JMFS
Journal of Management and Financial Sciences

Volume XVI
•
Issue 48
pp. 75–91
SGH Warsaw School of Economics
Collegium of Management and Finance

Łukasz Sroka
University of Economics in Katowice
Doctoral School of the University of Economics in Katowice
ORCID: 0000-0001-5721-2475

Applying the agglomerative method in hierarchical clustering for the medium-sized companies listed on the Warsaw Stock Exchange

Abstract

The purpose of this article is to use a hierarchical algorithm to reduce the number of companies in stock exchange portfolios, together with the identification of the most and least profitable groups of the companies. To prepare the research, the author decided to use a hierarchical clustering method to segment mWIG40 index entities. The conducted research contributed to the knowledge of the segments appearing on mWIG40 index and the profitability of the obtained clusters in the analyzed period. It was concluded that the hierarchical clustering method can divide the entities from mWIG40 index into six segments. The obtained groups differed from each other in terms of the analyzed features. Moreover, it was found that it was possible to identify more and less profitable segments in terms of the rate of return. What is more, only one segment was characterized by a higher rate of return than the benchmark. The findings can help investors to make better decisions during their investing process. In addition, the results can help companies to map their business in the market.

Keywords: hierarchical clustering, segmentation, medium-sized companies, Warsaw Stock Exchange
JEL Classification: C38, G11

Introduction

Business taxonomies are one of the most important and interesting knowledge management tools for investment activities. When investors are comparing different equity assets in the financial markets, they tend to classify companies according to their main business sector, financial performance, and the goods they produce. To discover companies with a high potential to grow across different industries, investors have to analyze different types of source data, such as statements, macroeconomic data, or companies' financial indicators. Having the possibilities of grouping firms according to the most essential financial criteria, investors can indicate the most profitable cluster. It is also the reason why developing of a large number of different business taxonomies is of primary interest of investors and mutual funds [Alford, 1992; Bai at al., 2019]. One of the methods used in business taxonomies is the agglomerative hierarchical clustering. This method occupies a prominent position in the science of classification and for this reason most standard references devote considerable space to its explication and evaluation [Day, Edelsbunner, 1984; Everitt, 1980].

Research shows that this grouping method is often used to build investment portfolios. The paper prepared by Korzeniewski [2017] showed that the hierarchical clustering method may have been competitive with other classic portfolio building methods in the stock market. Very similar conclusions were presented by Craighead and Klemesrud [2002], but in addition to the cluster analysis, the researchers also used the Kalman filtering method to remove some companies from 138 analyzed entities. Leon et al. [2017] presented the performance of seven portfolios created using the clustering method. They constructed a portfolio and measured the performance using the return on assets from a sample of the Russell 100 index. The researchers concluded that the algorithm produces stable results with similar volatility. Lahmiri [2016] adopted hierarchical clustering in order to present different sectors in the Casablanca Stock Exchange market. He observed that the general structure of stock exchange topology was considerably changing over the time periods. In addition to segmentation using financial and macroeconomic data, there were a lot of papers where the researchers focused on using the price movement or the rates of return in order to obtain clusters [Bin, 2020; Esmalifalak, 2015; Zuhroh et al., 2021]. As for the Polish market, beyond Korzeniewski's paper, also Pośpiech [2016] used clustering methods to create segments in the stock market. In that research four economic and financial indicators were used: return on sales, return on assets, return on equity, and profit per share. The research was conducted using companies from the mWIG40 index. It was proved that the clustering method used together with the synthetic measure allows selecting more profitable companies which belong to the same cluster.

As presented in the previous paragraph, there were significant numbers of papers concentrating on clustering analysis; however, there was a lack of research where medium-sized companies from Polish stock markets were segmented and based on the segments, the rates of return of each cluster were checked.

The purpose of this article is to use a hierarchical algorithm to reduce the number of companies in stock exchange portfolios, together with the identification of the most and least profitable groups of companies. The two research hypotheses are as follows: 1. there is a group of listed companies from the mWig40 index that will generate a higher rate of return than the broad WIG index; 2. the financial variables used in the study allow for the preparation of such groups of companies that will be characterized by similar financial parameters within a given group.

Methods

The aim of the research was to use a hierarchical algorithm to reduce the number of companies in stock exchange portfolios, together with the identification of the most and least profitable groups of companies. The research applied the financial data of the companies belonging to the medium-sized companies' index (mWIG40) listed on the Warsaw Stock Exchange (GPW). The mWIG40 index is the successor to the MIDWIG index, which was replaced on March 16, 2007. The index has a fixed number of 40 entities. The companies for the index are selected according to the rankings as the next 40 companies in terms of the ranking criteria after 20 entities in the WIG20 index. Table 1 presents the mWIG40 companies according to their operational segment.

Table 1. Companies from mWIG40 index with their operational sector

Operational sector	Company
Banks	ALIOR, HANDLOWY, INGBSK, MBANK, MILLENNIUM
Construction industry	BUDIMEX, DOMDEV, DEVELIA
Power engineering	ENA, PEP
Finance	KRUK, GPW, XTB
Retail and wholesale trade	EUROCASH, ASBIS, NEUCA
IT	11BIT, ASSECOSEE, COMARCH, DATAWALK, HUUUGE-S144, LIVECHAT, PLAYWAY, TSGAMES
Chemical and electromechanical industry	GRUPAAZOTY, CIECH, AMICA, FAMUR, KETY, INTERCARS
Pharmaceutical industry	BIOMEDLUB, CLNPHARMA, MABION, OAT, SELVITA
Other	AMREST, WIRTUALNA, KERNEL, BENEFIT, PKPCARGO

Source: own work.

