardzo obiecujące w sensie możliwości opracowania algorytmu opartego na analizie skupień, który prawie nie wymagałby interwencji inwestora.

Słowa kluczowe: analiza skupień, portfel inwestycyjny, liczba skupień, wskaźnik Sharpa

JEL: G20, G29, G35

Received: 2016-12-28; verified: 2017-06-07. Accepted: 2017-12-28