Journal of Business Economics and Management
ISSN 1611-1699 / eISSN 2029-4433
2019 Volume 20 Issue 6: 1238–1257
https://doi.org/10.3846/jbem.2019.11237

MEASURING PERFORMANCE BY INTEGRATING K-MEDOIDS WITH DEA: MONGOLIAN CASE

Batchimeg BAYARAA*, Tibor TARNOCZI, Veronika FENYVES

Institute of Accounting and Finance, Faculty of Economics and Business,
Károly Ihrig Doctoral School of Management and Business,
University of Debrecen, Debrecen, Hungary

Received 30 January 2019; accepted 18 July 2019

Abstract. Performance measurement encourages Decision Making Units (DMUs) to improve their level of performance by comparing their current financial positions with that of their peers. Data Envelopment Analysis (DEA) is a widely used approach to performance measurement, though it is susceptible when the data is heterogeneous. The main objective of this study is to examine the performance of Mongolian listed companies by combining DEA and a k-medoid clustering method. Clustering facilitates the characterization and patterns of data and identification of homogenous groups. This study applies the integration of k-medoids and performance measurement. The research used 89 Mongolian companies' financial statements from 2012 to 2015 – obtained from the Mongolian Stock Exchange website. The companies are grouped by k-medoids clustering, and efficiency of each cluster is evaluated by DEA. According to the silhouette method, the companies are classified into two clusters which are considered first cluster as small and medium-sized (80), and second cluster as big (9) companies. Both clusters are analyzed and compared by financial ratios. The mean efficiency score of big companies' is much higher than that of small and medium-sized companies. Integrated results show that cluster-specific efficiency provides better performance than pre-clustering efficiency results.

Keywords: financial performance, k-medoids clustering, data envelopment analysis, input efficiency, variable return to scale, decision making unit.

JEL Classification: C38, C14, L25.

Introduction

The basis of all types of analysis is data. It is possible to face various data in everyday life, with or without any prior knowledge. However, further analysis cannot be made without knowing the pattern and the characteristics of data. Data classification – by grouping or clustering – is one means of primary analysis. Clusters have been determined in many ways, yet there is no single determination which is globally accepted. Cluster analysis is used for understand-

*Corresponding author. E-mail: *bayaraa.batchimeg@econ.unideb.hu*

ing (finding meaningful groups of objects that share common characteristics) and utility (to abstract the representative object from among many others in the same clusters) (Wu, 2012).

Clustering techniques are divided into partitional and hierarchical types. The most popular and well-known partitional cluster technique is k-means, which is widely employed in research. Although k-means is a popular choice among partitional clusters, it is sensitive to outliers. On the contrary, the k-medoids algorithm is more robust and less sensitive to outliers. Research, which compared k-medoids with k-means, suggested the k-medoid was better in all aspects. For example, Arora and Varshney (2016) compared k-means and k-medoids in their research. Their results proved that k-medoids is better than k-means; as execution time, sensitivity to outliers and space complexity of overlapping are all less. Narayana and Vasumathi (2018) stated in their work that the k-medoids technique is more accurate and easier to understand than k-means clustering. Moreover, Patel and Singh (2013) studied a new approach for k-means and k-medoids algorithm and concluded that k-medoids improved accuracy. However, Arbin, Suhaimi, Mokhtar, and Othman (2016) evaluated k-means and k-medoids, and both methods were found to be good having mean errors less than three.

The K-medoids algorithm, which was proposed by Kaufman and Rousseeuw (1987), was developed and investigated by various researchers from different fields. For example, Ho-Kieu, Vo-Van, and Nguyen-Trang (2018) verified and compared the effectiveness and feasibility of the k-medoids method and algorithm with various other algorithms' through artificial and real datasets. The results revealed an outstanding performance through the evaluation criteria. Park and Jun (2009) recommended a simple and fast algorithm for k-medoids clustering. In this work, a new algorithm – which runs like the k-means algorithm – was proposed. Mohammad, Zadegan, Mirzaie, and Sadoughi, (2013) examined ranked k-medoids. They introduced a new k-medoids algorithm, which can find all Gaussian-shaped clusters. Zhang and Couloigner (2005) suggested a new k-medoids algorithm for spatial clustering in large applications. Gandhi and Srivastava (2014) presented an overview of the modified k-medoids algorithm to improve scalability and efficiency. Sood and Bansal (2013) surveyed the combination of the k-medoids algorithm and the bat algorithm. Mei and Chen (2011) investigated medoid-based fuzzy relational clustering.

Clustering itself is not the final result, rather it is a possible data input for further analysis. Therefore, the aim of the study was to improve the accuracy of performance analysis by integrating it with clustering. It must be mentioned that clustering methods have been used with efficiency analysis before. For example, Omrani, Shafaat, and Emrouznejad (2018) integrated a fuzzy clustering cooperative game with DEA model in an application to hospital efficiency. In their research, 288 hospitals from 31 provinces of Iran examined. Similarly, Jahangoshai, Rezaee, Jozmaleki, and Valipour (2018) integrated fuzzy C-means, DEA, and an artificial neural network. They obtained their data (from 2007 to 2012) from the Tehran Stock Market and used financial ratios as variables. Thakare and Bagal (2015) evaluated the performance of k-means with various metrics. They concluded that the performance of the k-means algorithm is based on the distance metrics as well the database used. Kim, Lee, and Kang (2018) integrated eight internal clustering efficiency measures based on DEA and proposed a new cluster validity index. Amin, Wan-Ismail, Abdul-Rasid, and Selemani (2014) analyzed some issues confront the DEA clustering algorithm. Kianfar Ahadzadeh Namin, Alam Tabriz, Na-

jafi, and Hosseinzadeh Lotfi (2017) examined a hybrid cluster and DEA. They concluded that there was a perceptible difference between the efficiency of DMUs with the upper bound and with the lower bound. Po, Guh, and Yang (2009) presented a new clustering approach using DEA. Bi, Song, and Wu (2014) proposed slack-based measure-based clustering method to classify the environmental performance of Chinese industry. Dai and Kuosmanen (2014) proposed benchmarking using the clustering method. They concluded cluster-specific efficiency ranking provides more efficient and meaningful benchmarking than the conventional approach. Moreover, some researchers integrated clustering with a parametric method such as, 'Bayesian clustering in stochastic frontier analysis' by Griffin (2011).