To prepare the segmentation, it was necessary to collect the financial information about all the companies from the mWIG40 index. To obtain all the needed data, the latest financial statements of the examined companies were checked. As there are different dates of publication of quarterly reports, the current statements as of November 17, 2021 were selected for the study. Because there was a wide range of choice, 36 financial characteristics were checked to select the variables on which segmentation could be performed. The following 8 variables presented in Table 2. were applied from the data preset.

Table 2. Variables used in companies' segmentation

Variable	Financial type
General debt	Indicator: Debt ratios
Receivable's rotation	Indicator: Activity
Price/Book Value	Indicator: Market value
Liability	Finance: Balance sheet
Sales revenue (sum of the last four quarters)*	Finance: P&L
Depreciation (sum of the last four quarters)	Finance: Cash flow
CAPEX intangible and tangible (sum of the last four quarters)	Finance: Cash flow
Capitalization	General: Rate and turnover

* For banks interest income and commission income were used as the revenue.

Source: own work.

In order to meet the aim of the article and to verify the hypotheses, classification methods such as the agglomerative method in hierarchical clustering analysis was used.

In hierarchical clustering the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to *k* clusters each containing a single object. The hierarchical clustering is subdivided into agglomerative methods, which proceed by series of fusions of the *k* objects into groups, and divisive methods, which separate *k* objects successively into finer groupings.

The agglomerative method is a popular tool used in the segmentation process and is a multidimensional technique that enables grouping multi-feature objects. The main purpose of the grouping is to aggregate objects into homogeneous classes so that objects similar in terms of the considered features are in the same class. Similarity is determined by distance: the shorter the distance, the more similar the objects [Davison, Ravi, 2005; Nitin et al., 2007; Marinova-Boncheva, 2008; Pośpiech, 2016]. The clustering procedure applying the agglomerative method in hierarchical clustering works according to the following scheme:

1. In the distance matrix, one should find a pair of clusters that are the most similar (the least distant from the adopted distance measure). Let assume that these are classes P1 and P2.
2. Next, the number of clusters should be reduced by one by combining P1 with P2.
3. The last step is transforming distances (according to the adopted cluster bonding method) between the combined clusters and the other clusters.

Repeat steps 1–3 until all the objects are in one segment. The result of grouping depends on the method of determining the distance between objects and the adopted method of combining the clusters [Kądziołka, 2018].

In this paper, in order to measure the distance between the selected variables the Euclidean distance with normalized data was applied, while for measuring distances between clusters the Ward method was used as a widely used combination for the hierarchical clustering process [Dokmanic et al., 2015; Kopczewska et al., 2009; Kubiczek, Hadasik, 2021]. The Euclidean distance is represented by the following formula:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2} \tag{1}$$

where: d(x, y) – distance, x_i and y_i are the value vectors of the features of comparable objects in a dimensional space. This kind of distance presents the measure which is simply the shortest distance in n-dimensional space (n-number of examined features) between the objects. To use the Euclidean distance metric, it is necessary to standardize or normalize the variables. To make the variables comparable, the following formula can be applied:

$$z_i=\frac{x_i-\underline{x}}{s_x} \tag{2}$$

where $\underline{x}$ is the mean and s_x is the standard deviation [Walesiak, 2011].

The Ward method belongs to the hierarchical agglomerative methods, in which at every stage of binding the objects or groups of objects are joined with the lowest degree of differentiation [Hair et al., 1998; Ward, 1963]. The measure of this difference is the error sum of squares (ESS), which is defined by the following formula [Stanisz, 2007, p. 122]:

$$ESS=\sum_{i=1}^{k}(x_i-\underline{x})^2 \tag{3}$$

where: x_i is the value of the variables, which is the segmentation criterion the i-th object, k is the number of objects in the cluster.

The Ward method uses the ANOVA approach to estimate the distance between the clusters. The method minimizes the increase in total within the cluster sum of squared error. This increase is proportional to the squared Euclidean distance between cluster centers [Szekely, Rizzo, 2005]. The method is very efficient in terms of creating small-sized clusters. Ward's method does better overall than other hierarchical methods, especially when the cluster proportions are approximately equal [Kuiper, Fisher, 1975; Ferreira, Hitchcock, 2009]. It is widely used in scientific research, mainly for classification purposes [Majerova, Nevima, 2017].

Beyond the methods described in the previous parts of the paper, the principal component analysis (PCA) was also used in the segmentation process, to be precise, in the data preparation step. The PCA is another way of representing the variance among observations in an ordination diagram, which can be seen as a spatial representation of the relationships among the variables [Boschetti, Massaron, 2017]. The PCA is a decomposition of the total variance of the data table, followed by selection of the axes that account for the largest portion of the variance; these axes are then used for representation of the observations in a smaller number of dimensions. From this reasoning, it can be seen that spatial (e.g. PCA) and clustering (e.g. Ward's) methods involve different yet complementary spatial and clustering models that are fit to the data using the same mathematical principle. This is why in practice the results of Ward's agglomerative clustering are likely to delineate clusters that visually correspond to regions of high densities of points in PCA ordination [Murthag, Legendre, 2014]. In this research the

PCA was used to reduce the dimensions of highly correlated financial variables to one variable, thus creating the new variable used later in the grouping process.

