*Article*

# Cluster Analysis of Financial Strategies of Companies

**Sergey Dzuba * and Denis Krylov**

FEFU Campus, Far Eastern Federal University, 10 Ajax Bay, Russky Island, 690922 Vladivostok, Russia;
krylov.dv@students.dvfu.ru
* Correspondence: dziuba.sa@dvfu.ru

**Abstract:** Measuring the value of companies and assessing their risk often relies on econometric methods that consider companies as a set of objects under study, homogeneous in the sense of their use of financial strategies. This paper shows that cluster analysis methods can divide companies into classes according to financial strategies that they employ. This indicates that homogeneity can be considered within these classes, while between-class companies should rather be perceived as heterogeneous. The clustering of companies has to be performed on quite a dense set of strategies, which requires a combination of formal and heuristic methods. To divide companies into classes, we used financial coefficients characterizing strategies for the 2030 largest non-financial companies within the time period from 2006 to 2018. As a result, a stable division into seven clusters/strategies was obtained. We revealed that some strategies were more characteristic for the companies of high-tech economy, while others were typical for the companies in basic industries. The dynamics of clusters is characterized by an increase in the share of risky strategies. A good meaningful interpretation of the resulting clustering confirms its consistency. The identified clusters can be used as dummy variables in econometric studies of companies to improve the quality of the results.

**Keywords:** cluster analysis; pattern analysis; cluster silhouette; financial strategy of the firm

## 1. Introduction

Attempting to increase their economic returns, companies can use a wide range of tools. In addition to long-term operational efficiency as a fundamental basis for cash flow, these tools may include advanced financial strategies that positively affect investors' expectations and thus the company's capitalization, such as active participation in mergers and acquisitions or share buybacks. Such strategies are closely related to financial risks, such as goodwill impairment, increased financial leverage, etc. Often, such risks are long-term in the sense that a company can stay in a risky position for quite a long time without direct negative consequences.

Models based on the economic value of a company can be tentatively divided into two classes: risk assessment and company value measurement.

Measuring the value of the company pursues an econometric assessment of the future value of the company from the perspective of stakeholders, based on a wide set of factors, including both the results of current activities, and the data characterizing the intellectual and social capital of companies. I.V. Ivashkovskaya's monograph [1] gives a comprehensive overview of such models. At the same time, a number of typical difficulties arise, the main of which are:

1.  The researcher needs to combine high-frequency market indicators and low-frequency fundamental ratios in the model, which complicates even static models.
2.  The researcher needs taking into account for lags between regressors and explained variables.

Risk research models have a narrower focus on estimating the company bankruptcy risk and originate from the well-known publication by Edward Altman [2], who presented a factor model called Altman's Z-coefficient, which was later developed by James Olsen [3]

and became known as the O-coefficient. The attention that those models received from econometricians led to the factorial models being transformed into more modern logit models (for example, [4]). Generally, such transformation does not change the meaning of the problem as a bankruptcy risk assessment expressed through indicators of the current state of the company. Information about the future state of the company is not used because the predicted negative events are determined by decisions that have already been implemented, rather than by planned decisions. Of course, additional information, for example, about the schedule of upcoming payments on liabilities can serve to increase the accuracy of estimates and increase the prediction horizon [5], but it is also directly determined by the current state of the company. Within the already established methodology, a more recent stream of work is aimed at obtaining estimates in local markets of individual countries and regions [6,7].

The value measurement models are characterized by the use of estimates for future cash flows and human capital, while the risk assessment is based on the structure of capital with an emphasis on debt load indicators. In general, both classes of models are characterized by the use of indicators of current efficiency, including market value. At the same time, cluster analysis is a rare research tool for such or similar problems, which makes it difficult to directly compare the results obtained in this paper. This can be explained by the fact that cluster analysis does not allow us to obtain econometric estimates but is designed to group objects in the space of selected indicators, as, for example in [8].

In the problem under consideration, special features of cluster analysis will be useful for classifying financial strategies of companies. This would probably be unnecessary if financial strategies were reduced only to the choice of capital structure. In that case, clustering would also be difficult from the technical point of view since efficiency and capital structure as model variables are expected to be dense unimodal distributions without explicit structure. In our paper, however, we want to show that the real picture is much more complex.

Let us consider two hypothetical companies of the same size generating the same cash flow. In this case, one of them creates production facilities using borrowed capital, while the other company used the borrowed capital to acquire another company (M&A deal) with the formation of goodwill. Then, depending on the conditions of the M&A deal, they may have different or similar capital structure. With the same structure, they would formally have the same level of risk in the econometric model, which is incorrect because it does not take into account goodwill depreciation risks. A more correct approach would be, for example, to separately compare (and evaluate) companies with and without goodwill and, accordingly, to develop different criteria for assessing their value or bankruptcy risk, i.e., to use the goodwill as a dummy variable. It is clear that in this case goodwill means not only any positive value on the balance sheet, but a value above a certain threshold, the correct determination of which requires methods of cluster analysis.

On this basis, the aim of this paper is to identify the financial strategy of a company as a stable classification indicator, which can be assessed using open data, such as corporate financial statements that allow for specification of the "markers" of their financial strategies using qualitative verifiable data. At the same time, clustering should be performed in such a way that the clusters were associated with different strategies, rather than some variations of the same strategy.

The resulting clustering proves useful not only as a tool for the specification of dummy variables, but also, first of all, as an independent economic study. In the Discussion, we show that clusters specify different strategic behavior of companies, which is stable in dynamics (companies rarely "migrate" between clusters) and has an industry structure. This means that strategies may be dictated not only by companies' preferences, but also by the technological characteristics of business or the competitive situation in the industry.

The key result of this research is justification of the heterogeneity of companies, which is caused by their own strategic preferences and by the constraints imposed by industry specialization. This heterogeneity should be taken into account in econometric studies.

In our opinion, this phenomenon has not received enough attention, since samples of companies are implicitly considered as some homogeneous sets. Another element of novelty is that the variables specifying strategies include not only the classical indicators characterizing companies' efficiency, asset structure, and debt burden, but also financial "markers" of M&A and share buyback activity. Such markers used to be associated with tactical-level decisions and have only become significant financial strategies in the last decade or so.

This paper is structured as follows. Section 2 introduces selected unique features of financial strategies used in clustering. In Section 3, we describe corporate reporting dataset. Section 4 gives insight into general problems of cluster analysis and our approaches to solving them. In Section 5, we present clustering results. In Section 6, we discuss the quality and interpretation of clusters. Finally, Section 7 presents our conclusions.

All computations in this paper were performed using the R with multiple additional packages [9]. The source code and the materials not included in the work are available at the Figshare platform at https://doi.org/10.6084/m9.figshare.12967976 (accessed on 8 November 2021).

## 2. Model Specification

The initial data for this study are the open corporate reporting of the world's largest companies in the non-financial sector. They allow us to calculate a large set of financial ratios of profitability, efficiency, liquidity, solvency and risk. It would be appropriate to use it as broadly as it possible if we were solving the task of predicting bankruptcy or falling (or rising) performance of firms using regression analysis methods. The difficulties associated with excessive dimensionality or the threats of multicollinearity are not considerable compared to the advantages of target breadth. However, if we want to obtain interpretable results in cluster analysis, we should focus on as few ratios preferably uncorrelated as is possible.

The list of six used financial indicators (see Table 1) can be attributed to the fundamental characteristics of activity, reflecting aspects of the effectiveness of companies and the risk of their financial strategies. First, we have deliberately rejected indicators characterizing future cash flows or human capital for the reasons of data availability from the companies' public corporate reports.

**Table 1.** The financial indicators used.

| Indicator | Description | Formula | Indicator's Objective |
|-----------|-------------|---------|----------------------|
| P | EBIT on Sales | $\frac{OI}{S}$ | Operational efficiency |
| R | Return on Capital Employed, ROCE | $\frac{NI+IE}{D+E}$ | Efficiency of invested capital |
| G | Goodwill Ratio | $\frac{GG}{GG+E}$ | Goodwill lost risk |
| T | Treasury Stock Ratio | $\frac{-TS}{E-TS}$ | Depreciation of own shares risk |
| F | Financial Assets Ratio | $1 - \frac{FA+GG}{NCA}$ | Risk of financial assets lost |
| L | Leverage | $\frac{D}{D+E}$ | Risk of borrowed funds lost |

Designations: *OI*—operation income, aka earnings before interest and taxes (EBIT); *NI*—net income; *S*—sales; *IE*—interest expenses; *D*—debt, long- and short-term; *E*—equity; *GG*—goodwill; *FA*—property, plant, equipment or gross fixed assets; *TS*—treasury stock; *NCA*—non-current assets.

To characterize the financial performance of the firm, we chose operating profitability P and return on invested capital R. Ratio P appears to be return on sales, but instead of net income, we chose operational income as the basis of ratio in order to cut off the financial part in total income. Therefore, the companies with massive financial assets and financial revenues (see F indicator of dataset) cannot appear to be profitable if its operational performance is poor. As for R indicator, we have preferred the return on invested capital ratio (ROIC) to return on equity ratio (ROE) because of second one can

be leveraged as opposed to first one (see L indicator of dataset). Both ratios specify profit maximization strategy of the company.

Indicators L, G, F, and T act as markers of financial strategies and reflect the risks for companies associated with their implementation. A high value of the indicator corresponds to a high risk. Let us characterize these strategies.

G. Mergers and acquisitions (M&A) activity will manifest itself in an increased amount of goodwill in the asset structure, especially after the 2001 "revolution" in its accounting [10]. The reason for these changes was the formation of large amounts of goodwill due to the systematic market risk premium over the fair value of the company to be acquired. From this point of view, the market bubble intervenes in the fundamental financial performance of companies, where goodwill is essentially a fictitious asset [11], through M&A. Unsuccessful acquisitions or internal management problems can lead to goodwill impairment. For example, a biggest one-time impairment can be seen in Time Warner's 2002 capital loss write-off of about USD 100 billion. The G-indicator reflects the magnitude of the threat to capital in such events.

