

Use of data envelopment analysis for clustering of decision making units

Hassanali Jami*

*Department of Mathematics
Khash Branch
Islamic Azad University
Khash
Iran
jami hassanali@gmail.com*

Sharad Gore

*Department of Statistics and Computer Sciences
Savitribai Phule
Pune University
Pune
India*

Abstract. In this paper, clustering based on data envelopment analysis (DEA) of eight educational groups, with each group having four inputs and outputs, as well as DMU_k ID, is specified. Minitab software and optimal weights for the formation of clusters of DMUs were used for the analysis. The statistical descriptions of DMUs showed, that the elementary educational group (DMU₂) with the best educational performance ranked first. The correlation between the variables of different DMUs was also investigated. In an analysis of the main variables, only three items of eigenvalue were greater than one. In clustering based on DMUs as highlighted in green, with the four clusters, more than 97% of variables could be retained. Based on observations as shown in red, considering the three clusters, more than 78% of information could be maintained and in this stage the abnormal variable of O_3 (scientific products) was removed from the process. Regarding the analysis, the results of path coefficients showed that DMU₂ had seen the best economy performance.

Keywords: data envelopment analysis, clustering, decision making units, cluster analysis, optimal weights.

1. Introduction

Data envelopment analysis (DEA) is a useful method for the clustering of variables of decision-making units (DMUs). DEA has been widely studied and applied in various areas for 38 years since [1] first proposed the method with the CCR model. One of the main forms of DEA models and their extensions includes those of the BCC model [2]. DEA is being used extensively by various researchers in different fields of interests including DMU clustering.

*. Corresponding author

Cluster analysis is a branch of statistical multivariate analysis and unsupervised learning in pattern recognition [3],[4],[5]. It is a method for classifying groups of variables set into the same cluster but not groups set into different clusters. The method of clustering is as old as science itself. Clustering is the task of grouping similar objects into clusters. The process involves classifying existing variables in such a way that the variation in the same-group variables is as low as possible while it is very high between groups.

More accurately, clustering can be defined as partitioning the variables set into subgroups called clusters so that variables in the same cluster are more similar to each other than variables in other clusters. Cluster analysis enables the analyst to find the underlying structure of variables and come up with hypotheses that would answer many questions regarding the variables. This study considers optimal weights corresponding to DMUs. It is thus obvious that every DMU will be represented by a weight vector. This provides us with a multidimensional representation of DMUs in the given problem. These weight vectors can then be used to form clusters of the DMUs. The use of these clusters will be used to treat all the DMUs separately. This paper aims to investigate clustering based on the DEA of educational groups from Islamic Azad University, Khash Branch, Iran. Based on the obtained results of the group's performance, it is necessary and compulsory to revise some of the input variables that affect the output results of each DMU of the university's Khash Branch in the future. It is important to find the best DMUs for managerial decision-making where decision-makers are interested in knowing the required changes in input resources to get the desired result.

The rest of this paper is organized as follows: Section 2 provides a survey of the relevant literature and introduces the DEA-related works that have been used for clustering of DMUs. Section 3 describes the related methodological framework, presents the data used in our study, the empirical results of analysis, and discussion of the clustering based on DEA. Finally, the conclusion is stated in Section 4.

2. Related work

In the clustering literature, several authors have discussed the robustness for clustering [6], [7]. The robustness of the DEA-based clustering algorithm is another interesting research issue. Some notable works include: Kaufman and Rousseeuw's [4] "finding groups in data: an introduction to cluster analysis" and Dave and Krishnapuram's [7] "Robust clustering methods: a unified view".