To determine the number of specific groups, dendrograms supported by the elbow method were applied. A dendrogram is a graphical illustration of the hierarchical structure of a set of objects grouped due to the decreasing similarity between them [Roman, 2016; Murthag, Legendre, 2014]. The elbow method is a method which looks at the percentage of variance explained as a function of the number of clusters. This method exists upon the idea that one should choose a number of clusters so that adding another cluster does not give much better modelling of the data. The percentage of variance explained by the clusters is plotted against the number of clusters. This method uses the intra-group sum of squared errors (distortions). It allows finding the number of segments for which the intragroup sum of squared errors stops rapidly decreasing, and adding another segment does not introduce much improvement in the distortion [Bholowalia, Kumar, 2014; Raschka, Mirjalili, 2019].

The course of the study

The first step of the analysis was to examine the basic characteristics of the variables distribution. Descriptive statistics were calculated both for the financial indicators, the balance sheet and the profit and loss data. Then the correlation coefficients were calculated for all the analyzed variables to decide what data transformations should be used. It was decided that the financial data would be logarithm and winsorization would be performed. The winsorization (lower and upper 2.5%) was used because of the sensitivity of hierarchical algorithms to outliers.

After the winsorization process the data was standardized and then, using the PCA two new variables were created. The first new variable was the DEVELOPMENT indicator created by combining: Depreciation, CAPEX and Sales revenue. The other new variable was the INDEBTEDNESS consisting of the following variables: General debt and Liability. Beyond the two new variables the following characteristics were applied: Receivable's rotation, Price/Book value indicator, and Capitalization.

The second stage of the analysis was performing the segmentation process. Using the agglomerative method (i.e. Ward's method, the Euclidean distance, logarithm, standardization, and winsorization), the segmentation of mWIG40 companies was prepared. To determine the optimal number of clusters the classification tree visualization and the elbow method graph were used. It was decided to divide the companies into six segments. After the segmentation process the created segments were described statistically to determine the differences between the clusters. Mean, standard deviation, quantile 0.25; 0.50; 0.75 statistics were used to receive information about the segments.

The third stage of the research was to check the rates of return of each segment in the period from 01/10/2020 to 31/09/2021 and to compare the results with a benchmark. As the benchmark the WIG index was used.

Results

The average value of the general debt indicator equals 0.55. The standard deviation of this characteristic is 0.26. This meant that companies from the mWIG40 index have, on average, low credit risk. The average value 14.32 with standard deviation equalling 31.18 of receivable's rotation indicator meant that enterprises in the analyzed index conducted, on average, a very restrictive debt collection policy. The price-to-book value ratio of the current market valuation of a listed company's assets was between 0.23 and 35.09. The average value of the liability was PLN 17,260.86k with the standard deviation equalling PLN 44,198.59k, while the mean value of sales revenue was equal to PLN 1,593.00k with the standard deviation equalling PLN 2,068.01. High variability, apart from liabilities and sales revenue, also occurred for depreciation and CAPEX. The average value for depreciation equalled PLN 208.79k with standard error equalling 332.06, while the average value and standard deviation was equal to PLN 274.48k and PLN 596.25k, respectively. These values meant a great variety of companies in mWIG40. The results were in line with the information provided in Table 1, namely that mWig40 index includes companies from different industries with varying levels of debt, receivables, CAPEX expenses, and depreciation.

The highest coefficient value of variation occurred for liability, while the smallest occured for general debt. Almost for all the analyzed characteristics the skewness was above zero (right-hand asymmetry). The largest asymmetry occurred for receivables rotation. The only one variable with left-hand asymmetry was general debt. In addition to general debt, the rest of the variables were characterized by leptokurtic distribution, which indicates a high concentration of observations around the mean value. The opposite situation occurred for general debt where the distribution was platykurtic. The remaining statistics for the examined variables are presented in Table 3.

Table 3. Descriptive statistics of the financial characteristics

	General debt	Receivable's rotation	Price/Book Value	Liability (in k PLN)	Sales revenue (in k PLN)	Depreciation (in k PLN)	CAPEX (in k PLN)	Capitalization (in k PLN)
Mean	0.55	14.32	5.08	172 60.86	1 593.00	208.78	274.48	4 421.86
Std	0.26	31.18	6.92	44 198.59	2 068.01	332.06	596.25	6 070.66
Min	0.05	0.03	0.23	7.97	0.20	0.23	1.36	494.15
Max	0.92	175.03	35.09	191615	7 029.55	1 576.67	3 115.66	3 4086.20
0.25	0.35	3.93	1.28	2.57	146.32	11.24	18.18	1 524.08
0.50	0.55	5.40	2.40	1 920.40	415.64	81.55	67.25	2 564.47
0.75	0.78	9.71	5.45	6 222.52	2 629.38	213.96	173.94	5 182.25
Coefficient of variation	47.87%	217.82%	136.13%	256.06%	129.75%	159.05%	217.23%	137.29%
Skewness	–0.21	4.14	2.84	3.14	1.32	2.50	3.57	3.60
Kurtosis	–0.96	17.27	8.48	8.93	0.56	6.43	13.01	13.84

Source: own work based on the data from www.biznesradar.pl

In the further part of the analysis, the correlation between the variables was checked. Because the data has the financial character, it was assumed that the Spearman correlation coefficient would be used. For the variables with the high coefficient value the principal component analysis was used to reduce the number of dimensions to one.

The first created variable was the INDEBTEDNESS. The high correlation coefficient between general debt and liability was noticed. Furthermore, both of these variables relate to the company's debt. The results of the correlation are presented in Table 4.