T. A share buyback strategy allows a company to replace expensive equity in its capital structure with cheaper debt [12]. This strategy is resorted to by companies that see no better investment for free cash flow [13] than investing in their own business, and not through assets (e.g., expansion of production facilities) but through liabilities, on the financial side—through equity participation. In this case, the risks are based on the fact that the buyout is carried out at market prices of capital, which may be many times higher than its book value, as reflected in the financial statements. If the company loses profitability, its problems may be worsened by the loss of capital due to the company's inability to sell the shares at the previously repurchased price. The formula for calculating the T-indicator implies that, according to accounting standards, the cost of repurchased shares is reflected with a minus sign in the capital structure.

F. A massive use of financial assets by industrial companies can also be considered as one of the strategies. Ford and General Motors, whose balance sheets before the 2008 crisis had many times more financial assets than industrial ones, are (were) the striking representatives. Accordingly, financial activity often brought more profit than the main business. The financial assets depreciated two years prior to the crisis, resulting in a total loss of capital for the companies. At the same time, having large financial assets is generally typical for the industrial companies that offer their products to customers on credit provided by their own financial units. Therefore, a high proportion of financial assets may be a business-specific feature, but it may also reflect a speculative activity of companies. The share of financial assets is calculated without taking into account goodwill (a type of intangible asset), such that its value could not affect the G and F indicators simultaneously.

L. The above strategies require the use of significant financial resources, which are attracted, as a rule, from borrowed sources. This entails the solvency risk. Usually, the value of financial leverage is calculated as $L = D/E$, but due to the possible $E \leq 0$ at the always non-negative $D$, the formula $L = D/(D + E)$ is applied, which helps avoid division by zero and, therefore, potential data outliers. The same considerations are present in the calculation of the index G.

Before completing the model, we turned to a broad set of ratios to check how they correlated with our indicators and to make sure we did not miss anything significant. Appendix A (Tables A1 and A2) presents a correlation matrix between our set and the standard calculated ratios of profitability, efficiency, liquidity and solvency. Our P and R indicators are correlated with profitability ratios, as it should be. Our solvency indicator L does not correlate with conventional solvency ratios as strongly as it should be due to specific our L formula. It also has a negative correlation with liquidity ratios. Therefore, it specifies not only long-term solvency risks, but also indirect liquidity. Both efficiency ratios are practically non-correlated. However, these are not financial strategy indicators. Amount of fixed assets or working capital usually depends on industry specifics of company (for example, large fixed assets for heavy industry and much lower ones for hi-tech) but not

on company's financial strategy. Opposite of it, goodwill and financial assets are financial strategy indicators and therefore we include there as F and G parameters.

Thus, our set of indicators reflects markers of companies' financial strategies. Some of them characterize the level of efficiency (P and R), debt burden (L), and asset structure (F), as is performed in econometric models for bankruptcy risk assessment. However, such variables as T and G are usually not taken into account in these models. This can partly be explained by the fact that until recently they have been used as tactical "maneuvers", rather than financial strategies. As we will see below, the inclusion of variables T and G in the model contributed greatly to the successful solution of the problem stated in this paper.

All the indicators that we use can be found in their public financial statements. Of particular interest is the inclusion in the dataset not only of the crisis year of 2009, but also the period of optimism and prosperity before it and the relevant data thereafter, when it was thought that the crisis had already been overcome.

## 3. Data

We used the publicly available corporate financial statements of the world's 2030 largest companies in the non-financial sector as the source data for this study. The companies have been listed in the Forbes Global 2000 ranking over the past decade. The observation period is 13 years, from 2006 to 2018 inclusive. This provides us with 26,390 observations. However, if we account for incomplete data (for example, because the company was acquired during the period, or conversely, recently entered the market) and emissions over six standard deviations, we obtain the number of 23,479 observations.

Table 2 shows the correlation matrix. It demonstrates that the indicators are mostly independent of each other (except for the expected dependence of P and R) reflecting a good specification of the model. Some of the dependencies will be useful for our analysis of the resulting clustering:

**Table 2.** Correlation matrix of indicators.

|   | **P** | **R** | **L** | **G** | **F** | **T** |
|---|---|---|---|---|---|---|
| P |   |   |   |   |   |   |
| R | 0.45 *** |   |   |   |   |   |
| L | −0.31 *** | −0.14 *** |   |   |   |   |
| G | −0.01 | −0.05 *** | 0.35 *** |   |   |   |
| F | −0.03 *** | −0.05 *** | −0.01 | −0.03 *** |   |   |
| T | 0.14 *** | 0.03 *** | 0.13 *** | 0.31 *** | 0.01 * |   |

NB. *, ***—significance at the level of 10%, 1%, respectively.

L and G indicate that the purchase of other companies with the formation of goodwill is often carried out at the expense of borrowed funds.

T and G indicate of the presence of companies specifically investing in their own business.

L and R. Formally, there should not be such a correlation since the R indicator (ROCE) should eliminate the financial leverage effect by its construction.

L and P. Constructively, there should not be such a correlation either, as the financial costs associated with L are not included in the calculation of P; most likely, the negative correlation indicates that companies with operational issues often try to compensate for them with borrowed funds.

The descriptive statistics of the dataset (see Figure 1 and Table 3) also indicate the presence of data features that should be taken into account when performing cluster partitioning.
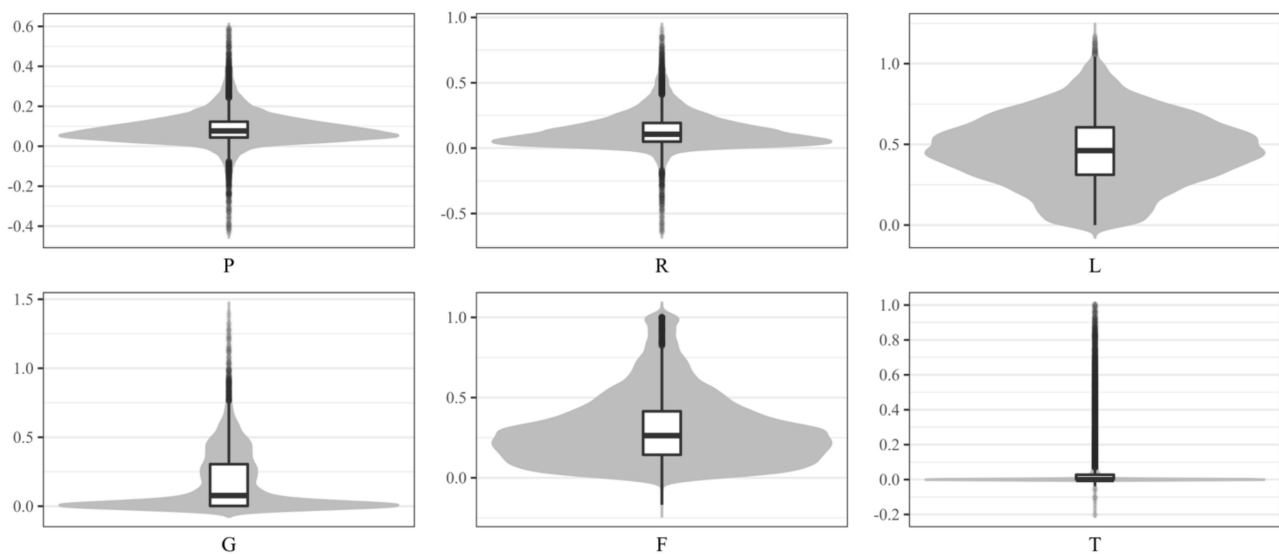
**Figure 1.** Boxplot statistics.

**Table 3.** Descriptive data statistics.

| Statistics | R | P | L | G | F | T |
|---|---|---|---|---|---|---|
| Minimum | −0.6451 | −0.4356 | −0.1212 | −0.7344 | −0.1683 | −0.2327 |
| 1 Quartile | 0.0495 | 0.0437 | 0.3123 | 0.0021 | 0.1435 | 0.0000 |
| Median | 0.1063 | 0.0766 | 0.4619 | 0.0762 | 0.2629 | 0.0000 |
| Mean | 0.1333 | 0.0880 | 0.4601 | 0.1720 | 0.3061 | 0.0639 |
| Standard deviation | 0.1341 | 0.0854 | 0.2196 | 0.2070 | 0.2175 | 0.1574 |
| 3 Quartile | 0.1926 | 0.1230 | 0.6070 | 0.3043 | 0.4153 | 0.0274 |
| Maximum | 0.9127 | 0.5916 | 1.7181 | 1.4004 | 1.0181 | 1.0055 |

1. The comparison of the maxima and minima with the boundaries of neighboring quartiles indicates that there are numerical outliers for almost all of the studied variables. Their rather high density hinders the development of clear cutoff criteria. Attempts to remove these outliers always ended in a loss of about 50% of the entire sample. Nearly all clustering methods are sensitive to outliers, i.e., the results of partitioning significantly depend on what is removed from the sample as an outlier or left in it.

2. The comparison of medians and averages indicates that P, R, L, and F are most likely unimodal, while G and T are not (see also Figure 1), i.e., the first set of strategies is used to some extent by all companies, while for the second set there is a class of companies that do not use them at all. Thus, the structure of data with respect to dimensions G and T "favors" the clustering, and with respect to P, R, L and F, does not.

3. A high value of standard deviation, giving a poor conditionality of the mean. This means that the choice of data normalization method will also be important for the results of analysis.

4. There is no option to decrease the problem's dimension. This issue can be indirectly predicted by the view of Table 2. A direct analysis of the key components shows that there is no such option.

We assume that distributions of the indicators shown in Figure 1 are consistent. Finding companies, for example, with low financial performance is of great interest for us, since the analysis of such clusters in dynamics can yield a lot of valuable information about the evolution of the financial position and strategies of companies.

### 4. Research Method

The problem of cluster analysis is defined as follows. Suppose we have a sample of $x \in X \subset R^n$ (in our case, $n = 6$). It is required to divide it into non-intersecting subsets $U_j$, $j = 1, \ldots, k$, with centers $\mu_j$, such that

$$L = \min_{\mu_j} \sum_{j=1}^{k} \sum_{x \in U_j} d(x, \mu_j) \tag{1}$$

is reached, where $d(x, \mu_j)$ is the distance metric value.