Other works include: McLachlan and Basford's [8] "Mixture models: inference and applications to clustering" and Wu and Yang's [9] "Alternative c-means clustering algorithms". The clustering is done by assigning different efficiency ratio grades [10],[11]. Generally, clustering methods can be divided into the following categories: hierarchical clustering [4] is a category of cluster analysis. K-means clustering algorithm [3],[9],[12] include clustering procedures that min-

imize dissimilarities and can also be considered as a feature analysis technique. In port cluster resources based on DEA model [13] and in a new clustering approach using DEA [14], using mixed DEA cluster model [15]. Besides, there are data envelopment-based clustering approach for public sugar factories in the privatizing process [16], clustering of DMUs using full-dimensional efficient facets (FDEFs) of PPS with BCC technology [17], and weight-based clustering in DEA [18], efficiency evaluation using DEA game, and cluster analysis [19]. In this study, in contrast to the previous literature on DEA, we present a new approach.

3. Research methodology

Several methods have been proposed to obtain the meaningful classification of variables. The DEA method can be used to cluster the variables with input and output items of the DMUs. In this study, we propose a clustering approach based on the input and output variables of educational groups from Islamic Azad University, Khash Branch, Iran, using machine learning, which involves DEA. The proposed DEA-based clustering approach employs machine learning techniques derived from the DEA method for clustering the observed variables with input and output items of all DMUs.

3.1 Data envelopment analysis

3.1.1 Theoretical foundation

Let there be n DMUs, $DMU_i, i = 1, 2, \dots, n$. Every DMU has m inputs and s outputs. The m inputs are denoted by X_1, X_2, \dots, X_m and the vector $\underline{X} = (X_1, X_2, \dots, X_m)$ is called the input vector. The s outputs are denoted by Y_1, Y_2, \dots, Y_s and the vector $\underline{Y} = (Y_1, Y_2, \dots, Y_s)$ is the output vector. These together give a vector of dimension $m+s$. Denoted by $\underline{T} = (\underline{X}, \underline{Y}), X \geq 0, Y \geq 0$, this vector is called the production vector.

3.1.2 Model M

Suppose there are n DMUs, each having m inputs, and are denoted by vector $\underline{X} \geq 0$ and the s outputs are denoted by vector $\underline{Y} \geq 0$. Therefore, the m coefficients $\underline{V} = (V_1, V_2, \dots, V_m), \underline{V} \geq 0$ for inputs and s coefficients $\underline{U} = (U_1, U_2, \dots, U_s), \underline{U} \geq 0$ for outputs are to be determined by solving the following constrained optimization problem.

Model M:

$$\max \frac{\sum_{r=1}^s u_r y_{ro}}{\sum_{i=1}^m v_i x_{io}}$$

Subject to the constraints:

$$(1) \quad \begin{aligned} \frac{\sum_{r=1}^s u_r y_{rj}}{\sum_{i=1}^m v_i x_{ij}} &\leq 1 && \text{for } j = 1, 2, \dots, n \\ u_r &\geq 0 && \text{for } r = 1, 2, \dots, s \\ v_i &\geq 0 && \text{for } i = 1, 2, \dots, m \end{aligned}$$

where u_r is the weight of the output y_{ro} , and r ($r = 1, 2, \dots, s$) is the number of the generated products, y_{ro} - the amount of the product r is generated by DMU_o ($o = 1, 2, \dots, n$), x_{io} - the amount of the resource i is used by DMU_o , v_i - the weight of the input x_{io} , i ($i = 1, 2, \dots, m$) is the number of the resources used, y_{rj} - the amount of the product r is generated by DMU_j , x_{ij} - the amount of the resource i is used by DMU_j , and j ($j = 1, 2, \dots, n$) is the number of DMUs.

To provide the optimal solution, model M produces vectors of optimal coefficients for all the n DMUs in the data. The optimal coefficient matrix is formed by treating these vectors as its rows. This matrix is denoted by T and can be written as follows:

$$(2) \quad T = \begin{bmatrix} v_{11}, & v_{12}, & \cdots & v_{1m}, & u_{11}, & u_{12}, & \cdots & u_{1s} \\ v_{21}, & v_{22}, & \cdots & v_{2m}, & u_{21}, & u_{22}, & \cdots & u_{2s} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{j1}, & v_{j2}, & \cdots & v_{jm}, & u_{j1}, & u_{j2}, & \cdots & u_{js} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{n1}, & v_{n2}, & \cdots & v_{nm}, & u_{n1}, & u_{n2}, & \cdots & u_{ns} \end{bmatrix}$$

The n rows of T correspond to n DMUs, the first m column of T corresponds to m inputs, and the last s column of T corresponds to s outputs. There are many distance functions that can be used for this purpose.