According to the Mongolian Stock Exchange's (MSE) research, 475 companies were registered for the last 10 years; however, 202 companies were delisted by March of 2019. This shows the need for the listed companies to be evaluated properly – by their performance – and make improvements based on best practices. From the researchers' point of view, there is no published research used in financial statements' parameters for k-medoids. Also, the data analyzed in this paper has high variability. Therefore, the algorithm of k-medoids is chosen instead of k-means; which is less sensitive to noise and outliers than k-means. Moreover, there is no published research using k-medoids for Mongolian companies, so this study applies k-medoids which is a comparably new clustering method.

The main goal is to measure the performance of Mongolian listed companies. To do so, the following steps are made:

- Fundamental statistical analysis to decide whether clustering is appropriate for the data;
- Determination of the optimal number of clusters, using the silhouette method;
- Identification of clusters by K-medoids;
- Analysis of each cluster by ratio analysis;
- Evaluation of the performance of each cluster, and comparison between the performance with that of pre-clustering performance.

The remainder of this paper is organized in the following way. After the introduction, the first section reviews the literature on performance measurement, the DEA method, clustering, and the k-medoids algorithm. The second section presents data sets and variables used during the analysis. Section three consists of empirical results and the conclusions are presented in section four.

1. Literature review

1.1. Performance measurement

Corporate performance is the measurement of what has been achieved by a company. Masri (2013) notes, 'Performance measurement system is used by an organization not just to determine whether its objectives have been met but also as a means of comparing their performance with that of other DMUs (decision-making units)'. DMUs can be firms, organizations, divisions, industries, projects or individuals.

Performance can be divided into two parts: efficiency and effectiveness, which are often confused. Neely, Gregory, and Platts (1995) described effectiveness as the extent to which

customers' requirements are met, while efficiency is a measurement of how economically the firm's resources are utilized when providing a given level of customer satisfaction. In contrast, Cooper, Seiford, and Tone (2006) described effectiveness as goal achievement and efficiency as the evaluations of the resources used. The scope of this study is to evaluate efficiency, not effectiveness.

Efficiency is the ratio calculated from input resources and output results, to evaluate whether the use of input resources is effectively employed for the outcome or not (Azadeh, Ghaderi, Miran, Ebrahimipour, & Suzuki, 2007; Ueasin, Liao, & Wongchai, 2015). The objective of efficiency measurement is to detect weak areas so that appropriate efforts can be devoted to improve performance.

Efficiency (cost efficiency or overall efficiency) has two components: technical efficiency (ability to avoid waste by producing as much output, as input usage allows), and allocative efficiency (combining input and output in an optimal proportion based on prices) (Munisamy-Doraisamy, 2004). Overall efficiency (cost efficiency) means the firm must be able to choose the right mix of inputs and use them in a technically efficient manner (Bogetoft & Otto, 2011)

$$OE = TE \times AE, \tag{1}$$

where *OE* is overall efficiency, *TE* is the technical efficiency, AE is the allocative efficiency.

Since allocative efficiency requires price information, this research concerns technical efficiency only. Technical efficiency signifies a level of performance that describes a process which uses the lowest amount of inputs to create the greatest amount of outputs. It is noteworthy that technical efficiencies can be gained by sacrificing quality, since higher quality can be attained by reducing productivity and increasing costs (Sudit, 1996). In the simplest case – where a process or unit has a single input and a single output – technical efficiency is defined as:

$$TE = \frac{y}{x}, \tag{2}$$

where x is input vector, y is the output vector.

Typically, DMUs use multiple inputs and outputs (Boussofiane, Dyson, & Thanassoulis, 1991) and in that case, efficiency is determined as:

$$\text{TE} = TE = \frac{\text{Weighted sum of outputs}}{\text{Weighted sum of inputs}}. \tag{3}$$

Efficient companies take a score of 1, so the efficiency score which is closer to 1 shows better performance. After calculating the efficiency score, inefficiency can be easily determined by subtracting the efficiency score from one. The smaller the inefficiency is, the better the performance is (Bogetoft & Otto, 2011).

1.2. Data envelopment analysis

Efficiency measurement methods can be divided into three main categories: ratio indicators, parametric and nonparametric methods (Vincová, 2005). A significant difference between the parametric and the non-parametric approaches is the estimation method. DEA is a non-parametric approach to weigh the inputs/outputs and to measure the relative efficiency of

DMUs (Ablanedo-Rosas, Gao, Zheng, Alidaee, & Wang, 2010). The idea of DEA was first introduced by Farrell (1957) and developed by Charnes, Cooper, and Rhodes (1978). The general idea of DEA is considering DMUs to have the same technology set. The technology set is the set of outputs that can be produced by using available inputs which takes zero or positive numbers as input and output variables (non-negative). DEA determines two orientations: input efficiency and output efficiency. Input efficiency is appropriate when one is interested in minimizing inputs, and output efficiency when one is interested in maximizing output.

Input efficiency:

$$E^0 = E(\left(x^0, y^0\right); \quad T^* = \min\{E \in R_+ \mid \left(Ex^0, y^0\right) \in = T^*\}, \tag{4}$$

where *(x, y)* means feasibility of the vector, and T^* the smallest set which is consistent with the data.

Input efficiency takes a value between 0.0 and 1.0. For example, a value of 0.6 obtained by the input-oriented method shows the possibility to produce the same output when the inputs are decreased by 40%.

The frontier scale of DEA consists of constant return to scale (CRS) and variable return to scale (VRS). VRS consists of increasing (IRS) and decreasing return to scale (DRS) (Fenyves, Tarnóczi, & Zsidó, 2015). Choosing between DRS and IRS depends on the firm's industry. In this research, the input efficiency VRS model by R statistical program is used.

Decreasing Return to Scale:

$$\left(x, y\right) \in T, \quad 0 \le \lambda \le 1 \Rightarrow \lambda\left(x, y\right) \in = T. \tag{5}$$

Increasing Returns to Scale:

$$\left(x, y\right) \in T, \quad \lambda \le 1 \Rightarrow \lambda\left(x, y\right) \in = T. \tag{6}$$

Although CRS is the basic DEA model, it is appropriate when DMUs freely produce under their optimal size. But the heterogeneity of the data shows that there is not a perfect competition among the DMUs. Therefore, VRS model is chosen which is more realistic in this study. When a dataset contains wide-ranging companies, the importance of the performance measurement maybe questionable, therefore, performance measurement is integrated with clustering.

1.3. Clustering

Clustering plays an essential role in helping people to analyze, describe and utilize the valuable information hidden in the groups (Wu, 2012). Cluster analysis is one of the data mining methods for discovering knowledge in multidimensional data. The primary goal of cluster analysis is to identify pattern or groups of objects within a data set which have high similarity. Clustering techniques are divided into the partitional and hierarchical. Partitional clustering methods directly divide data points into some pre-specified number of clusters without the hierarchical structure. In contrast, hierarchical clustering groups the data with a sequence of nested partitions, either from singleton clusters to a cluster including all of the individuals; or vice versa (Xu & Wunsch, 2008). Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning (Rokach & Maimon, 2010).