Problem (1) is nonconvex; thus, in general, $L$ is a local minimum, and the result of clustering depends on the chosen measure $d$, method of normalization of $x_i$, $i = 1, \ldots, n$, initial values of $x$ fed to the input of the algorithm for solving (1), and the algorithm itself. In addition, the solution exists for any $0 < k \leq m$, where $m$ is the number of elements in the sample, i.e., the number of clusters in the problem is undefined. Thus, the problem has many "degrees of freedom", and a neat approach to its solution requires discussion of each parameter.

1. The multi-scale data require normalization before the clustering procedure. The choice of normalization method is complicated by the objective lack of universal approaches [14]. The measure $z_{ri} = (x_{ri} - \min x_r)/(\max x_r - \min x_r)$ is often most convenient, bringing the ranges to 1 (here $r$ is the index number). However, our data contain many outliers and asymmetries of the indicators, which cannot be correctly eliminated (see Figure 1). As a result, such normalization can greatly distort the scale of the indicators. Therefore, a more common normalization by the standard deviation $z_{ri} = (x_{ri} - \overline{x}_r)/\sigma_r$, looks preferable in our case.

2. The choice of measure $d$ is closely related to the way it depends on the normalization of $x_i$. The classic metric is the Euclidean distance, but it is not always the best [15]. However, in our case, with almost uncorrelated data and the chosen normalization, there is no need to resort to generalized metrics of the Mahalanobis type, but it is reasonable settle with the Euclidean one.

3. There is no universal clustering algorithm [16]. In our case, an a priori choice of algorithm is impossible; thus, we used several algorithms, the results of which will be compared below.

4. There is no universal method for validating clustering results when the actual clusters of data are unknown [17]. Such a problem belongs to the class of internal cluster validation. Various internal cluster validation methods are essentially alternative clustering algorithms. The complexity of the problem here requires the use of alternative approaches due to the lack of a best-known approach. We focus on the silhouette and elbow metrics, the application of which contains both formalized and heuristic features.

The silhouette metric is defined [18,19] for each sample element: $s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$, where $a_i$ is the average distance from the element $x_i \in X$ to its cluster elements, $b_i$ is the average distance to the elements of the nearest cluster. By construction, $x_i \in [-1, 1]$. If $s_i = 1$, then the element belongs to its cluster. If $s_i = -1$, then the element is definitely located in the wrong cluster. If $s_i = 0$, then the element is located on the boundary of at least two clusters. For generalized evaluation of the clustering quality, we use $s_j = \overline{\overline{s}}_i$—the mean silhouette value over all cluster elements. A reasonable number of clusters is considered to be determined by the mean silhouette maximum. For the same reasons, when several clustering methods are used, the one with the maximum mean silhouette metric is recognized as the best one [19].

The elbow method is based on the comparative use of the total RMS distance $v_k = \sum_{j=1}^{k} \sum_{x_i \in X} (x_i - \mu_j)^2$ for various number of clusters (of the sum of within cluster variance with respect to the number of clusters) [17]. The sequence $v_k$ decreases with

respect to $k$, and the number of clusters is determined (as a rule, visually) as a transition from a large to a small change in the derivative of the resulting sequence.

Selecting a clustering algorithm, finding a certain value of the number of clusters and testing the silhouette metric for qualitative validation of clustering of our data are non-trivial tasks. To solve these problems, cluster analysis often uses visual data analysis as an auxiliary tool. However, for six-dimensional objects, as in our case, it is objectively difficult. As mentioned above, the reduction of the number of dimensions is not an option.

A way out of this situation was the method of data visualization in parallel coordinates [20]. We were inspired to use this tool by the paper "Pattern Analysis in Statics and Dynamics" [21,22]. Parallel coordinates allow for efficient visual analysis of multidimensional data because such a representation ensures no loss of information that the different dimensions contain [20].

Each object is represented in parallel coordinates with some given order of indicators, for example, PRLGFT. A brief illustration of the visualization mechanism is shown in Figure 2. The straight lines connecting neighboring indicators characterize the relationship between these indicators. The Value axis reflects the values for all indicators simultaneously.
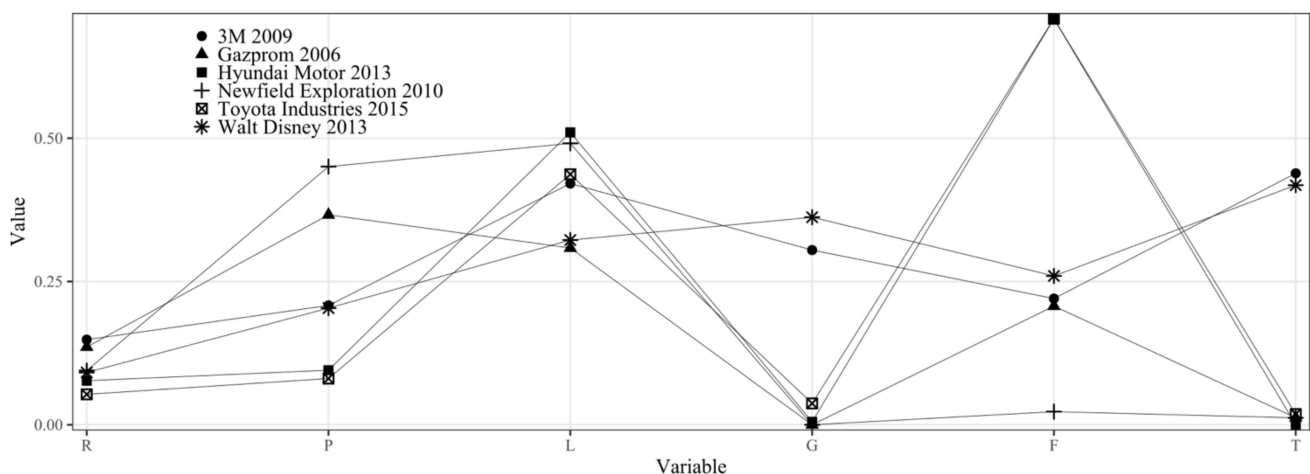


**Figure 2.** Illustration of clustering in parallel coordinates. Three clusters can be distinguished visually.

Figure 2 shows six companies at different points in time; visually, it is simple to divide them into three clusters:

1.  Toyota Industries 2015, Hyundai Motor 2013;
2.  3M 2009, Walt Disney 2013;
3.  Newfield Exploration 2010, Gazprom 2006.

The "k-means" algorithm returns the same clustering for $k = 3$.

Objects located close to each other in a n-dimensional space visually will have similar inclination angles of lines and locations of points on axes of indicators in parallel coordinates. For the objects located at a large distance from each other, the opposite picture will be observed; the objects will have different inclination angles and location of points on axes of indicators [20,22,23].

The order in which the measurements are presented can be a matter of discussion, because ideally the data should be presented in parallel coordinates in three variations with the following orders of indicators: RPTLFG, PLRGTF, LGPFRT. This is necessary in order to reflect all possible relationships between the variables [20]. In our case, the result turned out to be fundamentally independent of the order, and we settled on PRLGFT.

Classification of tens of thousands of objects by this method is hardly possible, but the visual evaluation of clusters obtained by some formal algorithmic method can help choose the best option.

Thus, to select the optimal clustering algorithm and the hyperparameter for the number of clusters $k$, we use the following procedure:

1.  Cluster the data by various methods;
2.  Compare the silhouette metric for various methods and various $k$;
3.  Selectively, according to the best in silhouette combinations of clustering algorithms and the number of clusters, present the results of clustering in parallel coordinates, visually analyze the quality of the resulting clusters;
4.  In the case of finding several configurations close to each other by the value of the silhouette, study in detail the resulting clusters in parallel coordinates and determine the (see below) partitioning of the best quality;
5.  Check the results of the silhouette metric and visual analysis for consistency in the conclusions about the clustering quality to understand whether the silhouette metric fits our data case or not.

By high-quality clustering, we mean the fulfillment of the following conditions listed in order of importance:

1.  The maximum of the objects included in one cluster have most similar slopes and line locations in parallel coordinates with respect to each other;
2.  The mean value of the silhouette metric is maximal (or close to maximal) comparing to other clustering variants;
3.  The resulting clusters have clear economic interpretations.

The alternative algorithms that we use are:

*   k-means [24,25];
*   Agglomerative clustering with Ward's linkage (agnes ward.D2) [26];
*   k-medoids (PAM) [27,28].

Of course, the above list is not complete. Testing all existing algorithms on our data does not make much sense in view of the huge number of such algorithms and the fact that the high-quality clustering problem has been solved by this set of algorithms. We tested our data on several other popular algorithms: spectral clustering, mixed Gaussian model, and applied other agglomerative clustering communication methods. However, the obtained clusters were of low quality; thus, we do not include these results in the discussion.

It is well known that clustering by the k-means algorithm strongly depends on the initial cluster centers, which are chosen randomly by default [17]. To find the best starting cluster centers in the k-means algorithm, we carry out the following procedure:

1.  For each considered number of clusters $k$, use a random number generator to select a large number (in this case, up to $k = 100$) of random objects from the sample as cluster centers (each center forms its own cluster);
2.  Use the k-means method for clustering with all random cluster centers for all $k$ under consideration;
3.  Determine the maximum mean silhouette value for each considered $k$ and analyze all unique clustering results through their representation in parallel coordinates;
4.  Memorize all initial cluster centers that provide the best data clustering for each $k$ (further, they will be used as the initial cluster centers in the k-means algorithm).

## 5. Clustering Results

The key issue of the clustering procedure is the number $k$ of clusters obtained. It should be noted right away that a clear boundary for the parameter $k$ could not be formed, and the value obtained was the result of several complementary compromise considerations.