3.2 Clustering of decision making units

The first step toward the formation of clusters is to construct the distance matrix. It is a square symmetric matrix of order $n \times n$, where the $(i, j)^{\text{th}}$ element denotes the distance between the i^{th} row and the j^{th} row of matrix T . The distance function is a non-negative, symmetric, reflexive, and transitive function that may desirably satisfy triangle inequality.

There are some s standard distance functions that are commonly used because they are meaningful. Euclidean distance, Manhattan distance, binary distance, Minkowski distance, and Canberra distance are some of the commonly used distance functions. Matrix D can be written as $D_{nn} = ((d_{ij}))$, where

$$(3) \quad d_{ij} = d(\underline{T}_i, \underline{T}_j) \quad \text{for } i, j = 1, 2, \dots, n.$$

Here d is the distance function and $d(\underline{T}_i, \underline{T}_j)$ denotes the distance between the i^{th} and j^{th} rows of the matrix T .

3.2.1 Data analysis

Clustering is one of the most important topics, and in peer groups its classification has many applications. In this section, using the data obtained from the research titled “Assess the relative efficiency of educational groups of Islamic Azad University Khash Branch” in academic year 2011-12, eight educational groups are specified, with each academic department having four inputs (I_i) and four outputs (O_j), and DMU_k ID ($i, j = 1, \dots, 4; k = 1, \dots, 8$). In order to evaluate DMUs regarding all the inputs and outputs of the eight DMUs using the clustering method, Minitab 17 statistical software was used. As mentioned above, the inputs and outputs of the DMUs and their values in this study are presented in Tables 1 and 2.

Table 1: Inputs and outputs for each DMU & field of study.

Inputs: I_i	Outputs: O_j	DMU _k	Field of study
I_1 : Head of department	O_1 : Number of students	DMU ₁	Persian literature
I_2 : Specialty books	O_2 : Number of graduates	DMU ₂	Elementary Educational
I_3 : Faculty members	O_3 : Scientific products	DMU ₃	Management
I_4 : Area of department	O_4 : Accepting in higher academic level	DMU ₄	Geography
		DMU ₅	Civil engineering
		DMU ₆	Computer
		DMU ₇	Electro-technique
		DMU ₈	Physical education

Table 2: The values of inputs and outputs for each of educational groups.

DMU _k	Score of HOD	Number of library books	Faculty members	Area of department	Number of students	Number of graduates	Scientific product	Accepting in higher academic level
DMU ₁	2.973	2000	12	20	298	54	0	65
DMU ₂	2.671	3300	6	40	250	126	0	65
DMU ₃	2.590	3000	5	20	431	31	0	65
DMU ₄	2.806	2000	5	20	125	10	3	20
DMU ₅	2.500	2000	2	20	29	4	0	75
DMU ₆	2.806	1500	4	20	77	9	0	50
DMU ₇	2.447	1000	3	30	35	8	0	35
DMU ₈	2.198	700	5	20	33	10	1	35

3.2.2 Description

By applying the described process on the eight educational groups of the DMUs, the average allocation to each of these groups was used. These descriptions are provided in Table 3 for the different DMUs.

Table 3: Descriptive statistics of DMUs.

DMU _k	DMU ₁	DMU ₂	DMU ₃	DMU ₄	DMU ₅	DMU ₆	DMU ₇	DMU ₈
Mean	20.91	31.70	17.21	9.15	14.31	12.18	11.12	10.40
St. Dev	24.75	44.40	21.88	7.10	25.31	16.43	13.50	11.62

The mean of the obtained values in Table 3 show that the maximum amount related to DMU₂ is allocated to the elementary education group. This shows that the DMU₂ has the best academic performance due to input variables compared to other first-ranked DMUs. DMU₁ is ranked second after DMU₂ shows. However DMU₄ did not give good performance in comparison with other DMUs.