Based on the way to approach the center, cluster analyses are classified as: hard (crisp) clustering and soft (fuzzy) clustering (Ho-Kieu et al., 2018). The most common and well-known hard clustering is k-means. The k-means is a simple and fast clustering method. Moreover, a k-means algorithm has the excellent ability to handle a large number of investigated data. The k-means algorithm applies a standard distance measure formula, to determine the similarity of the data repetitively, to obtain the high inter-cluster distance among clusters (Arbin et al., 2016). K-means clustering iteratively finds the k centroids and assigns every object to the nearest centroid (Park & Jun, 2009). The centroids are updated by taking the average of all data. Therefore, if there are outliers in data, the centroids will be pushed to the outliers. Extremely high values might substantially distort the distribution of data, which is the drawback of k-means.

In contrast to k-means, k-medoids uses the most centrally located object in a cluster – instead of center mass – which helps to overcome the k-means' drawback.

1.4. K-medoids algorithm

K-medoids algorithm is computationally harder than that of the k-means due to computing the medoids using the frequency of occurrences. Clustering tendency, which shows whether the clustering is appropriate for the data, must be assessed, before employing a clustering algorithm. Afterwards, the number of clusters and algorithms must be determined. Finally, cluster validation (goodness of clustering results) should be done.

According to Kassambara (2017), the most common algorithm of k-medoids clustering is the Partitioning Around Medoids (PAM) and is as follows:

- Initialize. Randomly select k (number of the cluster) of the n data points as the 'mediods'. Like k-means, k-medoids requires a pre-set number of cluster (k). There is no final approach to determining the number of clusters, but identifying an inappropriate number of clusters can lead to meaningless clusters (which do not exist). A useful approach to assess the optimal number of clusters is the silhouette method (Kassambara, 2017).
- Calculate the dissimilarity matrix. A standard way to express similarity is through a set of distances between pairs of objects (Hartigan, 1989). Data within the group (intra-cluster) are similar, while data between the groups (inter-cluster) are different, based on the specific criteria.
- Assign every object to the closest medoid. The closest medoid is defined by using any valid distance metric, most commonly: Euclidean distance (the root-sum-of-squares of differences), Manhattan distance (the sum of absolute distances) or Minkowski distance. Manhattan is more robust than Euclidean distances when data contains outliers.

If any of the objects of the cluster decreases the average dissimilarity coefficient, select the entity that reduces this coefficient the most as the medoid for this cluster.

Each clustering algorithm creates a different cluster for the same data, therefore, cluster validation is essential to determine whether the clusters are meaningful or just artifacts of the clustering algorithm.

- There are three categories of validation:
- External; used to select a suitable clustering algorithm;

– Internal: measures the compactness (within cluster variation), and the connectedness and the separation (how well-separated) of the cluster partitions;
– Relative criteria (Kassambara, 2017).

2. Data and variables

Mongolian companies are organized as public or non-public. Since public companies' financial reports are required to be audited, their data is more reliable (than that of non-public companies) and is publicly available.

In this research, 89 public companies' financial statements (from 2012 to 2015) were obtained from the MSE (Mongolian Stock Exchange) and used as data. In 2009, the MSE started providing downloadable financial statements; though only nine companies' financial statements were available at that time. In 2010, the number of publicly available financial statements rose dramatically to 100. However, the form of financial statements was changed in 2012, which made it difficult to compare the financial statements before and after 2012. Although there are 334 registered public companies, not all the companies were suitable sources for data. Some companies' financial reports were deducted from research due to bankruptcy, lack of annual reports, and zero values in financial data. Only 137 companies out of 334, reported publicly their financial statements of 2015. The financial statements which are used in the research met the requirements of consistency, comparability, and accuracy. In this paper, total assets and revenue are chosen for k-medoids. Data includes companies in different sectors, including, services, mining, manufacturing, etc. The majority of costs in the service sector are operational costs, while the costs of goods sold were greater in manufacturing. Similarly, current assets constitute more in the service sector and less in the mining sector. Considering the characteristics of sectors, revenue and after-tax profit are chosen as output variables; while non-current assets, current assets, cost of goods sold and operational costs are input variables in the DEA.

The descriptive statistics of each year are attached in Appendix 1. As we can see from Appendix 1, the coefficients of variation are remarkably high (> 200%), especially for the last variable (after-tax profit > 400%). The high values of the coefficients of variation indicate that the variability and heterogeneity are also extremely high. From Appendix 1, it is apparent that the values of the kurtosis indicator are high, positive values, which means that the data are predominating around the average value. The values of the skewness indicator are also positive which indicates the density functions of the variables have longer tails on the right side, and the mass of the distribution is concentrated on the left side. Figure 1 presents the high variability of the investigated variables which can be seen in Appendix 1.

The descriptive statistics of the four years' average values are presented in Table 1. The Interquartile Range (IQR) indicators, which show the range of the middle 50% of the data, were calculated by using the average values of the variables. It is apparent from Table 1 that the IQRs of given variables represent only a fraction of their total ranges (0.72–7.43%). Additionally, more than 90% of the total range of the variables is in the fourth quartile. Overall, the statistical characteristics show similar tendencies to the annual data in Appendix 1. The variables have huge variability, and the more significant part of their total range is located in