Table 4. Spearman correlations of the INDEBTEDNESS variables

	General debt	Liability (in k PLN)
General debt	1	0.81***
Liability (in k PLN)	0.81***	1

*** All the correlations are significant at the 0.01 level

Source: own work based on data from www.biznesradar.pl

On the basis of the data presented in Table 4, it can be concluded that there was a strong straight positive dependence. After reducing the two-dimensional variables to one dimension using the PCA, the 92% of variability was explained by the new INDEBTEDNESS characteristic.

The second prepared variable was DEVELOPMENT. This variable was created by reducing the dimensions of the following characteristics: Depreciation, CAPEX, and Sales revenue. The variables were also strongly correlated with each other. From financial perspective, it can be assumed that the higher the revenues, the higher the development costs in the form of increasing depreciation and investment expenditure on product development should be noticed. The results of the correlation are presented in Table 5.

Table 5. Spearman correlation of the FINANCIAL variables

	Depreciation	CAPEX	Sales revenue
Depreciation	1	0.93***	0.82***
CAPEX	0.93***	1	0.79***
Sales revenue	0.82***	0.79***	1

*** All the correlations are significant at the 0.01 level

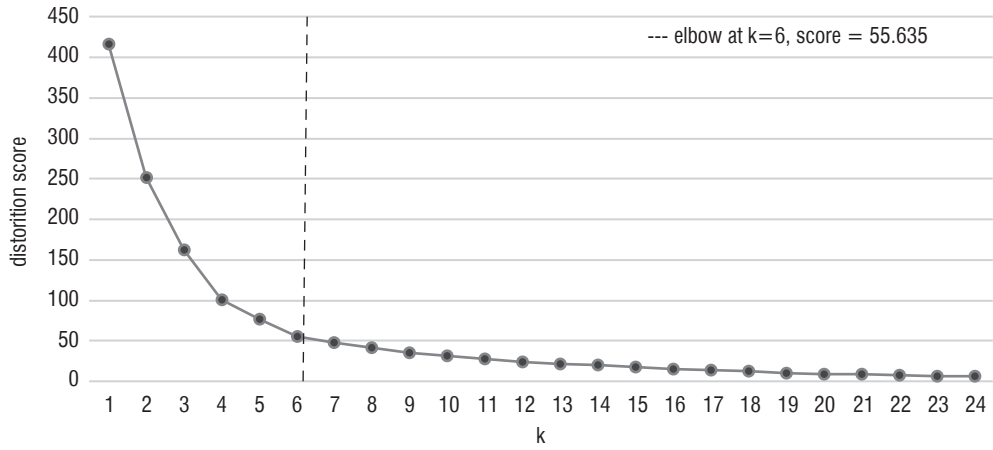
Source: own work based on data from www.biznesradar.pl

For the data presented in Table 5 also a strong positive straight of dependence could be noticed. Reducing the three-dimensional variables to one dimension using the PCA, the 86% of variability was explained by the new DEVELOPMENT characteristic.

The segmentation of the mWIG40 companies was performed using the hierarchical cluster analysis based on the following characteristics: INDEBTEDNESS, DEVELOPMENT, Receivables rotation, Price/Book Value, and Capitalization. Before starting the segmentation process the variables were winsorized (2.5%), logarithmized, and standardized. The numbers

of specified clusters are results from the analysis of the elbow method graph and the classification tree visualization.

Figure 1. The elbow method graph



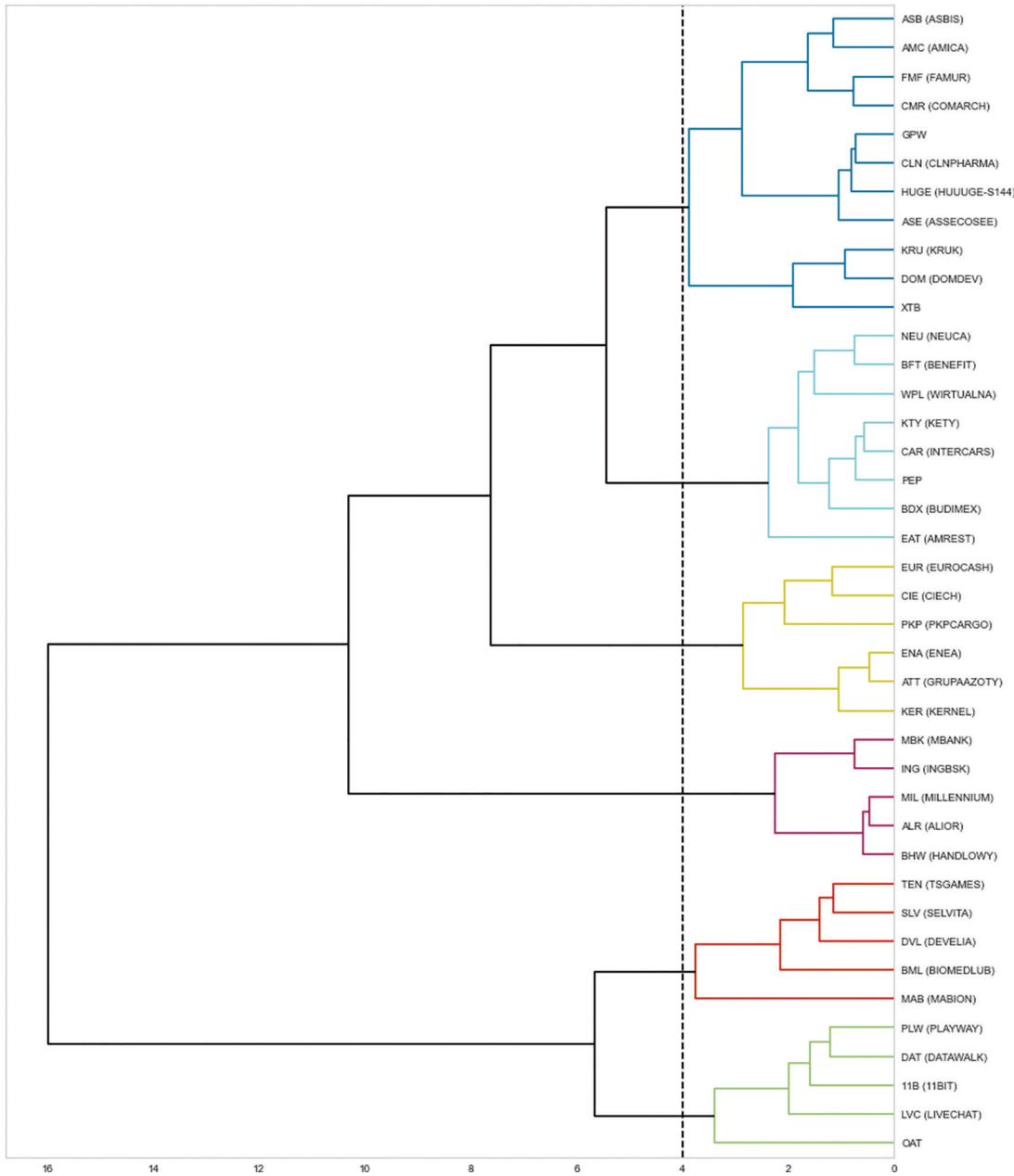
Source: own work.