3.2.3 Correlation

To find out the relationship and the correlation between the input and output of each DMU, variables related to the analysis and a covariance matrix are used as measurement variables.

This matrix is symmetric, hence a bottom triangle. Since the diagonal matrix is always 1, the same variables have a value of 1 and have a full relationship. Thus, there is complete solidarity and a significant amount. Therefore, it is automatically eliminated from the analysis of the matrix. Thus, it seems that the obtained correlation matrix of Table 4 shows a matrix of 7×7 .

Table 4: Correlation and significance.

C \ R	I ₁	I ₂	I ₃	I ₄	O ₁	O ₂	O ₃
I ₂	0.422 0.298						
I ₃	0.582 0.130	0.208 0.621					
I ₄	-0.051 0.905	0.384 0.348	-0.048 0.910				
O ₁	0.435 0.282	0.766 0.027	0.563 0.146	0.071 0.867			
O ₂	0.288 0.489	0.709 0.049	0.459 0.252	0.762 0.028	0.533 0.173		
O ₃	0.058 0.891	-0.157 0.711	-0.044 0.917	-0.269 0.519	-0.206 0.625	-0.276 0.509	
O ₄	0.165 0.696	0.573 0.138	0.177 0.675	0.111 0.793	0.444 0.271	0.420 0.300	-0.757 0.030

In each cell of this matrix, two values are visible showing the higher correlation criterion, and the lower amount shows the significance level of the test. In addition, that the calculated P value of the test is less than or equal to 0.05 shows a high correlation between the respective variables. According to the above matrix, the values marked in purple represent a high correlation between variables.

3.2.4 Analysis of the main variables

Based on the correlation matrix analysis, the eigenvalues in Table 5 are marked as numbers for eight variables. It should be noted that the eigenvalue should be more than 1 to have enough credibility. According to Table 5, there are only three eigenvalue items that are greater than one.

Table 5: Correlation and significance.

Principal Component Analysis	Eigen analysis of the Correlation Matrix							
Eigenvalue	3.5834	1.6777	1.2549	0.7377	0.4981	0.2277	0.0204	-0.000
Proportion	0.448	0.210	0.157	0.092	0.062	0.028	0.003	-0.000
Cumulative	0.448	0.658	0.814	0.907	0.969	0.997	1.000	1.000

According to Figure 1, a part from the first three clusters, the rest of the clusters have gradually become worthless.

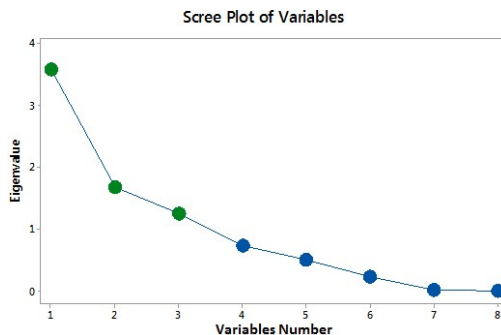


Figure 1: Scree plot for eigenvalue of main variables.

3.2.5 Clustering

Cluster analysis is a process in which observations in a sample are divided into groups so that observations in each group are similar or close to each other. This division is based on values of the observed variables, and these variables can include input and output variables in a DEA problem. Clustering is achieved by calculating the distance matrix of order $n \times n$ or correlation matrix of order

$n \times d$. Each row describes an index. The algorithm output can be grouped as indicators in distinct sets or a tree hierarchical clustering to divide variables.

3.2.6 Clustering based on educational groups

A description of this clustering and the steps using the Minitab software are presented in Table 6.

Table 6: Cluster analysis of groups.