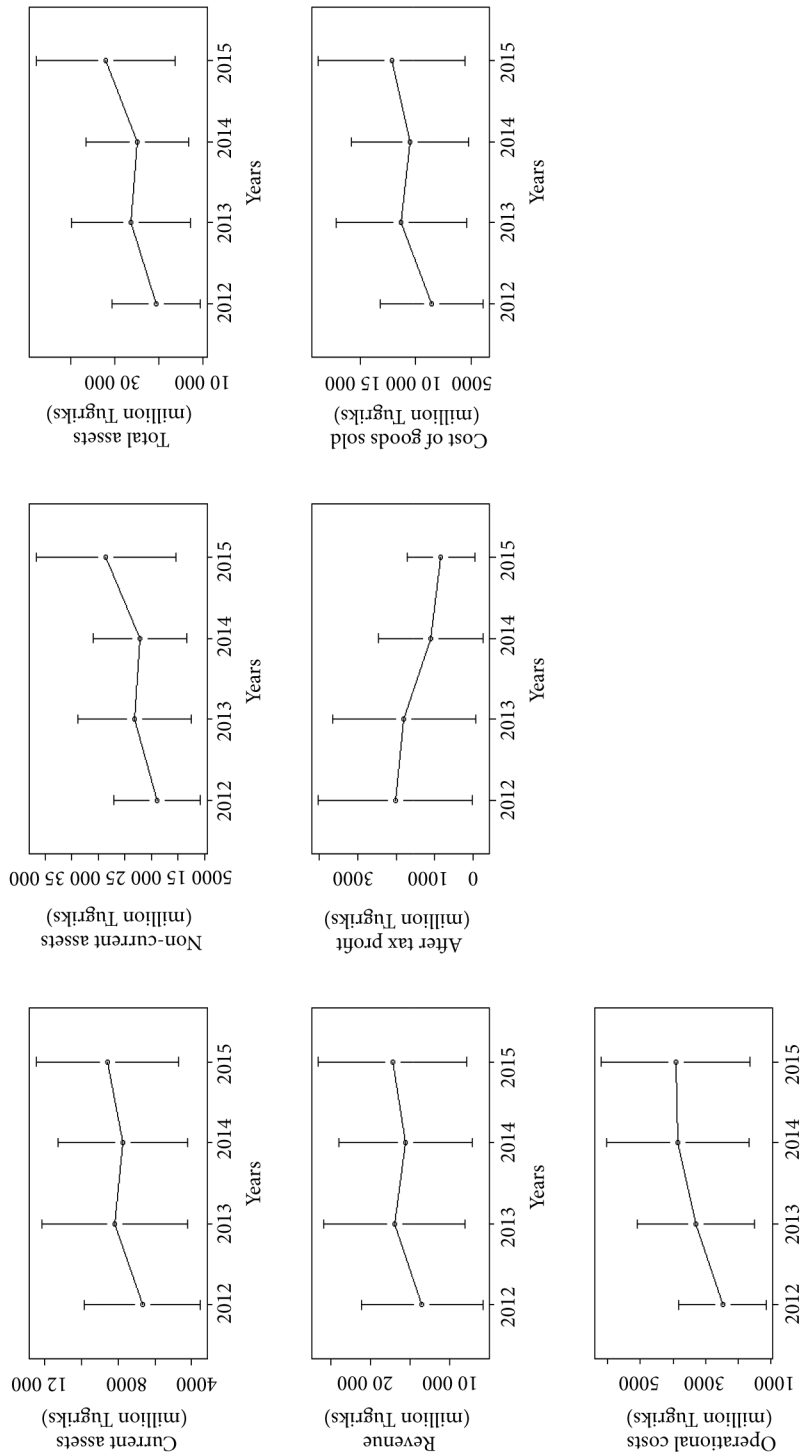


Figure 1. Heterogeneity across years – differences among the companies (source: author's calculation by R Studio)

the fourth quarter regarding revenue. It shows that further analysis could possibly be incorrect without using classification (clustering) method, and it could lead to misinterpretations. It is essential to group them into clusters and use clusters individually for further analysis.

Table 1. Descriptive statistics of variables investigated (source: own calculation by R Studio (average of 4 years)) (million tugriks)

Statistical characteristics	Current assets	Non-current assets	Assets	Revenue	Cost of goods sold	Operational costs	After-tax profit
Minimum	4	5	22	8	0	0	–3,719
1. Quartile	233	534	927	309	104	120	-65
Median	1,129	1,764	4,391	1,419	696	485	17
3. Quartile	7,543	9,693	17,236	7,259	5,064	2,228	328
Maximum	98,399	266,458	290,458	239,834	145,738	56,705	50,955
Total range	98,395	266,454	290,437	239,826	145,738	56,705	54,674
IQR	7,310	9,158	16,309	6,950	4,960	2,107	393
IQR / Total range	7.43%	3.44%	5.62%	2.90%	3.40%	3.72%	0.72%
Mean	7,801	18,204	26,005	15,821	10,629	3,385	1,437
Standard deviation	16,720	45,353	57,482	39,065	25,087	8,956	6,513
Coefficient of variation	214.32%	249.14%	221.04%	246.92%	236.03%	264.58%	453.09%
Skewness	3.36	3.50	3.22	3.81	3.40	4.08	5.79
Kurtosis	12.04	12.75	10.36	16.05	12.32	17.45	37.49

3. Analysis and results

The study applies PAM (Partitioning Around Medoids) algorithm – the most common k-medoids clustering method – to determine clusters; and uses DEA to evaluate the performance of each cluster. Various packages of the R statistical program are used for the analysis; such as "facto extra", "fpc", "cluster", "kmed", and "cluster" packages for k-medoids, while the Benchmarking package is used for DEA. The R statistical program (from R Studio) was used because it provides a more user-friendly platform than the original R software (Gandrud, 2015).

It is worth recalling that the clustering results are not the final results of research in general, but are possible inputs for other calculations. This research aims to integrate the k-medoids' results with DEA. The first step is to determine the clusters by the k-medoids algorithm and the second step is to make the DEA calculations using clusters.

The k-medoids clustering requires the number of clusters to be calculated. The optimal number of clusters was established by the "factoextra" package of R statistical program. For cluster analysis, revenue and total assets were selected as variables.

Appendix 2 shows the cluster validation executed by the silhouette method. When the average silhouette width (asw) is closer to 1, it indicates the object is well clustered. The

dashed vertical lines in the graphs in Appendix 2 indicate the optimal number of clusters by the silhouette method; which determined three clusters in 2012 and two clusters in the other years. Therefore, two clusters are made according to the obtained annual asw values.

The k-medoids cluster analysis was performed for each year and for the average of four years, using the revenue and the total assets as variables. The results of the k-medoids cluster analysis are presented in Table 2. It can be seen from the Table 2 that the number of cluster elements in the first and third years (2012 and 2014) is nearly identical, while in the second and the fourth years (2013 and 2015) the numbers of the cluster elements are equal. However, the table also shows that there are significant differences among the central values of cluster variables in different years. The differences between the medoids of the two clusters are extremely high, 45 to 95 times more in the case of total assets, and 43–212 times more in the case of revenues. Based on Table 2, it can be concluded that small and medium-sized enterprises (SME) are in the first cluster, while the second cluster includes large corporations (big companies).

Table 2. The results of k-medoids cluster analysis (source: own calculation by R Studio)
(million tugriks)

Years	SMEs			Big companies		
	Number of companies	Total assets medoid	Revenue medoid	Number of companies	Total assets medoid	Revenue medoid
2012	79	1,733	863	10	108,883	59,619
2013	84	2,825	1,145	5	267,065	193,716
2014	80	2,831	1,535	9	141,444	65,750
2015	84	4,018	990	5	307,662	210,111
Average	80	2,903	1,332	9	131,264	77,088

The k-medoids cluster analysis was performed by using the average data of the given years. Table 3 illustrates the main statistical characteristics of big companies. It can be seen from Table 3 that the maximum values of the variables, along with their total range values, have been significantly reduced. Not only the minimum value but also the quartile values of the variables are reduced. These decreases are extremely significant in the case of the maximum and total range values.