The results of testing different clustering algorithms by silhouette metric are shown in Figure 3. For $k > 15$, the mean value of the silhouette metric does not exceed the best values. High values of the silhouette metric also have higher visual characteristics in parallel coordinates and a good meaningful interpretation.
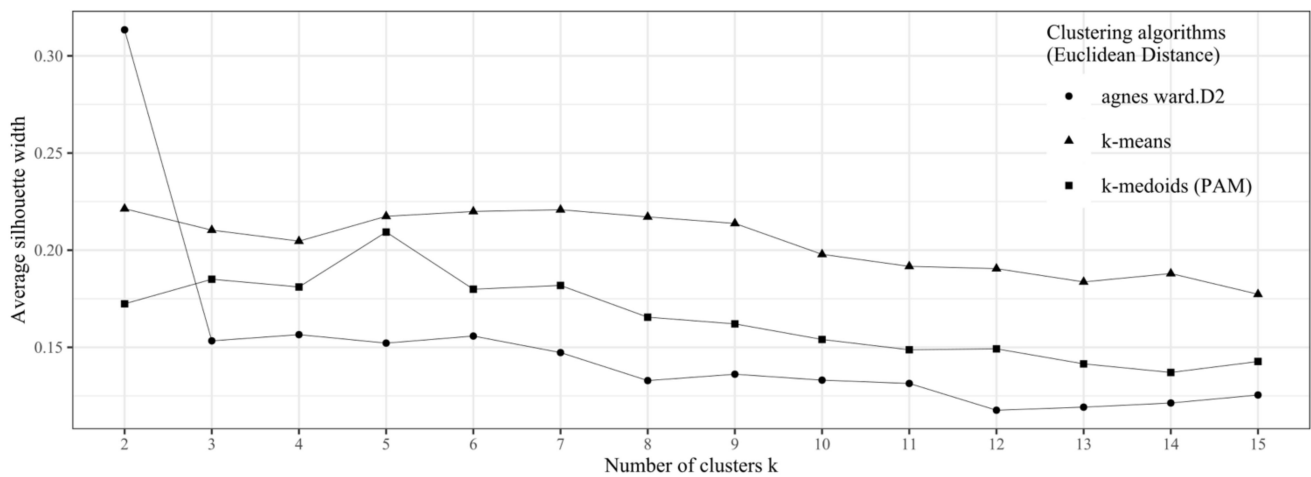
**Figure 3.** The mean silhouette value depending on the algorithm and *k* (the higher the better).

An alternative check of the number of clusters by the elbow method (Figure 4) also indicates the optimal number of clusters in the range from six to nine. However, in general the silhouette metric is more informative; thus, it is used as the main one. Among the algorithms, the k-means clustering algorithm proved to be the best, because it achieves the maximum silhouette values compared to other clustering algorithms for all considered *k*, except for *k* = 2 when we used the agnes ward.D2 algorithm.
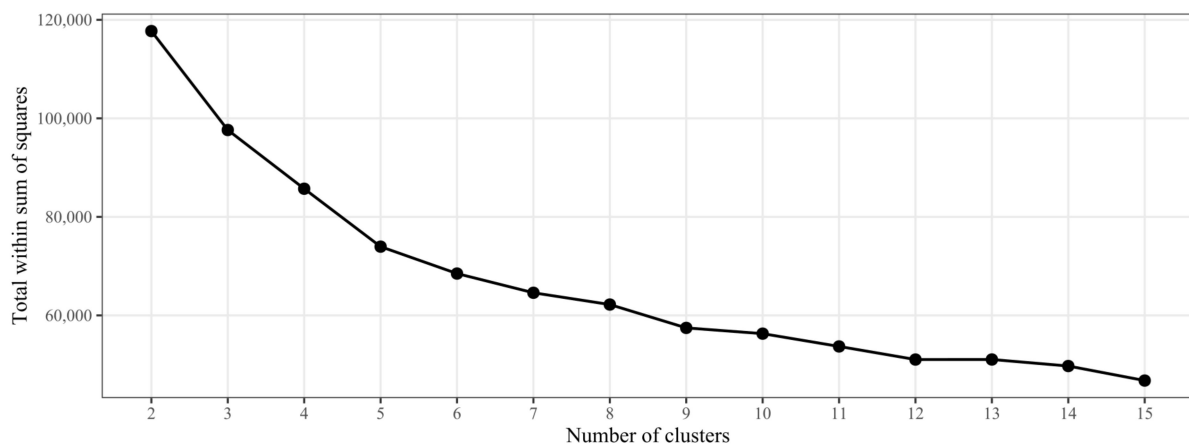


**Figure 4.** Mean variation of clusters $v_k$ (the elbow method) indicates the number of clusters from six to nine.

In Figure 3, the mean silhouette value for the agnes ward.D2 algorithm and *k* = 2 is 0.372. It may seem that such a set of methods provides the most optimal clustering, but at *k* = 2 the agnes ward.D2 algorithm divides data into giant and small clusters. An analysis of this clustering in parallel coordinates shows that the small cluster contains observations whose indicator T is non-zero for completely different values of the other indicators. Of course, this does not reveal the internal structure of the data; thus, it is of no value for the purposes of this study.

This situation with the silhouette metric is not unique. For example, when searching for the optimal number of clusters in the known Fisher iris dataset [29] using this metric with the Euclidean distance and almost any clustering algorithm, the maximum mean value of silhouette is reached at *k* = 2, when there are actually three clusters. It is in such a complex case when the parallel coordinate cluster analysis becomes helpful, because it ensures a more flexible selection of the hyperparameter *k*.

The maximum mean value of the silhouette (0.221) is achieved with the k-means algorithm and when *k* = 7. At the same time, Figure 3 shows that the silhouette values

for the k-means algorithm and the number of clusters of six, seven, and eight are quite close. In this complex case, the final choice of the hyperparameter *k* is not obvious. To solve this problem, we present the clustering results for k-means and the number of clusters six, seven, and eight in parallel coordinates and check them against the high-quality clustering conditions we introduced earlier through visual analysis.

The results of clustering are presented graphically in Figures 5–7. The ordinate axis (Z-score) reflects standardized values for all indicators. All clusters are assigned names that reflect the values of the indicators that distinguish them. They were formed based on the analysis of differences in the mean values of the indicators in the clusters in parallel coordinates and their standard deviations. All figures have a straight dotted line over the data, which runs through zero along the Z-score axis; it is used as a ruler to simplify the perception of parallel coordinates given the small size of the figures. The degree of transparency of the lines in the graphs is directly proportional to the density of objects in space.
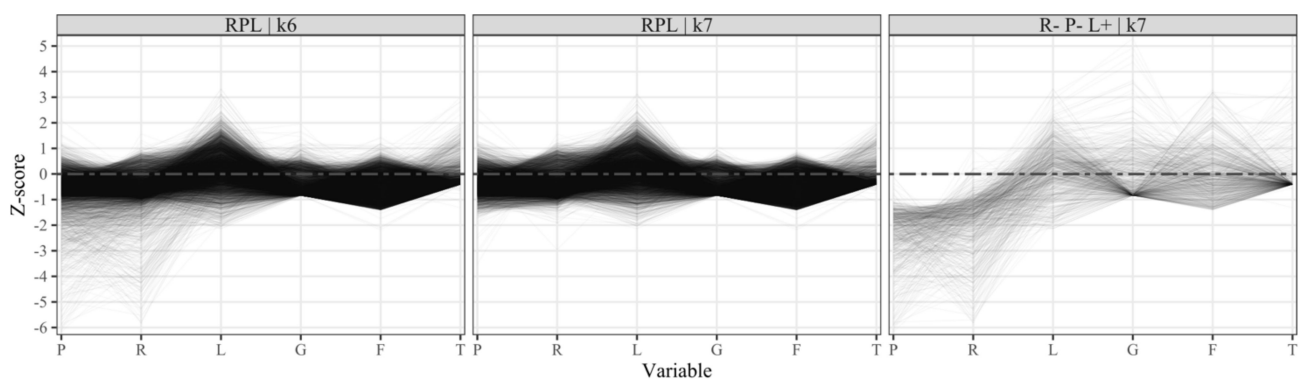


**Figure 5.** When partitioned into seven clusters, a new R-P-L+ cluster was separated from the RPL cluster. This can be seen from the comparison of the RPL cluster when partitioned into six clusters (**left**) and the pair of RPL and R-P-L+ clusters when partitioned into seven clusters (**right**).
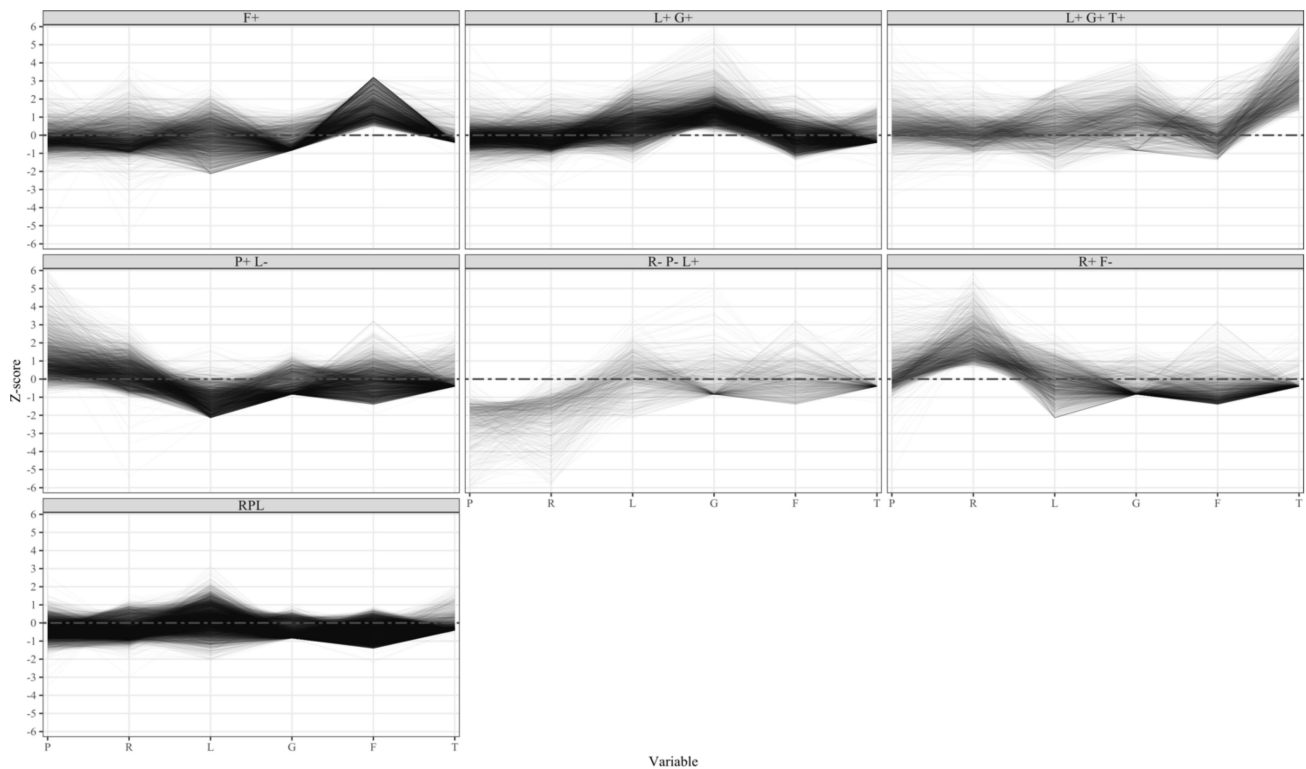