As presented in the elbow method chart above, it can be noticed that from the k=6 the intragroup sum of squared errors equaled 55.635 and stopped rapidly decreasing, and adding another segment did not introduce much improvement in the distortion. Therefore, it was decided to divide the data into six segments.

As presented in Figure 2, the entities form mWIG40 were divided into six segments. The first segment contained eleven entities (ASBIS, AMICA, FAMUR, COMARCH, GPW, CLNPHARM, HUUGE-S144, ASSECOSSE, KRUK, DOMDEV and XTB), in the second there were eight companies (NEUCA, BENEFIT, WIRTUALNA, KĘTY, INTERCARS, PEP, BUDIMEX and AMREST), in the third segment there were six entities (EUROCASH, CIECH, PKP CARGO, ENEA, GRUPA AZOTY and KERNEL), and the next three segments contained five companies each (the fourth: MBANK, INGSK, MILENIUM, ALIOR and HANDLOWY; the fifth: TSGAMES, SELVITA, DEVELIA, BIOMED LUBLIN and MABION; the sixth: PLAYWAY, DATAWALK, 11BIT, LIVECHAT and OAT).

The next step of the research was to prepare a statistical description of the clusters. Table 6 shows the results of the following statistics: mean, standard deviation, min, max, quantile 0.25, 0.50, 0.75 for each obtained segment.

Figure 2. The visualization of the hierarchical clustering process



Source: own work.

Table 6. Statistics of the segments obtained using the hierarchical method

	General debt	Receivable's rotation	Price/Book Value	Liability (in k PLN)	Sales revenue (in k PLN)	Depreciation (in k PLN)	CAPEX (in k PLN)	Capitalization (in k PLN)
Segment 1								
Mean	0.46	26.16	2.19	1265.50	547.10	52.47	52.99	2265.73
Std.	0.19	50.88	0.96	1013.40	791.29	48.39	35.05	1408.16
Min	0.26	2.19	0.86	171.22	57.67	8.70	5.04	945.47
Max	0.78	175.03	4.08	2857.39	2860.67	175.00	118.00	6165.48
25%	0.30	4.43	1.76	391.65	219.00	17.55	27.13	1618.22
50%	0.42	5.27	2.16	840.52	331.50	37.91	46.13	1830.01
75%	0.60	17.26	2.50	2277.12	415.64	69.44	75.61	2422.01
Segment 2								
Mean	0.55	7.75	15.30	461.27	92.30	7.84	13.74	1388.29
Std.	0.19	7.42	11.38	700.19	96.95	6.96	5.31	755.03
Min	0.35	0.03	8.25	45.02	1.59	1.53	8.66	494.15
Max	0.86	18.53	35.09	1708.50	219.59	19.29	20.95	2568.27
25%	0.43	3.98	8.87	148.23	6.55	3.58	9.75	1116.75
50%	0.55	4.22	9.36	195.59	68.89	5.94	11.73	1295.68
75%	0.56	11.98	14.94	209.02	164.89	8.88	17.59	1466.60
Segment 3								
Mean	0.66	9.12	4.05	3341.49	1603.17	248.49	176.32	4833.13
Std.	0.16	7.19	1.47	2517.32	1157.65	346.71	160.44	1604.55
Min	0.43	4.64	2.21	510.10	225.10	74.87	59.59	2123.88
Max	0.85	26.48	6.56	8140.01	3170.32	1099.82	465.08	6577.84
25%	0.52	5.46	3.25	1603.76	505.83	90.27	68.89	3745.77
50%	0.66	6.90	3.94	2922.87	1766.22	138.58	92.27	5142.87
75%	0.80	8.26	4.90	3907.57	2492.50	166.22	232.76	6073.39
Segment 4								
Mean	0.62	8.45	0.81	8714.26	3971.93	741.55	1276.83	2793.51
Std.	0.14	4.30	0.75	5109.23	2595.22	440.85	1104.60	1637.36
Min	0.48	4.99	0.23	3791.45	834.71	345.05	205.91	698.68
Max	0.90	16.96	2.26	17111.94	7029.55	1576.68	3115.66	4941.04
25%	0.56	6.49	0.36	5127.95	1758.03	476.68	700.85	1689.35
50%	0.59	7.13	0.56	7306.41	4318.94	655.28	762.51	2666.54
75%	0.63	7.93	0.84	11102.25	5898.51	763.72	1791.20	3985.13
Segment 5								
Mean	0.10	24.43	11.77	21.10	25.90	4.54	12.49	1821.49
Std.	0.03	43.74	8.60	15.98	25.03	5.21	9.81	1210.90
Min	0.05	2.90	3.89	7.97	0.20	0.24	1.36	552.45
Max	0.13	102.56	25.28	48.25	56.35	10.45	27.43	3373.25
25%	0.09	3.78	6.36	10.87	10.62	0.63	6.72	1172.65
50%	0.11	3.98	8.36	19.18	13.42	1.38	12.26	1180.35
75%	0.12	8.95	14.94	19.23	48.91	10.00	14.69	2828.76