In Table 3, the skewness and kurtosis indicators were also decreased significantly which means the distributions of the first cluster (SMEs) data are approximate to the normal distribution. Although the standard deviation and coefficient of variation are still high, after-tax profit is decreased. IQR to total range indicator is also improved significantly. These changes in the statistical characteristics confirm the necessity of the separation of companies. Clustering gives possibilities to analyze a more homogeneous database.

Table 4 represents the main statistical characteristics of big companies, which shows changes similar to those in Table 3. However, the changes in Table 4 are more significant for the following indicators: IQR to total range, the coefficient of variant, skewness, and kurtosis. These changes also justify the necessity of clustering. Appendix 3 illustrates the significant differences between the clusters.

Table 3. Descriptive statistics of SMEs (source: own calculation by R Studio (average of 4 years))(million tugriks)

Statistical characteristics	Current assets	Noncurrent assets	Assets	Revenue	Cost of goods sold	Operational costs	After-tax profit
Minimum	4	5	22	8	0	0	–3,719
1. Quartile	151	475	832	272	93	116	–66
Median	1,007	1,289	3,009	792	506	406	7
3. Quartile	3,323	5,961	10,270	3,762	2,977	1,231	165
Maximum	25,911	54,938	64,360	41,538	31,599	14,191	3,152
Total range	25,907	54,933	64,338	41,530	31,599	14,191	6,871
IQR	3,172	5,486	9,438	3,490	2,884	1,116	231
IQR / Total range	12.2%	9.9%	14.6%	8.4%	9.1%	7.8%	3.3%
Mean	3,554	5,758	9,312	5,136	3,798	1,278	121
Standard deviation	5,891	10,218	14,140	9,715	7,485	2,431	822
Coefficient of variation	165.7%	177.4%	151.8%	189.1%	197.0%	190.2%	679.0%
Skewness	2.21	2.80	2.17	2.41	2.44	3.53	–0.34
Kurtosis	4.23	8.38	4.29	4.82	5.05	14.24	7.86

Table 4. Descriptive statistics of Big companies (source: own calculation by R Studio (average of 4 years)) (million tugriks)

Statistical characteristics	Current assets	Noncurrent assets	Assets	Revenue	Cost of goods sold	Operational costs	After-tax profit
Minimum	9,503	7,969	75,033	45,146	15,208	2,228	–2,686
1. Quartile	17,356	110,474	119,977	62,815	43,977	6,103	1,021
Median	51,837	118,780	132,450	80,163	65,261	15,261	6,502
3. Quartile	67,064	171,305	269,686	133,706	98,842	35,521	18,997
Maximum	98,399	266,458	290,458	239,834	145,738	56,705	50,955
Total range	88,896	258,490	215,426	194,688	130,530	54,477	53,642
IQR	49,708	60,831	149,709	70,891	54,865	29,418	17,976
IQR / Total range	55.92%	23.53%	69.49%	36.41%	42.03%	54.00%	33.51%
Mean	45,555	128,833	174,389	110,795	71,344	22,117	13,139
Standard deviation	30,795	79,230	82,554	67,572	41,904	19,615	16,987
Coefficient of variation	67.60%	61.50%	47.34%	60.99%	58.73%	88.68%	129.29%
Skewness	0.23	0.10	0.30	0.82	0.36	0.42	1.08
Kurtosis	–1.53	–1.08	–1.81	–1.01	–1.27	–1.52	–0.07

Tables 3 and 4 explain the descriptive analyses of two clusters. Companies with total assets less than MNT64.3 billion and revenues less than MNT41.5 billion are classified in the first cluster (as SMEs). The second cluster's companies (big companies) earn approximately

108 times more profit than SMEs (on average) which also demonstrates the substantial difference between the two clusters.

Table 5 reveals the financial ratios of two clusters and the whole dataset. SME's return on sales (ROS) ratio is 46.5% higher than that of big companies, which means smaller companies pay greater attention to the cost management than the larger ones. In contrast to the ROS, the value of return on assets (ROA) ratio is about ten times higher in the second cluster than the first cluster which means the efficiency of asset management in larger companies is much better than smaller companies. The return on equity (ROE) ratio shows a nearly threefold difference between the two clusters. For the gross profit margin (GPM) ratio, smaller companies also perform better, but in this case, the difference is much smaller, which may indicate that smaller companies have relatively higher fixed costs. Based on the debt to total assets (DTA) ratio, it can be stated that smaller companies are indebted approximately 18% more than the larger ones which is not a big difference.

Table 5. Financial ratios of the clusters (source: author's calculation (average of 4 years)) (percentage)

Titles	ROS	ROA	ROE	GPM	CATA	DTA
SMEs	16.03	0.81	4.31	34.53	39.49	40.56
Big companies	10.94	8.32	12.52	33.48	31.34	34.48
Total	15.52	1.57	5.14	34.43	38.66	39.94

ROA	Return on assets	ROE	Return on equity
GPM	Gross profit margin	ROS	Return on sales
CATA	Current assets to total assets ratio	DTA	Debt to total assets ratio
OCR	Operational costs to revenue ratio		

As a final step in the analysis, DEA was applied for the entire population and the two clusters separately. The main interest of the study was identifying the effects of cluster-specific analysis influenced the number of effective companies. For better comparability, some of the statistical characteristics of the efficiency coefficients are calculated. The results of DEA are presented in Table 6.