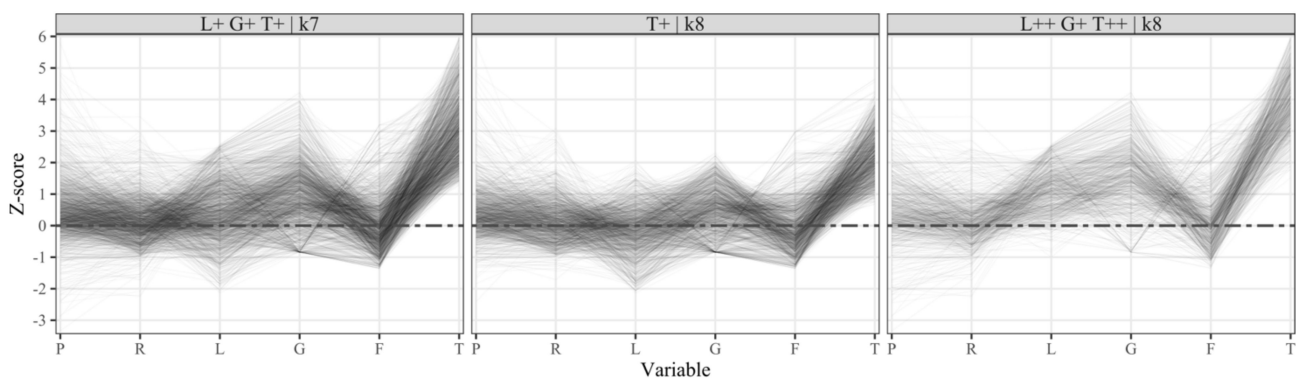**Figure 6.** Seven-cluster partitioning assumed as a final one.

**Figure 7.** In the eight-cluster partitioning, the old L+G+T+ cluster (**left**) splits into two similar clusters (**right**) T+ and L+G+T++.

Comparison of the results at *k* = 7 and *k* = 8 (Figures 6 and 7) shows that all clusters formed at *k* = 7 enter the eight-cluster partitioning while preserving the same general appearance of lines, except the L+G+T+ cluster. This cluster splits into two clusters: L+G+T++ (99% of objects used to be in L+G+T+) and T+ (75.6% of objects earlier were in L+G+T+). As a result, there is no separate different financial strategy but only a refinement of the structure of the already existing one. From this point of view, the clustering is less consistent with the original formulation of the problem.

The designation of the resulting clusters is based on their meaningful interpretation. To this end, the names of key indicators characterizing the cluster and the signs "+" or "−" are used, reflecting, respectively, its relatively high or low mean value. Figure 6 shows the following seven clusters:

F+—companies with an increased share of financial assets;

L+G+—companies active in mergers and acquisitions using borrowed funds;

L+G+T+—companies repurchasing their own shares (indicators L and G are high, but not independent);

P+L− —companies with high operating efficiency;

R+F− —companies with high capital efficiency;

R-P-L+—a cluster of companies with low profitability indicators;

RPL—companies that do not use risky financial strategies F, G and T;

The issue of choosing partitioning into seven rather than into six or eight (or more) clusters requires additional discussion. Let us consider the differences in partitions into six and seven clusters (Figures 5 and 6). All clusters formed at *k* = 6 are present with the same general line type in the partitioning into seven clusters. A new cluster appears, which is distinguished by low values of R and P and, on average, by elevated values of L. When divided into six clusters, the objects comprising the R-P-L+ cluster are mainly contained in the RPL cluster (77% of R-P-L+ objects were in this cluster) and, to a lesser extent, in the F+, L+G+, and L+G+T+ clusters.

The objects forming the new cluster R-P-L+ can be called outliers with a high degree of certainty if we consider them within the partitioning into six clusters, because these objects differ greatly in R and P from the mean values of the indicators in the clusters obtained at *k* = 6. The economic interpretation of the R-P-L+ cluster is clear and simple; it includes companies in certain years with weak financial fundamentals. This cluster includes many objects in crisis years, such as 2008, and few in years of economic growth and stability. Based on this, a seven-cluster partitioning is definitely better than a six-cluster one.

Graphical results of partitioning into 9, 10, etc. clusters are not given, but a similar effect occurs there: instead of identifying new strategies, detailing of the existing ones begins. Against this background, the boundaries between clusters become increasingly blurred, which leads to a decrease in the mean value of the silhouette. This explains why we stopped at *k* = 7. Figure 8 visualizes the information about the silhouette values for

each individual company, presented as a bar whose height equals the silhouette value, and Figure 9 presents the mean values of the indicators for each cluster in parallel coordinates.
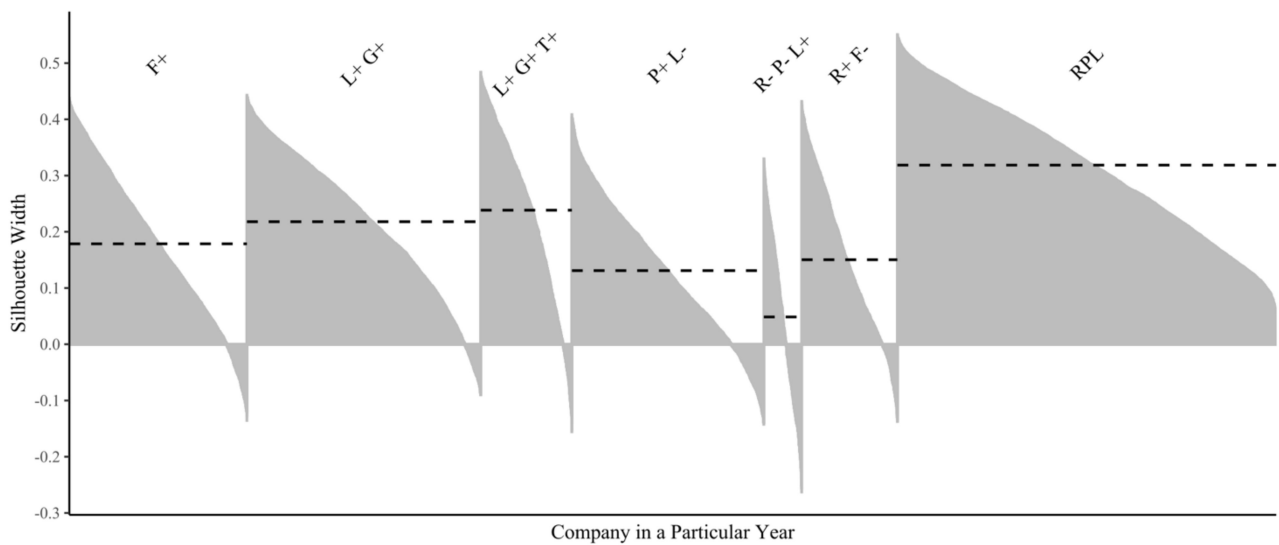


**Figure 8.** Silhouette metric values for all companies by cluster (the dotted line shows the mean silhouette value for each cluster).
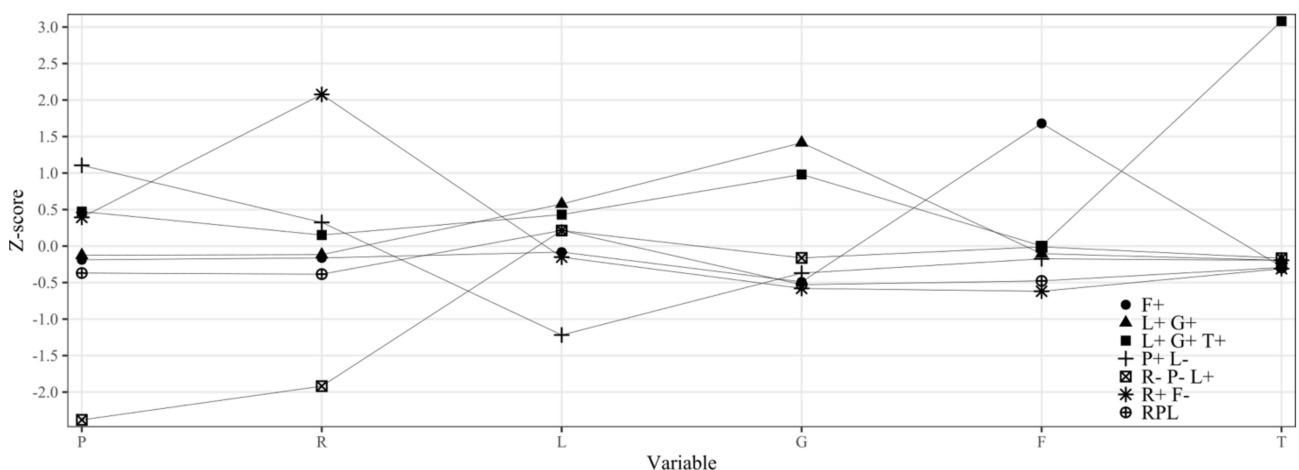


**Figure 9.** The mean values of the indicators over clusters in parallel coordinates. All clusters are unique in terms of their mean indicators.