Correlation Coefficient		Distance	Complete Linkage		Amalgamation Steps		
Steps	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	7	99.4184	0.011632	5	6	5	2
2	6	99.0167	0.019666	4	7	4	2
3	5	97.4019	0.051962	5	8	5	3
4	4	97.2872	0.054257	1	3	1	2
5	3	89.1056	0.217888	4	5	4	5
6	2	84.5944	0.308111	1	2	1	3
7	1	66.8580	0.662841	1	4	1	8

As highlighted with green in above Table 4, with the adoption of the four clusters, more than 97% of variables can be retained. So, with respect to Figure 2, on the basis of similarities that can be determined for the variables, DMUs, DMU₁ and DMU₃ are located in a cluster. DMU₄ and DMU₇ adopted a common feature and together constitute a cluster.

DMU₅, DMU₆, and DMU₈ have formed a cluster with the highest subscription. Owing to the unique nature of DMU₂, this group is in a separate cluster. The clustering of the groups based on the criteria of input and output is shown in Figure 2.

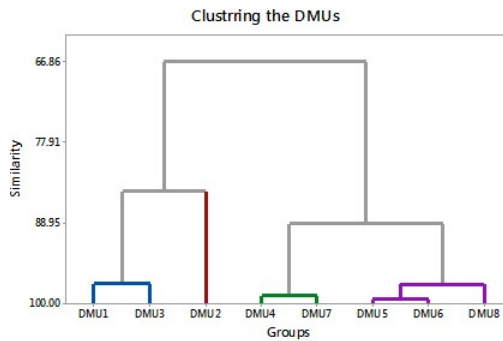


Figure 2: Clustering main variables based on educational groups.

3.2.7 Clustering based on observations

As evident in Table 7, this output provides a structure of cluster analysis variables.

Table 7: Cluster analysis of observations.

Euclidean Distance		Complete Linkage		Amalgamation Steps		Number of obs. in new cluster	
Steps	Number of clusters	Similarity level	Distance level	Clusters joined	New cluster		
1	7	95.7546	110.88	4	5	4	2
2	6	89.5105	273.97	1	4	1	3
3	5	88.5067	300.19	7	8	7	2
4	4	86.0799	363.58	2	3	2	2
5	3	78.9891	548.78	1	6	1	4
6	2	49.1619	1327.82	1	7	1	6
7	1	0.0000	2611.87	1	2	1	8

According to Table 7, the criterion of similarity level of this process at level 110 arrives from 95.75 in the first stage of the process to the amount of 78.99 at level 548.78 in the fifth stage and then this with a huge disparity in the sixth stage reached almost 49. Therefore, as shown in Table 7 in red, considering the three clusters, more than 78% of information can be maintained. This amount can be a desirable value for selecting the three clusters from this clustering. The results of the clustering are shown in Figure 3.

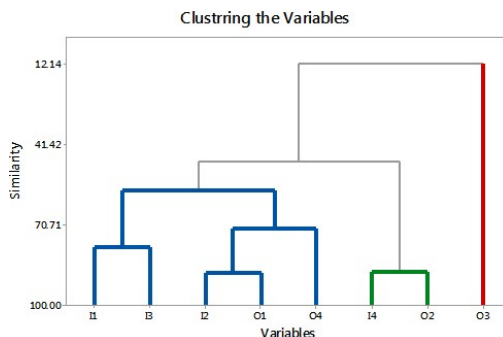


Figure 3: Clustering of main variables based on observations.

According to Figure 3, variables I_1 , I_2 , I_3 , O_1 , and O_4 are in the same blue cluster, the variables I_4 and O_2 are in a green cluster, and O_3 alone is in a red cluster. Of the five variables that have been identified in distinct blue clusters, I_1 and I_3 have the same structure and characteristics of variables I_2 , O_1 , and O_4 . They are also similar in terms of features and structure to fit together in a group. Moreover, according to the clustering of variables plot shown in Figure

3, variable O_3 is clearly not compatible with other variables, which may be due to a lack or shortage of variables. Figure 4 shows a lack of consistency of O_3 variable. Therefore, to optimize the model and prevent skewness, O_3 is deleted from all educational groups to arrive at actual results.