Table 6. Efficiency results before and after clustering (source: author's calculation by R studio (average of 4 years))

Efficiency range	Whole data set	SMEs	Big companies	Sum of clusters
0.0–0.2	2	2	0	2
0.2–0.3	0	0	0	0
0.3–0.4	5	5	0	5
0.4–0.5	1	1	0	1
0.5–0.6	3	3	0	3
0.6–0.7	8	4	0	4
0.7–0.8	11	3	0	3
0.8–0.9	12	14	0	14

End of Table 6

Efficiency range	Whole data set	SMEs	Big companies	Sum of clusters
0.9–1.0	13	16	1	17
1.0	34	32	8	40
Minimum	0.141	0.141	0.937	0.222
1st quartile	0.772	0.811	1.000	0.830
Median	0.903	0.936	1.000	0.943
3rd quartile	1.000	1.000	1.000	1.000
Maximum	1.000	1.000	1.000	1.000
Mean	0.834	0.850	0.993	0.864

Table 6 presents the efficiency coefficients of all companies and the companies of the SMEs and the big companies, by DEA along with its descriptive statistics. Table 6 shows very clearly that the performance coefficients of the investigated companies were changed. In the case of all companies, efficiency scores are calculated using the original data set which consists of four-year average data of 89 companies. In the first cluster, 80 companies (of 89) are used. Before clustering, 34 companies were determined as efficient. In contrast, DEA defined 40 efficient companies after clustering which, is higher by 17.6%. As a result of the clustering, the statistical characteristics improved. Average efficiency also grew slightly. The number of companies performing – under the efficiency coefficient of 0.8 – has been significantly reduced, from 28 to 18. It can be stated – based on the results of Table 6 – that the combined method provides better and valuable performance coefficients for company performance evaluation. By integrating DEA with clustering, improved performance evaluation and homogeneous comparison are achieved, which is consistent with Lemos, Lins, and Ebecken (2005). Based on the research, the combination of k-medoids and DEA is assumed to give more reliable results which are based on the more homogeneous comparison.

Conclusions

This study evaluates the performance of Mongolian companies by integrating cluster analysis with DEA. The research is based on Mongolian listed companies' data; however, the empirical methodology is not limited to Mongolian companies' but is also internationally applicable.

The research consists of two main parts. Initially, the k-medoids clustering method is chosen to reduce the high values of coefficients of the variant. Based on the silhouette method's result, companies are divided into two clusters, the first cluster consists of 80 companies which are considered as SMEs, while the second cluster consists of only nine companies which are big enterprises. Companies in the second cluster (big companies) earn approximately 108 times more profit than SMEs, which proves that there is a substantial difference between the clusters. After clustering, the maximum values of the variables, along with their total range values, have been significantly reduced and the distributions of SMEs' data shifted more, into a normal distribution. Subsequently, each cluster is analyzed and compared by ratio analysis and by DEA.

According to the ratio analysis, the average ROA ratio of big companies (8.32%) is approximately ten times higher than that of SMEs' (0.81%), which indicates that the big companies' asset management is more efficient. Likewise, big companies have a much high value of ROE (12.52%) than SMEs (4.31%). Although big companies have a much higher value of ROA than SMEs, both big companies and SMEs have similar GPM ratios, 33.48%, and 34.53% respectively. In contrast, big companies' ROS ratios are lower (10.94%) than those of SMEs (16.03%). The higher ROS ratio of SMEs' is possibly caused by their operational costs which shows the smaller companies pay greater attention to cost management than the larger ones. SMEs have higher values – than big companies – in other ratios (i.e., GPM, CATA, and DTA), see Table 5.

According to the DEA results, 34 companies out of 89 are determined as efficient before clustering, while 40 companies are efficient after clustering. Clustering also resulted in better statistical characteristics. The number of companies performing within the efficiency range of 0.8 has been significantly reduced, from 28 to 18. In the case of big companies', only one company is inefficient, and the efficiency coefficient is 93.7%. The efficiency score of SMEs was 85%, which is also high. From these results, it is assumed that the combined method provides better performance measurement.

Combining the clustering method with the DEA possibly complements the drawbacks of each method. Clustering results are not the final results; it is possible to use as data inputs for further research. As for DEA, the results are dependent on the database. If the data has too high a standard deviation, it is required to group the data for proper and accurate results, which is also supported by this analysis. Based on the results of the DEA, the clustering of companies has improved their efficiency rating.

The main limitation of the study is that the empirical analysis is restricted to listed companies, which might limit the scope for generalizations about all Mongolian companies. Future research can be extended by either considering both listed and unlisted companies or by covering a longer time period.

References

Ablanedo-Rosas, J. H., Gao, H., Zheng, X., Alidaee, B., & Wang, H. (2010). A study of the relative efficiency of Chinese ports: A financial ratio-based data envelopment analysis approach. *Expert Systems*, *27*(5), 349-362. https://doi.org/10.1111/j.1468-0394.2010.00552.x

Amin, M., Wan-Ismail, W.-K., Abdul-Rasid, S. Z., & Selemani, R. D. A. (2014). The impact of human resource management practices on performance: Evidence from a Public University. *The TQM Journal*, *26*(2), 125-142. https://doi.org/10.1108/TQM-10-2011-0062

Arbin, N., Suhaimi, N. S., Mokhtar, N. Z., & Othman, Z. (2016). Comparative analysis between k-means and k-medoids for statistical clustering. In *Proceedings – AIMS 2015, 3rd International Conference on Artificial Intelligence, Modelling and Simulation* (pp. 117-121). https://doi.org/10.1109/AIMS.2015.82

Arora, P., & Varshney, S. (2016). Analysis of K-means and K-Medoids algorithm for big data. *Procedia – Procedia Computer Science*, *78*, 507-512. https://doi.org/10.1016/j.procs.2016.02.095

Azadeh, A., Ghaderi, S. F., Miran, Y. P., Ebrahimipour, V., & Suzuki, K. (2007). An integrated framework for continuous assessment and improvement of manufacturing systems. *Applied Mathematics and Computation*, *186*(2), 1216-1233. https://doi.org/10.1016/j.amc.2006.07.152

Bi, G., Song, W., & Wu, J. (2014). A clustering method for evaluating the environmental performance based on slacks-based measure. *Computers & Industrial Engineering*, *72*, 169-177. https://doi.org/10.1016/j.cie.2014.03.016

Bogetoft, P., & Otto, L. (2011). *International Series in Operations Research & Management Science: Vol. 157. Benchmarking with DEA, SFA, and R.* Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-1900-6

Boussofiane, A., Dyson, R. G., & Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, *52*(1), 1-15. https://doi.org/10.1016/0377-2217(91)90331-O

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429-444. https://doi.org/10.1016/0377-2217(78)90138-8

Cooper, W., Seiford, L., & Tone, K. (2006). *Introduction to data envelopment analysis and its uses*. Retrieved from http://link.springer.com/content/pdf/10.1007/0-387-29122-9.pdf

Dai, X., & Kuosmanen, T. (2014). Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega*, *42*(1), 179-188. https://doi.org/10.1016/j.omega.2013.05.007

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, *120*(3), 253. https://doi.org/10.2307/2343100

Fenyves, V., Tarnóczi, T., & Zsidó, K. (2015). Financial performance evaluation of agricultural enterprises with DEA method. *Procedia Economics and Finance*, *32*(15), 423-431. https://doi.org/10.1016/S2212-5671(15)01413-6