cont. Table 4

	General debt	Receivable's rotation	Price/Book Value	Liability (in k PLN)	Sales revenue (in k PLN)	Depreciation (in k PLN)	CAPEX (in k PLN)	Capitalization (in k PLN)
Segment 6								
Mean	0.91	0.06	1.32	119016.91	4097.31	254.99	238.72	16095.24
Std.	0.02	0.01	0.37	64282.05	2033.73	120.29	243.26	11707.90
Min	0.88	0.05	1.02	52952.44	1394.20	111.95	92.63	7402.41
Max	0.92	0.08	1.95	191615.63	6508.00	435.97	670.81	34086.20
25%	0.91	0.06	1.12	71601.28	3434.77	202.33	119.37	7643.59
50%	0.92	0.06	1.19	95812.72	3449.85	232.50	139.20	9456.25
75%	0.92	0.07	1.32	183102.50	5699.75	292.20	171.57	21887.75

* in k PLN

Source: own work.

The first segment was described as one of the lowest debt ratios and the highest receivables turnover ratio. This means that companies from this segment were characterized by low overall debt and long crediting to their recipients. The market value ratio in this segment, calculated using the price-to-book value ratio, is one of the lowest. These enterprises were also characterized by one of the lowest average liabilities value and sales revenues Moreover, the average value of depreciation and CAPEX was lower compared to other segments. Such a result may indicate a lower possibility of generating income in the future due to the reduced amount of the investment.

The second segment was characterized by the highest average value of the price-to-book value ratio with, at the same time, one of the lowest average sales revenues. Similarly to the first segment, the average value of CAPEX and depreciation is at a low level. Moreover, the market capitalization is the lowest in comparison with the other clusters.

The entities from the third segment were characterized by one of the highest average capitalizations. The remaining features are at an average level in comparison to the other groups.

The companies from the fourth segment, unlike the first and the second cluster, were characterized by the high average values of expenditure on depreciation and CAPEX. Such levels of the variables may have indicated that the companies placed a strong emphasis on further development or had a significant number of fixed assets. In the long term, it is possible to increase sales revenues by companies from this cluster thanks to the investments and product development.

The fifth segment consisted of four IT companies and one pharmaceutical company. It is worth noting that most of the examined features were characterized by the lowest average values. This is due to the business profiles of the entities in this segment. On the other hand, this segment had one of the highest receivables turnover rates.

The last, sixth segment includes only banks. A very high average value of liabilities and the general debt ratio were noticed for the companies from this cluster. Moreover, this segment included the entities with a very low average value of the receivables turnover ratio. The

average value of sales revenues (in this case, revenues related, inter alia, to the bank's lending and interest activities) and capitalization were at the highest level among all the segments.

The last stage of the research was to check the rates of return of each of the segments and to compare the results with a benchmark – WIG index (broad market index). The analysis was prepared using daily logarithmic rates of return for the period from 01/10/2020 to 30/09/2021 (250 stock quotes). The results are presented in Table 7.

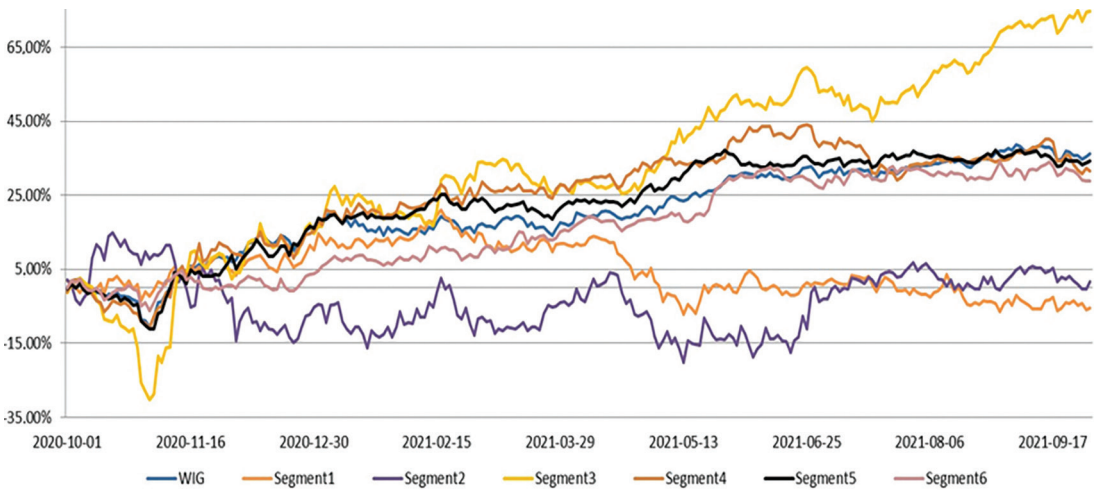
Table 7. Rate of return in portfolio analysis