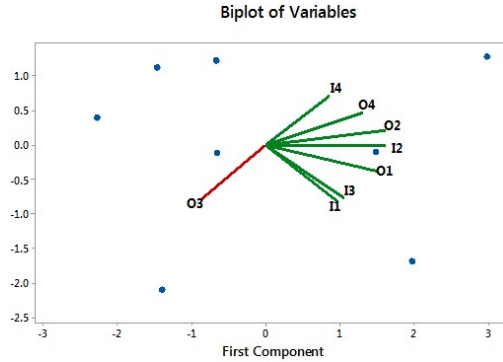


Figure 4: Clustering of main variables based on observations.

3.2.8 Structural equation modeling

The results of the various paths to connect the variables are shown in Figure 5.

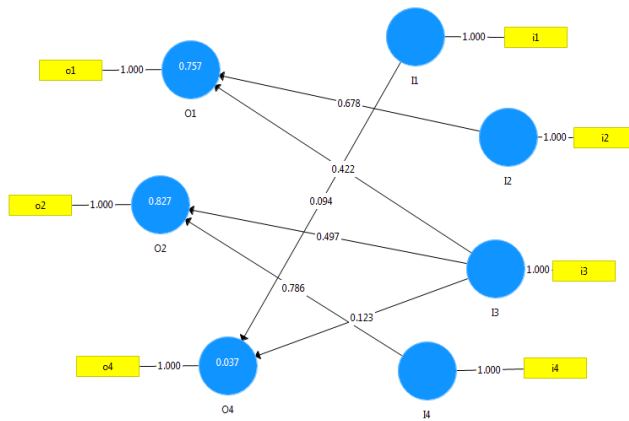


Figure 5: Different possible paths for the relationships between different variables.

In the above figure, different paths may be specified for different variables' relationships and it has been shown how and with what relationship and amount each of the various inputs (I_i) is connected with the different outputs (O_j).

3.2.9 Analysis of path coefficients

According to the plot, the values of the path coefficients are highly desirable in most cases. This is because the route from I_2 to O_1 has the coefficient of 0.678, the route from I_3 to O_1 has the coefficient of 0.422, the variable I_3 to O_2 has the coefficient of 0.497, and the route from I_4 to O_2 has the coefficient of 0.786. Thus, the relationship between I_4 to O_2 shows the highest correlation.

Instead, the correlation between I_1 and O_4 has the minimum value of 0.094. The results of the variables and interpretation show a good fit of the coefficients.

3.3 Results and discussion

In order to investigate the performance of DMUs from Islamic Azad University, Khash Branch, using data analysis, and to compare the values described in Table 3, the main focus is on DMU_2 . Table 4 shows the state of the correlation coefficient and the significance level of the different DMUs. In an analysis of the main variables, based on Table 5, only three items of eigenvalue are greater than one, therefore, we can choose only three clusters, as Figure 1 shows.

In a cluster analysis of DMUs, based on the green-highlighted part in Table 6 and with the adoption of the four clusters, more than 97% of data can be retained, but in this stage with respect to Figure 2, the DMU_2 is in a separate cluster. Therefore, in clustering based on observations regarding the data marked in red in Table 7, more than 78% of information can be maintained by considering the three clusters, because in this stage, regarding Figure 3 and Biplot 4, the variable O_3 is eliminated from all DMUs.

Finally, Figure 5 shows that the path values of I_4 to O_2 with the coefficient of 0.786 is the maximum. Instead, the relationship between I_1 and O_4 has the minimum value of 0.094. According to the results of data interpretation, the model in Figure 5 shows good utility.

4. Conclusions

According to the results of this study, DMU_2 and DMU_1 have been ranked as first and second, respectively, and have shown the best performance. The other DMUs need to be revised and changed to accommodate some effective input variables, including faculty members, for not showing a good performance. In Figure 3, O_3 was eliminated from the cluster. Figure 5 shows that the path coefficients of I_4 to O_2 with a value of 0.786 has the highest correlation. In contrast, I_1 to O_4 is associated with the minimum value (0.094) and has the lowest correlation. The model in Figure 5 shows good utility.