Analysis of Figures 6 and 8 allows us to estimate approximately the quality of partitioning for certain clusters. The average silhouette value for all objects is 0.221, which is not high in absolute values, if we recall that the silhouette metric takes values from $-1$ to 1. The reason for this is that there is almost no "empty space" between clusters. This peculiarity of the data was discussed when considering their descriptive statistics. This circumstance imposes certain restrictions on the interpretation and further analysis of clustering results. Therefore, when analyzing if specific companies belong to a particular cluster, it is always necessary to check the silhouette value for this company. If it is almost equal or below zero, then it is necessary to illuminate this object in parallel coordinates and determine on the border of which clusters this company is located.

Clusters were named based on the analysis of the means in Figure 9 and the standard deviations of the indicators, which can be found together with other cluster statistics in the Appendix A (Table A3). For each cluster, we singled out indicators with mean values different to the mean of the corresponding indicators in other clusters and with relatively small standard deviations. The exception is the RPL cluster; this large cluster mostly

contains objects whose indicators are close to the mean values of the indicators of the whole sample, i.e., this cluster does not stand out as special.

We also conducted a test of stability of the obtained partitioning through clustering without a part of the sample. The test algorithm is as follows:

1.　Randomly select 25% of the companies from the entire sample, delete them;
2.　Following the algorithm we described, perform the k-means clustering of the truncated sample multiple times (we performed 50 iterations);
3.　Calculate the corrected rand index (ARI) between the final clusterization vector and clusterization vectors of the truncated sample, removing from the final clusterization those companies that were randomly removed from the sample.

A visualization of the fifty derived ARIs in the form of a box with a mustache is given in Appendix A P2. ARI takes values from −1 to 1, where −1 is completely different clustering, 0 is random matches, 1 is the same clustering. In our case, the median index is 0.947, the mean is 0.903. In general, the clustering proved to be stable.

## 6. Discussion

The meaningful interpretation of the clusters relies, first of all, on the graphical representation of the results in parallel coordinates shown in Figure 5. In addition, we will use the data on the of the clusters collected in Table 4. This industrial splitting is important because the representation of industries in the clusters is rather uneven. The unevenness of representation for both industries in the cluster and the cluster in the industry is important.

**Table 4.** Industrial representation in clusters.

| Industry/Cluster | F+ | L+G+ | L+G+T+ | P+L− | R-P-L+ | R+F− | RPL |
|---|---|---|---|---|---|---|---|
| Automobile industry | 280 | 131 | 42 | 132 | 32 | | 375 |
| Chemical industry | 102 | 197 | 126 | 216 | 14 | 56 | 565 |
| Construction | 623 | 309 | 24 | 135 | 47 | 300 | 570 |
| Electronics | 453 | 453 | 238 | 751 | 163 | 94 | 469 |
| Food | 152 | 341 | 136 | 264 | 11 | 42 | 335 |
| Power | 142 | 271 | 23 | 46 | 61 | 272 | 1046 |
| Consumer goods | 129 | 142 | 127 | 276 | 3 | 13 | 87 |
| Machine-building | 493 | 560 | 236 | 248 | 55 | 22 | 583 |
| Medicine | 271 | 439 | 240 | 348 | 38 | 47 | 121 |
| Metallurgical industry | 132 | 81 | 4 | 193 | 75 | 302 | 812 |
| Oil | 54 | 67 | 42 | 238 | 109 | 380 | 906 |
| Trade | 300 | 884 | 398 | 596 | 68 | 76 | 768 |
| Telecommunications | 211 | 503 | 84 | 208 | 27 | 89 | 191 |
| Transport | 112 | 171 | 53 | 84 | 33 | 169 | 542 |
| Total | 3454 | 4549 | 1773 | 3735 | 736 | 1862 | 7370 |

The F+ cluster includes companies with a high share of financial assets, provided that markers G and T indicate the absence of these strategies, and marker L can have any value. About 70% of companies in the cluster are in the following industries (in the descending order): construction, automobile industry, machine building, medicine, electronics, and trade. At the same time, industries have different "motivation" to use the F-strategy. For example, for construction, automotive, and machine-building industries it is typical to manufacture products purchased by the buyer on credit, which, in turn, is provided by the financial units of the manufacturer. Thus, the presence of financial assets is an objective necessity of such business, especially in construction where not only creation of real estate (construction itself) but also real estate management is often represented by financial assets. Electronics, trade, and medicine industries are another case, where financial assets are not "an obligatory attribute" but act as an object of the free cash flow investment. Therefore, it is not surprising that the share of the F + cluster in the first three industries is higher than in the last three.

The L+G+ cluster unites companies actively acting as buyers in mergers and acquisitions which occur with the formation of the goodwill G and, as a rule, are carried out with massive borrowing that increases L.

The leader here is the telecommunications industry, with more than 38% of observations represented in this cluster. Next come medicine, trade, food production, and machine building with shares of 25–25%. The presence of industries in the cluster indicates that they are undergoing active consolidation of companies. Thus, the telecommunications industry started to develop (if we take the old wireline companies out of the equation) with the mass emergence of regional mobile operators, subsequently acquired by major players. Similar processes based on the digitalization of logistics took place in trade. The industries represented in the cluster are characterized by the fact that the era of information technology and biotechnology "launched" business consolidation. Conversely, the automotive, oil, chemical and metallurgy industries, i.e., all the "old" industries, which passed the stage of consolidation much earlier, are poorly represented in this cluster.

The L+G+T+ cluster is structurally similar to the cluster L+G+, but it has a pronounced indicator T of using the strategy of own shares redemption and more blurred contours of indicators L and G. In addition, unlike L+G+, it has a rather modest size. There is no pronounced industry composition in this cluster. However, if we look at the names of companies representing this cluster, we see that many "blue chips" of the American stock market can be seen there: Colgate–Palmolive, Procter & Gamble, IBM, Pfizer, ExxonMobil, Coca-Cola, etc. Using the example of this cluster, we see that stock buybacks are a common strategy of companies with good cash flow and lack of investment ideas, forcing them to invest in their own business, and they invest not in its production, but in finance.

The emergence of L+G+ and L+G+T+ clusters was quite expected. This was "hinted at" by the correlation between L and G, as well as by the data structure of the T indicator. Less expected was the separation of clusters P+L− and R+F−, since there is a high correlation between P and R. Nevertheless, this can be explained by the fact that the low P value in the R+F− cluster is associated with high company revenue and relatively low capital (Recall that profit is included in the numerator of both P and R, but in the first case the denominator is revenue, while in the second case it is capital). There are relatively few such companies (observations), because the R+F− cluster is quite small as opposed to the large P+L− cluster. There is also a striking industry distinction between the clusters. R+F− represents the "old" economy: construction, energy, oil production, metallurgy, and transportation. While the P+L− cluster comprises companies of the "new" economy: electronics, medicine, trade, food, and consumer goods production.

The R-P-L+ cluster is the smallest among the revealed ones and represents observations with (very) poor financial performance. This cluster cannot be associated with a targeted financial strategy, rather, it is a financial situation in which the company cannot stay for a long time. Many companies in this cluster "did not make it" to 2018. For this cluster, the effect of a short-term "stay" is most pronounced. A similar, but not convex, effect is characteristic of the "successful" P+L− and R+F− clusters. It is expressed by relatively few companies, which would be in one of the clusters during the whole period from 2006 to 2018. At the same time, clusters F+, L+G+ and L+G+T+ are characterized by a noticeably higher number of such companies (see Table 5).

For the P and R type companies, the RPL cluster acts as a location. More than 30% of the "old" economy companies are located in it, and this share reaches 50% for the energy, metallurgy, and oil industries and over 40% for the chemical and transport industries. At the same time, the "new" economy sectors are relatively poorly represented in it.

Thus, we can say that the stable classification indicators of the financial strategy are clusters F+, L+G+, L+G+T+, and RPL in which a relatively large number of companies is observed during all (or almost all) periods of observation. However, on this basis, the remaining clusters cannot be regarded as "superfluous". Thus, clusters P+L− and R+F− include the companies most successful in implementing the RPL strategy, and, respectively, R-P-L+ as the ones failing to do so.

**Table 5.** Number of companies that were in a particular cluster for a certain number of years between 2006 and 2018.

| Number of Years/Cluster | F+ | L+G+ | L+G+T+ | P+L− | R-P-L+ | R+F− | RPL |
|---|---|---|---|---|---|---|---|
| 13 consecutive years | 53 | 81 | 32 | 37 | 0 | 16 | 120 |
| At least 12 years | 78 | 126 | 50 | 60 | 0 | 26 | 209 |
| At least 11 years | 108 | 168 | 70 | 76 | 0 | 39 | 284 |
| Cluster's share (The share relative to the total of Table 4. To calculate it, we should take into account that the companies of the first line of Table 5 occur 13 times in the cluster, etc.) | 38% | 45% | 48% | 25% | 0% | 25% | 47% |

The stability can also be looked at from another perspective. Figure 10 shows the details of the resulting clustering year-by-year. The changes in the dynamics that we observe are quite interesting:
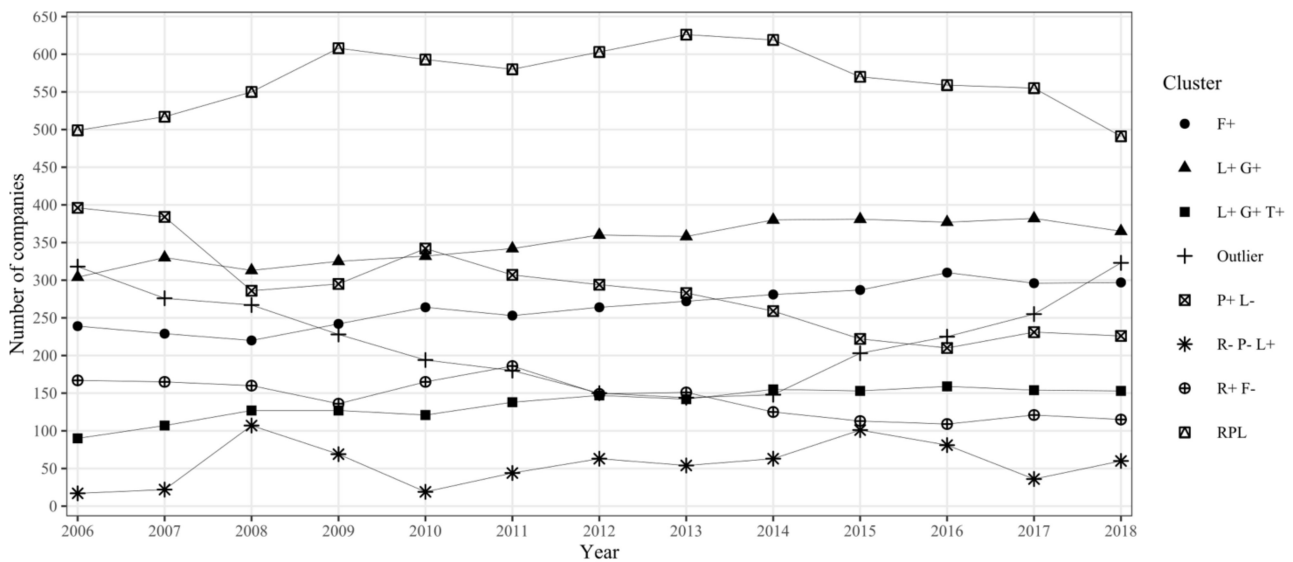