	WIG	Segment1	Segment2	Segment3	Segment4	Segment5	Segment6
Total rate of return	36.08%	–5.52%	1.61%	74.63%	31.48%	34.12%	28.84%
Average daily rate of return	0.14%	–0.02%	0.01%	0.30%	0.13%	0.14%	0.12%
Std. of daily rate of return	1.13%	1.57%	2.64%	2.37%	1.54%	1.14%	1.13%
Maximum daily rate of return	4.30%	4.36%	9.76%	13.56%	5.68%	4.18%	3.94%
Minimum daily rate of return	–4.75%	–4.41%	–11.55%	–9.39%	–5.15%	–4.45%	–4.20%

Source: own work.

As shown in Table 7, the third segment (segment with banks only) was characterized by the highest overall rate of return. This means that by investing in this cluster throughout the analyzed period, the investor could earn 74.63%. This result was also higher than the benchmark. The rest of the segments had much worse rates of return and did not perform better than the WIG index. It was noticed that the first segment presented a negative total rate of return. The highest daily volatility was in the second and the third segment. It is worth noting that the lowest daily rate of return was in the second cluster and the highest in the third cluster. Figure 3 shows the chart of the daily logarithmic rate of return for each cluster compared with the benchmark.

Figure 3. The logarithmic rates of return for the analyzed segments and the benchmark



Source: own work.

Discussion

The results of the research confirm that they are consistent with the results of other scientists. As in Korzeniowski [2017], it has been proven that the hierarchical method allows for the division of companies in such a way that it is possible to create stock exchange portfolios from them. These portfolios are characterized by a variable level of return and can be used by investors with a different approach to investment.

If investors invested in Segment 3, they could earn approximately 75% in the analyzed period, however, when the investor prefers short selling, he/she could sell a bench of companies from Segment 1 and as a result he/she would receive about 6% of profit. Therefore, it can be concluded that this research could help make decisions for each kind of investors. If investors do not like to take the risk, the companies from Segment 5 should be bought because on average these industries are characterized by the lowest daily standard deviation. On the other hand, Segment 2 can be classified as the one with the highest deviation.

The conducted research also confirmed the results presented by Pośpiech [2016]. The use of financial and economic variables allows for the preparation of cluster analysis on the Polish stock exchange.

When it comes to comparing the results of the created segments to the results of the entire broad market in the form of the WIG index, in contrast to the results presented by Leon et al. [2017], the groups created using the hierarchical method give different results compared to the benchmark. Only segment 3, consisting solely of banks, allowed obtaining rates of return higher than the WIG index. The rates of return of segments 4 and 5 were at a similar level as the WIG index, while the remaining groups performed clearly worse than the benchmark.

Summary

The purpose of this article was to use a hierarchical algorithm to reduce the number of companies in stock exchange portfolios, together with the identification of the most and least profitable groups of companies. The two research hypotheses were as follows: 1. there is a group of listed companies from the mWig40 index that will generate a higher rate of return than the broad WIG index; 2. financial variables used in the study allow for the preparation of such groups of companies that will be characterized by similar financial parameters within a given group.

Using the agglomerative hierarchical clustering method, the companies from the mWIG40 index were divided into six segments. Importantly, the segments differed from each other in terms of the analyzed features, thanks to which it was possible to designate a segment with high average capitalization, debt, or sales revenues. The information on the differences in sectors is important from the point of view of investors who are looking for methods to optimize their

investments on the stock exchange and reduce the number of the companies in their portfolio. What is more, after analyzing the rates of companies' returns in given sectors, it was possible to select groups of industries whose historical rates of return were above the benchmark and those segments where the rates of return were lower, or negative in case of one sector.

The research showed that an investor does not have to purchase the entire WIG broad index to generate satisfactory rates of return. It is enough that, using the segmentation strategy proposed in this study, they will focus on purchasing only selected companies from a given group of companies. This means that the presented method allows reducing the number of companies in the portfolio, which is associated with lowering the investor's costs.

Moreover, the research hypotheses provided at the beginning of the article were confirmed. There is a group of companies listed in the mWIG40 index, which gives higher rates of return than the WIG. Such a group of companies are entities from segment 3. Also, the financial variables used in the study allow preparing of such groups of companies that were characterized by similar financial parameters within a given group (similar level of debt, price-to-book value, capitalization, CAPEX, values, and liabilities).

However, some limitations of hierarchical clustering should be noted. This method is very sensitive to outliers, therefore, in order to segment in such a way that the groups do not contain only one observation, either the data should be winsorized or the outliers should be excluded from the dataset. In addition, this method works only with numerical values, which have to be standardized. If there are categorical observations in the dataset, the researcher has to map them appropriately using the dummy 0–1 values before applying the method.