Gandhi, G., & Srivastava, R. (2014). Analysis and implementation of modified K-medoids algorithm to increase scalability and efficiency for large dataset. *IJRET: International Journal of Research in Engineering and Technology*, *3*(6), 150-153. https://doi.org/10.15623/ijret.2014.0306027

Gandrud, C. (2015). *Reproducible research with R and R studio* (2nd ed.). New York: Chapman & Hall/CRC. https://doi.org/10.1201/b18546

Griffin, J. E. J. P. A. (2011). Bayesian clustering of distributions in stochastic frontier analysis. *Journal of Productivity Analysis*, *36*(3), 275-283. https://doi.org/10.1007/s11123-011-0213-7

Hartigan, J. A. (1989). *Clustering algorithms* (pp. 331-335). John Wiley & Sons, Inc. https://doi.org/10.1002/0471725382.scard

Ho-Kieu, D., Vo-Van, T., & Nguyen-Trang, T. (2018). Clustering for probability density functions by new k-Medoids Method. *Scientific Programming*, *2018*, 7. https://doi.org/10.1155/2018/2764016

Jahangoshai Rezaee, M., Jozmaleki, M., & Valipour, M. (2018). Integrating dynamic fuzzy C-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange. *Physica A: Statistical Mechanics and Its Applications*, *489*, 78-93. https://doi.org/10.1016/j.physa.2017.07.017

Kassambara, A. (2017). *Practical guide to cluster analysis in R: unsupervised machine learning*. STHDA.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical Data Analysis Based on the L 1-Norm and Related Methods. First International Conference*. Elsevier Science Ltd.

Kianfar, K., Ahadzadeh Namin, M., Alam Tabriz, A., Najafi, E., & Hosseinzadeh Lotfi, F. (2017). Hybrid cluster analyzing and data envelopment analysis with interval data. *Scientia Iranica*, *25*(5), 2904-2911. https://doi.org/10.24200/sci.2017.4482

Kim, B., Lee, H., & Kang, P. (2018). Integrating cluster validity indices based on data envelopment analysis. *Applied Soft Computing*, *64*, 94-108. https://doi.org/10.1016/j.asoc.2017.11.052

Lemos, C. A. A., Lins, M. P. E., & Ebecken, N. F. F. (2005). DEA implementation and clustering analysis using the K-Means algorithm. *WIT Transactions on Information and Communication Technologies*, *35*, 321-329.

Masri, M. H. (2013). *Performance measurement systems in service SME: A Brunei case study*. (PhD Thesis). The University of Manchester, United Kingdom.

Mei, J. P., & Chen, L. (2011). Fuzzy relational clustering around medoids: A unified view. *Fuzzy Sets and Systems*, *183*(1), 44-56. https://doi.org/10.1016/j.fss.2011.06.009

Mohammad, S., Zadegan, R., Mirzaie, M., & Sadoughi, F. (2013). Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems*, *39*, 133-143. https://doi.org/10.1016/j.knosys.2012.10.012

Munisamy-Doraisamy, S. (2004). *Benchmarking the performance of UK electricity distribution network operators: A study of quality, efficiency and productivity*. *Quality*. Warwick Business School.

Narayana, G. S., & Vasumathi, D. (2018). An Attributes similarity-based K-Medoids clustering technique in data mining. *Arabian Journal for Science and Engineering*, *43*(8), 3979-3992. https://doi.org/10.1007/s13369-017-2761-2

Neely, A., Gregory, M., & Platts, K. (1995). Performance measurement system design. *International Journal of Operations & Production Management*, *15*(4), 80-116. https://doi.org/10.1108/01443579510083622

Omrani, H., Shafaat, K., & Emrouznejad, A. (2018). An integrated fuzzy clustering cooperative game data envelopment analysis model with application in hospital efficiency. *Expert Systems with Applications*, *114*, 615-628. https://doi.org/10.1016/j.eswa.2018.07.074

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336-3341. https://doi.org/10.1016/j.eswa.2008.01.039

Patel, A., & Singh, P. (2013). New Approach for K-mean and K-medoids Algorithm. *International Journal of Computer Applications Technology and Research*, *2*(1), 1-5. https://doi.org/10.7753/IJCATR0201.1001

Po, R.-W., Guh, Y.-Y., & Yang, M.-S. (2009). A new clustering approach using data envelopment analysis. *European Journal of Operational Research. 1*. https://doi.org/10.1016/j.ejor.2008.10.022

Rokach, L., & Maimon, O. (2010). Chapter 15 – Clustering methods. In *The data mining and knowledge discovery handbook* (pp. 321-352). Springer. https://doi.org/10.1007/0-387-25465-X_15

Sood, M., & Bansal, S. (2013). K-Medoids clustering technique using bat algorithm. *International Journal of Applied Information Systems*, *5*(8), 20-22. https://doi.org/10.5120/ijais13-450965

Sudit, E. F. (1996). *Effectiveness, quality and efficiency: a management oriented approach*. Springer. https://doi.org/10.1007/978-94-009-1828-3

Thakare, S. Y., & Bagal, S. B. (2015). Performance evaluation of K-means clustering algorithm with various distance metrics. *International Journal of Computer Applications*, *110*(11), 975-8887. https://doi.org/10.5120/19360-0929

Ueasin, N., Liao, S. Y., & Wongchai, A. (2015). The technical efficiency of rice husk power generation in Thailand: comparing data envelopment analysis and stochastic frontier analysis. *Energy Procedia*, *75*, 2757-2763. https://doi.org/10.1016/j.egypro.2015.07.518

Vincová, I. K. (2005). Using dea models to measure efficiency. *Biatec*, (1), 24-28.