**Figure 10.** Clustering results with year-by-year detailing.

Clusters F+, L+G+ and L+G+T+ grow (almost) monotonically;

The R-P-L+ cluster has characteristic spikes in 2008 (global financial crisis) and 2015 (crisis of falling oil prices). The P+L− cluster has a dip in 2008. The negative consequences occurred in 2008, which allows us to assume the financial nature of the losses, since the production and consumer phases of the crisis took place in 2009 (This can be seen, for example, in the GDP dynamics of the countries affected by the crisis, where the fall happened in 2009). The financial phase of the crisis happened in 2008;

The number of companies in clusters P+L− and R+F− tend to decrease;

The RPL cluster has a trajectory mirroring the "cluster" of outliers, which includes not only companies with data outliers, but also those with missing or defective data. This suggests that the data losses come mainly from the RPL cluster.

In general, the low and explainable volatility of the trajectories demonstrates that the resulting clustering can be characterized as stable in terms of dynamics.

## 7. Conclusions

The clustering of companies by indicators R, P, L, G, F, T aims to identify the financial strategy of the company as a stable classification attribute. The initial information was the

corporate financial statements. The resulting dataset has a rather dense structure that does not favor clustering.

The clustering was carried out using the k-means method with a Euclidean measure supplemented by the procedure of forming a stable set of starting points. The number of clusters is the key issue here, and it was addressed by using the cluster silhouette metric augmented by visual analysis of the results in parallel coordinates. The low value of the average cluster silhouette provides a good description of the complexity of the problem under study. The result of the partitioning was seven clusters. This partitioning is stable and reproducible on truncated subsets of the original data.

Each of the resulting clusters is associated with a particular financial strategy. Clusters F+, L+G+, L+G+T+, and RPL are characterized by a prolonged positioning of companies in the cluster over the entire study period from 2008 to 2018. Although the first three clusters are associated with risky strategies, we see that these risks have not been realized for a long time. In contrast, the P+L−, R+F− clusters rarely have long stays, and the R-P-L+ cluster has the shortest stays. Therefore, it is the first mentioned group of clusters that can be characterized as a stable classification attribute of the financial strategy.

In addition, clusters can be differentiated by industry. Thus, in clusters L+G+ and P+L−, the companies of the "new" economy, which is characterized by intensive consolidation, are strongly represented. While F+, R+F−, RPL have a "bias" toward the companies of the "old" economy, represented by the basic industries, where business consolidation has been largely completed.

Thus, the present work solves the problem of identifying a financial strategy as a stable classification attribute by methods of cluster analysis. The analysis of the evolution and dynamics of companies is beyond the scope of this paper and is a task for a separate study.

## Appendix A

**Table A1.** Correlation matrix of additional indicators.

| | Return on Assets | Return on Equity | Return on Sales | Working Capital Turnover | Fixed Assets Turnover | Cash Ratio | Quick Ratio | Current Ratio |
|---|---|---|---|---|---|---|---|---|
| Return on equity | 0.71 | | | | | | | |
| Return on sales | 0.66 | 0.46 | | | | | | |
| Working capital turnover | 0.03 | 0.02 | −0.07 | | | | | |
| Fixed assets turnover | 0.10 | 0.13 | −0.09 | 0.13 | | | | |
| Cash ratio | 0.25 | 0.06 | 0.27 | −0.01 | 0.02 | | | |
| Quick ratio | 0.27 | 0.07 | 0.24 | 0.05 | 0.08 | 0.90 | | |
| Current ratio | 0.26 | 0.05 | 0.22 | 0.07 | 0.07 | 0.74 | 0.84 | |
| Debt to equity ratio | −0.26 | 0.01 | −0.15 | −0.07 | −0.09 | −0.24 | −0.27 | −0.28 |
| Debt to capital ratio | −0.04 | 0.10 | −0.06 | −0.02 | 0.05 | −0.17 | −0.18 | −0.17 |
| Interest coverage ratio | 0.46 | 0.28 | 0.26 | 0.06 | 0.11 | 0.19 | 0.23 | 0.23 |
| Debt coverage ratio | 0.08 | 0.02 | 0.00 | 0.05 | 0.14 | 0.08 | 0.10 | 0.10 |
| P | 0.49 | 0.33 | 0.79 | −0.10 | −0.15 | 0.23 | 0.19 | 0.17 |
| R | 0.93 | 0.76 | 0.57 | 0.04 | 0.15 | 0.17 | 0.18 | 0.16 |
| L | −0.19 | −0.07 | −0.26 | −0.07 | −0.08 | −0.35 | −0.37 | −0.39 |
| G | −0.05 | 0.09 | −0.10 | 0.01 | 0.27 | −0.14 | −0.08 | −0.13 |
| F | −0.03 | 0.01 | 0.01 | −0.02 | 0.28 | 0.09 | 0.07 | 0.08 |
| T | 0.03 | 0.22 | 0.02 | 0.05 | 0.13 | 0.03 | 0.06 | 0.05 |

**Table A2.** Correlation matrix of additional indicators, continuation.

| | Debt to Equity Ratio | Debt to Capital Ratio | Interest Coverage Ratio | Debt Coverage Ratio | P | R | L | G | F |
|---|---|---|---|---|---|---|---|---|---|
| Debt to capital ratio | 0.25 | | | | | | | | |
| Interest coverage ratio | −0.27 | −0.02 | | | | | | | |
| Debt coverage ratio | −0.15 | 0.04 | 0.51 | | | | | | |
| P | −0.07 | −0.04 | 0.25 | −0.07 | | | | | |
| R | −0.24 | 0.00 | 0.40 | 0.03 | 0.45 | | | | |
| L | 0.60 | 0.26 | −0.40 | −0.27 | −0.31 | −0.14 | | | |
| G | 0.29 | 0.04 | −0.10 | −0.15 | −0.01 | −0.05 | 0.35 | | |
| F | 0.02 | 0.08 | 0.02 | 0.15 | −0.03 | −0.05 | −0.01 | −0.03 | |
| T | 0.11 | −0.07 | 0.05 | −0.03 | 0.14 | 0.03 | 0.13 | 0.31 | 0.01 |