Continuation of the research can follow using different clustering methods in segmentation of the companies from the mWIG40 index or other variables and comparing results given by both analyses. Two stage clustering methods can be selected for the companies for the most numerous Segment 1. As presented in Figure 2, the other division of the second cluster may give different points of view in case of most and least profitable companies in the data set.

References

1. Alford, W.A. (1992). The effect of the set of comparable firms on the accuracy of price earnings valuation method. *Journal of Accounting Research*, 30(1), pp. 94–108.
2. Bai, H., Xing, Z.F., Cambria, E., Huang, W. (2019). Business Taxonomy Construction Using Concept-Level Hierarchical Clustering. arXiv preprint arXiv:1906.09694.
3. Bholowalia, P., Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 105(9), pp. 17–24.
4. Bin, S. (2020). K-Means Stock Clustering Analysis Based on Historical Price Movements and Financial Ratios Scholarship. Retrieved from: claremont://scholarship.claremont.edu/cmc_theses/2435/ [accessed: 1.12.2021].
5. Boschetti, A., Massaron, L. (2016). *Python Data Science Essentials*. 2nd Edition. Gliwice: Helion.

6. 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, pp. 12–16.
7. Davison, I., Ravi, S.S. (2005) *Agglomerative Hierarchical Clustering with Constraints*: *Theoretical and Empirical Results*, Proc. 15th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005), Porto, Portugal, pp. 59–70.
8. Day, W.H., Edelsbrunner, H. (1984). Efficient Algorithms for Agglomerative Hierarchical Clustering Methods. *Journal of Classification*, Vol. 1, pp.1–24.
9. Dokmanic, I., Parhizkar, R., Ranieri, J., Vetterli, M. (2015). Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6), pp. 12–30.
10. Esmalifalak, H., Ajirlou, A.I., Behrouz, S.P., Esmalifalak, M. (2015). (Dis) integration levels across global stock markets: A multidimensional scaling and cluster analysis. *Expert Systems with Applications*, 42(22), pp. 8393–8402.
11. Everitt, B. (1980). Cluster analysis. *Quality & Quantity: International Journal of Methodology*, 14(1), pp. 75–100.
12. Ferreira, L., Hitchcock, D.B. (2009). A Comparison of Hierarchical Methods for Clustering Functional Data. *Communications in Statistics – Simulation and Computation*, 38(9), pp. 1925–1949.
13. Hair, J., Anderson, R., Tatham, R. Black, W. (1998). *Multivariate data analysis*. 5th Edition, New Jersey: Prentice Hall.
14. Kądziołka, K. (2018). Zastosowanie metod grupowania hierarchicznego w strategiach portfelowych. *Zeszyty Naukowe Zachodniopomorskiej Szkoły Biznesu Firma i Rynek*, 1(53), pp. 115–124.
15. Kopczewska, K., Kopczewski, T., Wójcik, P. (2009). *Metody ilościowe w R, Aplikacje ekonomiczne i finansowe*. 2nd Edition. Warszawa: CeDeWu.
16. Korzeniewski, J. (2017). *Zastosowanie analizy skupień do konstrukcji portfela akcji na GPW*. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 468, pp. 108–115.
17. Kubiczek, J., Hadasik, B. (2020). Segmentation of the electric scooter market in Poland. Econometrics. *Ekonometria. Advances in Applied Data Analytics*, 24(4), pp.50–65.
18. Kuiper, F.K., Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics*, 31, pp. 777–783.
19. Lahmiri, S. (2016). Clustering of Casablanca stock market based on hurts exponent estimates. *Physica A: Statistical Mechanics and its Applications*, 456, pp. 310–318.
20. Leon, D., Araggn, A., Sandoval, J., Hernnnde, G. (2017). Clustering Algorithms for Risk – Adjusted Portfolio Construction. *Procedia Computer Science*, 108, pp. 1334–1343.
21. Majerova, I., Nevima, J. (2017). The measurement of human development using the Ward method of cluster analysis. *The Journal of international studies*, 10(2), pp. 239–257.
22. Marinova-Boncheva, V. (2008). Using the Agglomerative Method of Hierarchical Clustering as a Data Mining Tool in Capital Market. *International Journal Information Theories & Applications*,15, pp.382–386.
23. Nitin, G., Patel, R., Bruce, P.C. (2007). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. 2nd Edition. New Jersey: John Wiley & Sons.

24. Pośpiech. E. (2016). *Comparative analysis of selected clustering methods of listed companies*. Zeszysty Naukowe Uniwersytetu Ekonomicznego w Katowicach, 297, pp. 153–165.
25. Raschka, S., Mirjalili, V. (2017). *Python mashine learning*. 2nd Edition. Gliwice: Helion. Roman, W. (2016). Zastosowanie hierarchicznej metody aglomeracyjnej do grupowania państw OECD ze względu na efektywność wykorzystania energii. *Roczniki Kolegium Analiz Ekonomicznych*, 40, pp. 411–424.
26. Stanisz, A. (2007). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe*. Kraków: Statsoft.
27. Szekely, G.J., Rizzo, M.L. (2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification*, 22, pp. 151–183.
28. Walesiak, M. (2011). *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*. Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.
29. Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), pp. 236–244.
30. Zuhroh, I., Rofik, M., Echchabi, A. (2021). Banking stock price movement and Macroeconomic indicators: k-means clustering approach. *Cogent Business and Manaement*, 8(1).