Wu, J. (2012). *Springer Theses: recognizing outstanding (Phd Research). Advances in K-means Clustering: a data mining thinking*. Springer. https://doi.org/10.1007/978-3-642-29807-3

Xu, R., & Wunsch, D. C. (2008). *Clustering*. Wiley. https://doi.org/10.1002/9780470382776

Zhang, Q., & Couloigner, I. (2005). A new and efficient K-Medoid algorithm for spatial clustering. In O. Gervasi, et al. (Eds.), *Computational Science and Its Applications – ICCSA 2005. ICCSA 2005. Lecture Notes in Computer Science* (vol. 3482, pp. 181-189). Springer. https://doi.org/10.1007/11424857_20

APPENDIX 1

Descriptive statistics of variables (Million tugrik)

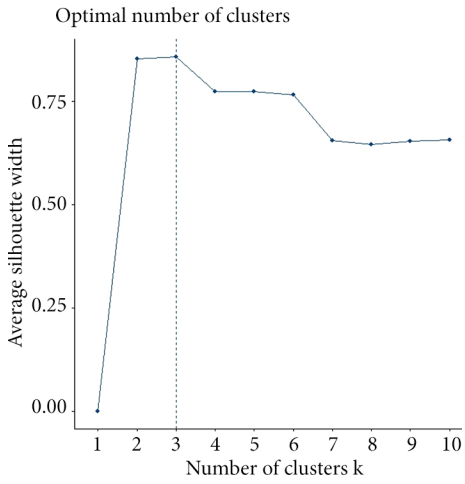
Years	The name of statistical characteristics	Current assets	Non-current assets	Assets	Revenue	Cost of goods sold	Operational costs	After-tax profit
2012	Minimum	0	0	9	0	0	0	–8,840
	1. Quartile	165	420	782	200	19	94	0
	Median	854	1,286	2,154	924	376	422	11
	3. Quartile	5,281	8,306	13,588	8,658	4,353	1,224	453
	Maximum	105,495	267,542	291,633	229,514	142,790	39,669	80,021
	Mean	6,689	13,945	20,633	13,492	8,562	2,478	2,022
	Standard deviation	15,083	38,744	47,659	36,149	21,906	6,306	9,455
	Coefficient of variation	225.50%	277.84%	230.98%	267.92%	255.86%	254.45%	467.67%
	Skewness	4.09	4.58	3.67	4.25	3.92	3.96	6.66
	Kurtosis	20.53	23.50	14.82	19.70	17.33	16.58	49.95
2013	Minimum	3	5	21	2	0	0	–5,259
	1. Quartile	171	492	892	258	96	104	–44
	Median	1,095	1,296	3,390	1,291	800	430	21
	3. Quartile	6,161	10,409	14,928	7,411	6,201	2,029	298
	Maximum	105,217	360,649	402,889	233,290	155,390	48,554	62,096
	Mean	8,185	18,160	26,345	16,966	11,295	3,301	1,792
	Standard deviation	18,880	50,712	63,663	42,024	27,915	8,509	8,849
	Coefficient of variation	230.68%	279.25%	241.65%	247.69%	247.15%	257.79%	493.81%
	Skewness	3.56	4.54	3.87	3.42	3.54	3.81	5.52
	Kurtosis	13.34	23.73	16.38	11.78	12.95	14.58	31.46
2014	Minimum	2	5	23	4	0	0	–14,174
	1. Quartile	168	510	1,000	372	107	117	–41
	Median	1,233	1,692	5,063	1,535	750	441	11
	3. Quartile	7,490	8,858	16,276	9,145	7,355	2,334	463
	Maximum	101,021	233,295	334,316	255,895	154,619	66,141	43,794
	Mean	7,745	17,161	24,907	15,618	10,505	3,849	1,101
	Standard deviation	16,728	41,785	55,186	39,714	25,103	10,424	6,425
	Coefficient of variation	215.98%	243.48%	221.57%	254.29%	238.96%	270.82%	583.78%
	Skewness	3.48	3.43	3.61	4.10	3.77	4.07	4.63
	Kurtosis	13.15	11.88	14.36	18.66	15.82	17.54	26.48

End of Appendix 1

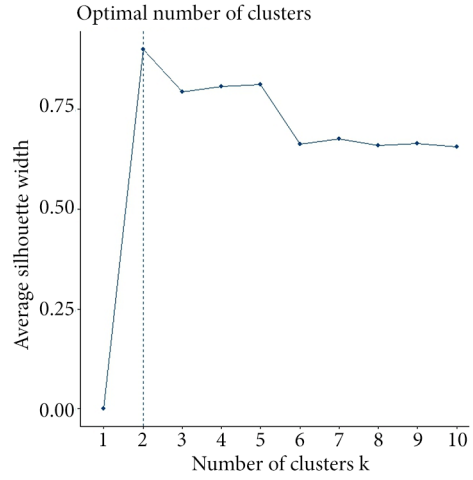
Years	The name of statistical characteristics	Current assets	Non-current assets	Assets	Revenue	Cost of goods sold	Operational costs	After-tax profit
2015	Minimum	1	5	24	0	0	0	–3,505
	1. Quartile	205	587	1,153	234	82	143	–61
	Median	1,386	1,783	4,707	1,053	618	432	22
	3. Quartile	8,326	10,634	19,092	8,444	6,378	2,036	348
	Maximum	88,985	420,903	448,809	240,637	191,124	72,455	30,072
	Mean	8,586	23,549	32,135	17,208	12,154	3,912	835
	Standard deviation	18,465	62,275	74,707	44,021	31,411	10,846	4,128
	Coefficient of variation	215.06%	264.44%	232.48%	255.83%	258.44%	277.27%	494.11%
	Skewness	3.24	4.04	3.50	3.58	3.77	4.32	5.51
	Kurtosis	10.26	18.83	13.02	12.86	15.18	20.19	32.52

APPENDIX 2

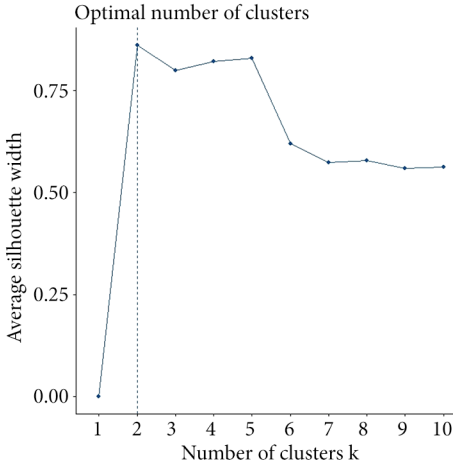
Determination of the optimal cluster numbers by silhouette method



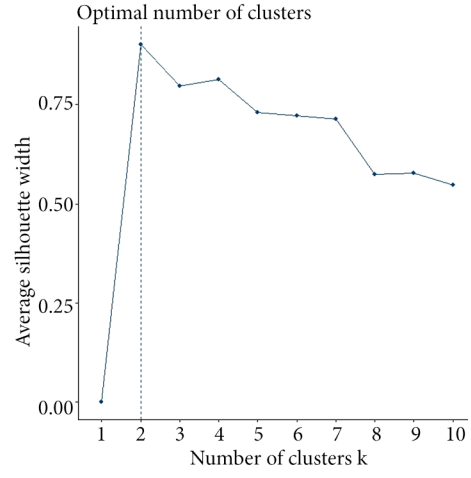
Optimal cluster number in 2012



Optimal cluster number in 2013



Optimal cluster number in 2014



Optimal cluster number in 2015

APPENDIX 3

The clustering plot