**Table A3.** Descriptive statistics of indicators by clusters.

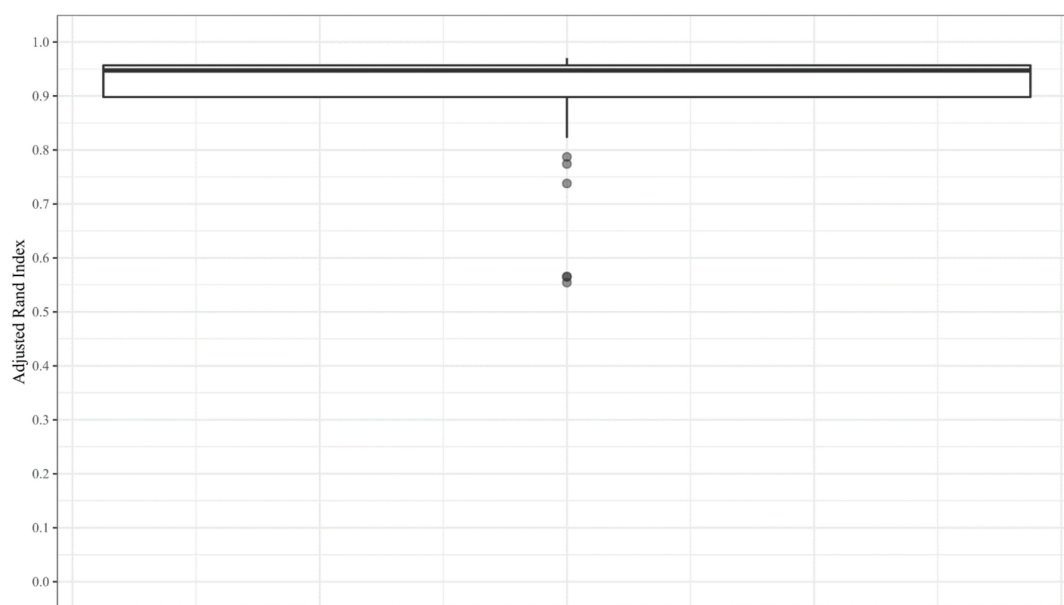| Cluster | F+ | | | | | | |
|---|---|---|---|---|---|---|---|
| **Statistics Indicator** | **Min** | **Q1** | **Median** | **Mean** | **Std.Dev** | **Q3** | **Max** |
| R | −0.59 | 0.04 | 0.09 | 0.11 | 0.11 | 0.17 | 0.65 |
| P | −0.35 | 0.04 | 0.06 | 0.07 | 0.05 | 0.10 | 0.42 |
| L | 0.00 | 0.30 | 0.44 | 0.44 | 0.21 | 0.59 | 1.09 |
| G | 0.00 | 0.00 | 0.02 | 0.07 | 0.10 | 0.11 | 0.45 |
| F | 0.40 | 0.53 | 0.64 | 0.67 | 0.17 | 0.79 | 1.00 |
| T | −0.10 | 0.00 | 0.00 | 0.02 | 0.05 | 0.01 | 0.57 |
| **Cluster** | **L+G+** | | | | | | |
| **Statistics Indicator** | **Min** | **Q1** | **Median** | **Mean** | **Std.Dev** | **Q3** | **Max** |
| R | −0.26 | 0.06 | 0.11 | 0.12 | 0.08 | 0.16 | 0.55 |
| P | −0.18 | 0.05 | 0.07 | 0.08 | 0.05 | 0.10 | 0.51 |
| L | 0.05 | 0.45 | 0.57 | 0.58 | 0.17 | 0.69 | 1.16 |
| G | 0.17 | 0.37 | 0.44 | 0.46 | 0.15 | 0.53 | 1.40 |
| F | −0.07 | 0.18 | 0.27 | 0.28 | 0.14 | 0.37 | 0.91 |
| T | −0.20 | 0.00 | 0.00 | 0.03 | 0.07 | 0.03 | 0.39 |
| **Cluster** | **L+G+T+** | | | | | | |
| **Statistics Indicator** | **Min** | **Q1** | **Median** | **Mean** | **Std.Dev** | **Q3** | **Max** |
| R | −0.17 | 0.09 | 0.14 | 0.15 | 0.09 | 0.20 | 0.60 |
| P | −0.20 | 0.09 | 0.12 | 0.13 | 0.08 | 0.16 | 0.58 |
| L | 0.00 | 0.41 | 0.54 | 0.55 | 0.20 | 0.69 | 1.01 |
| G | 0.00 | 0.23 | 0.37 | 0.38 | 0.21 | 0.51 | 1.04 |
| F | 0.01 | 0.18 | 0.28 | 0.31 | 0.18 | 0.40 | 1.00 |
| T | 0.27 | 0.42 | 0.53 | 0.55 | 0.16 | 0.65 | 1.01 |
| **Cluster** | **P+L−** | | | | | | |
| **Statistics Indicator** | **Min** | **Q1** | **Median** | **Mean** | **Std.Dev** | **Q3** | **Max** |
| R | −0.59 | 0.11 | 0.17 | 0.18 | 0.09 | 0.24 | 0.55 |
| P | 0.01 | 0.12 | 0.16 | 0.18 | 0.09 | 0.22 | 0.59 |
| L | 0.00 | 0.09 | 0.18 | 0.20 | 0.13 | 0.29 | 0.86 |
| G | 0.00 | 0.00 | 0.04 | 0.10 | 0.12 | 0.17 | 0.55 |
| F | −0.11 | 0.13 | 0.25 | 0.27 | 0.17 | 0.37 | 1.00 |
| T | 0.00 | 0.00 | 0.00 | 0.03 | 0.08 | 0.02 | 0.48 |
| **Cluster** | **R-P-L+** | | | | | | |
| **Statistics Indicator** | **Min** | **Q1** | **Median** | **Mean** | **Std.Dev** | **Q3** | **Max** |
| R | −0.65 | −0.19 | −0.08 | −0.12 | 0.15 | −0.03 | 0.27 |
| P | −0.44 | −0.15 | −0.08 | −0.11 | 0.09 | −0.05 | 0.08 |
| L | 0.00 | 0.36 | 0.52 | 0.50 | 0.22 | 0.65 | 1.17 |
| G | 0.00 | 0.00 | 0.05 | 0.14 | 0.19 | 0.24 | 1.26 |
| F | 0.00 | 0.14 | 0.27 | 0.30 | 0.21 | 0.42 | 1.02 |
| T | −0.14 | 0.00 | 0.00 | 0.04 | 0.10 | 0.01 | 0.66 |
| **Cluster** | **R+F−** | | | | | | |
| **Statistics Indicator** | **Min** | **Q1** | **Median** | **Mean** | **Std.Dev** | **Q3** | **Max** |
| R | 0.20 | 0.31 | 0.38 | 0.41 | 0.13 | 0.49 | 0.91 |
| P | −0.35 | 0.07 | 0.10 | 0.12 | 0.08 | 0.16 | 0.59 |
| L | 0.00 | 0.29 | 0.43 | 0.42 | 0.20 | 0.56 | 1.14 |
| G | 0.00 | 0.00 | 0.00 | 0.05 | 0.10 | 0.06 | 0.61 |
| F | 0.00 | 0.05 | 0.12 | 0.17 | 0.17 | 0.23 | 1.00 |
| T | −0.02 | 0.00 | 0.00 | 0.02 | 0.05 | 0.00 | 0.51 |
| **Cluster** | **RPL** | | | | | | |
| **Statistics Indicator** | **Min** | **Q1** | **Median** | **Mean** | **Std.Dev** | **Q3** | **Max** |
| R | −0.26 | 0.03 | 0.07 | 0.08 | 0.07 | 0.12 | 0.34 |
| P | −0.21 | 0.03 | 0.05 | 0.06 | 0.04 | 0.08 | 0.31 |
| L | 0.00 | 0.39 | 0.50 | 0.50 | 0.16 | 0.61 | 1.13 |
| G | 0.00 | 0.00 | 0.02 | 0.06 | 0.08 | 0.11 | 0.40 |
| F | −0.17 | 0.11 | 0.20 | 0.20 | 0.12 | 0.29 | 0.50 |
| T | −0.11 | 0.00 | 0.00 | 0.02 | 0.05 | 0.01 | 0.38 |

**Figure A1.** Boxplot of 50 values of the adjusted rand index obtained by comparing the similarity of the clustering of truncated samples with the final k-means partitioning into seven clusters.

## References

1.  Ivashkovskaya, I.V. *Financial Dimensions of Corporate Strategies. Stakeholder Approach. M.*; INFRA-M: Moscow, Russia, 2013; 320p. (In Russian)
2.  Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
3.  Ohlson, J.A. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [CrossRef]
4.  Hillegeist, S.A.; Keating, E.K.; Cram, D.P.; Lundstedt, K.G. Assessing the probability of bankruptcy. *Rev. Account. Stud.* **2004**, *9*, 5–34. [CrossRef]
5.  Philosophov, L.V.; Batten, J.A.; Philosophov, V.L. Predicting the event and time horizon of bankruptcy using financial ratios and the maturity schedule of long-term debt. *Math. Financ. Econ.* **2008**, *1*, 181–212. [CrossRef]
6.  Chouhan, V.; Chandra, B.; Goswami, S. Predicting financial stability of select BSE companies revisiting Altman Z score. *Int. Lett. Soc. Humanist. Sci.* **2014**, *15*, 92–105. [CrossRef]
7.  Tian, S.; Yu, Y. Financial ratios and bankruptcy predictions: An international evidence. *Int. Rev. Econ. Financ.* **2017**, *51*, 510–526. [CrossRef]
8.  Capece, G.; Cricelli, L.; Di Pillo, F.; Levialdi, N. A cluster analysis study based on profitability and financial indicators in the Italian gas retail market. *Energy Policy* **2010**, *38*, 3394–3402. [CrossRef]
9.  R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2020; Available online: https://www.R-project.org/ (accessed on 8 November 2021).
10. Ramanna, K. The implications of unverifiable fair-value accounting: Evidence from the political economy of goodwill accounting. *J. Account. Econ.* **2008**, *45*, 253–281. [CrossRef]
11. Dzuba, S.A. Goodwill: A toxic asset with a plausible name. *EKO* **2016**, *3*, 165–183. (In Russian)
12. Bonaimé, A.A.; Hankins, K.W.; Jordan, B.D. The cost of financial flexibility: Evidence from share repurchases. *J. Corp. Financ.* **2016**, *38*, 345–362. [CrossRef]
13. Nedelkin, I. Pandemic Left Buffett One Direction to Invest. Expert Online. 2020. Available online: https://expert.ru/2020/11/9/pandemiya-ostavila-baffetu-odno-napravlenie-dlya-investitsij/ (accessed on 8 November 2021). (In Russian)
14. De Souto, M.C.P.; De Araujo, D.S.A.; Costa, I.G.; Soares, R.G.F.; Ludermir, T.B.; Schliep, A. Comparative study on normalization procedures for cluster analysis of gene expression datasets. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 8–13 July 2018; pp. 2792–2798. [CrossRef]
15. Shirkhorshidi, A.S.; Aghabozorgi, S.; Wah, T.Y. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS ONE* **2015**, *10*, e0144059. [CrossRef]
16. Jain, A.; Murty, M.; Flynn, P. Data Clustering: A Review. Association for Computing Machinery. *ACM Comput. Surv.* **1999**, *31*, 264–323. [CrossRef]
17. Aggarwal, C.C. Cluster Analysis. In *Data Mining*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 153–204. [CrossRef]

18. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **2013**, *46*, 243–256. [CrossRef]
19. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
20. Inselberg, A. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009. [CrossRef]
21. Aleskerov, F.T.; Belousova, V.Y.; Egorova, L.G.; Mirkin, B.G. Pattern analysis in statics and dynamics, Part 1: Literature review and concept clarification. *Bus. Inform.* **2013**, *3*, 3–18. (In Russian)
22. Aleskerov, F.T.; Belousova VYu Egorova, L.G.; Mirkin, B.G. Pattern analysis in statics and dynamics, Part 2: Examples of application to the analysis of socio-economic processes. *Bus. Inform.* **2013**, *4*, 3–20. (In Russian)
23. Chou, S.; Lin, S.; Yeh, C. Cluster identification with parallel coordinates. *Pattern Recognit. Lett.* **1999**, *20*, 565–572. [CrossRef]
24. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28*, 100. [CrossRef]
25. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
26. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
27. Kaufman, L.; Rousseeuw, P.J. Clustering by means of medoids. In *Statistical Data Analysis Based on the L1 Norm and Related Methods*; Dodge, Y., Ed.; North-Holland: New York, NY, USA, 1987; pp. 405–416.
28. Schubert, E.; Rousseeuw, P.J. Faster K-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms. In *Similarity Search and Applications*; Springer: Cham, Switzerland, 2019; pp. 171–187. [CrossRef]
29. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]