Acknowledgements

I am grateful for anonymous referees for their constructive comments that improve the quality of this manuscript. I have produced this work at my own expense because it is a compulsory requirement for annual promotion.

References

- [1] A. Charnes, W.W. Cooper, and E. Rhodes, *Measuring the efficiency of decision making units*, Eur. J. Oper. Res., 2 (1978), 429-444.
- [2] R.D. Banker, A. Charnes, and W.W. Cooper, *Some models for estimating technical and scale inefficiency in data envelopment analysis*, Manage Sci., 3 (1984), 1078-1092.
- [3] R.O. Duda, and P.E. Hart, *Pattern classification and scene analysis*, NY, USA, Wiley- Intersci. Publ., 1973, 271-273.
- [4] L. Kaufman, and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, NY, USA, Wiley, 1990.
- [5] A.K. Jain, R.P.W. Duin, and J. Mao, *Statistical pattern recognition: a review*, Ieee T. Pattern Anal., 22 (2000), 4-37.
- [6] J.M. Jolion, P. Meer, and S. Bataouche, *Robust clustering with applications in computer vision*, Ieee T. Pattern Anal., 13 (1991), 791-802.
- [7] N. Dave, and R. Krishnapuram, *Robust clustering methods: A unified view*, Ieee T. Fuzzy Syst., 5 (1997), 270-293.
- [8] G.J. McLachlan, and K.E. Basford, *Mixture models: inference and applications to clustering*, Marcel Dekker, New York, 1988.
- [9] W. Kuo-Lung, and Y. Miin-Shen, *Alternative c-means clustering algorithms*, Pattern Recogn., 35 (2002), 2267-2278.
- [10] G.R. Jahanshahloo, F. Hosseinzadeh Lotfi, N. Shoja, G. Tohidi, and S. Razavyan, *A one-model approach to classification and sensitivity analysis in DEA*, Appl. Math. Comput., 169 (2005), 887-896.
- [11] W.D. Cook, and K. Bala, *Performance measurement and classification data in DEA: Input oriented model*, Omega-Int. J. Manage S., 35 (2007), 39-52.
- [12] X. Wang, Y. Wang, and L. Wang, *Improving fuzzy c-means clustering based on feature weight learning*, Pattern Recogn. Lett., 25 (2004), 1123-1132.
- [13] Z. Li, X. Zhao, W. Zhang, and B. Wang, *Analysis on integration efficiency of port cluster resources based on DEA model*. In: IEEE International Conference on service Operations and Logistics, and Informatics, 2 (2008), 2604-2608.

- [14] R.W. Po, Y.Y. Guh, and M.S. Yang, *A new clustering approach using data envelopment analysis*, Eur. J. Oper. Res., 199 (2009), 276-284.
- [15] W. Yawei, and L. Baosong, *Efficiency evaluation of city circular economy based on the super-efficient mixed dea cluster model*, In: International Conference on Management and Service Science, (2010), 1-4.
- [16] E.A. Demirtas, *A data envelopment-based clustering approach for public sugar factories in privatizing process*, Math. Probl. Eng., (2011), 11.
- [17] G.M.R. Moazami, and A.M.R. Jaber, *Clustering decision making units (DMUs) using full dimensional efficient facets (FDEFs) of PPS with BCC technology*, Appl. Mathl. Ski., 6 (2012), 1431-1452.
- [18] L.A. Alves, and J.C.C.B.S. Mello, *Weights based clustering in data envelopment analysis using kohonen neural network: an application in brazilian electrical sector*, Ieee Lat. Am. T., 13 (2015), 188-194.
- [19] L.G. Machado, J.C.C.B.S. de Mello and M. C. Roboredo, *Efficiency evaluation of brazililian electrical distributors using DEA game and cluster analysis*, Ieee Lat. Am. T., 14 (2016), 4499-4505.

Accepted: 11.02.2019