Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 687–696

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Application of DEA and Group Analysis using K-means; compliance in the context of the performance evaluation of school networks

Lorran Santos Rodrigues[a,*], Marcos dos Santos[a], Claudio de Souza Rocha Junior[b]

[a]*Instituto Militar de Engenharia (IME), Rio de Janeiro, RJ*
[b]*Centro de Análises de Sistemas Navais (CASNAV), Rio de Janeiro, RJ*

Abstract

Data Envelopment Analysis (DEA) is already established in the literature for presenting itself as a robust and practical method for efficiency analysis. The present work aims to use the DEA in conjunction with data mining techniques such as K-media and principal component analysis, to find homogeneous groups, within a dataset aiming to ensure a DEA that respects different business realities, within the context of student enrollment in the private school sector in the state of Rio de Janeiro, Brazil.

Keywords: DEA ; K-means; Education;

1. Introduction

Problems involving efficiency analysis, to support decision-making have wide relevance in the context of today's organizations. According to Farrel [1], "it is important to know how far a particular industry can increase its production simply by increasing its efficiency." Because of this, several tools are developed to provide managers with the proper support in their activities, in addition to obtaining these estimates. Given this scenario, operational research and many of its subfields thrive, since they are most concerned about making good business decisions. Some previous work studies the use of Multi-Criteria Decision Methods (MCDM) in real-world problems [2], [3], [4]; also the concerning of making such methods available to the general public [5]. Among them, Data Envelopment Analysis (DEA) presents itself as a robust tool with a wide range of uses [6]. The DEA proposes to obtain relative efficiency between two or more production units that use multiple inputs to produce goods and/or services (outputs). However, when there is a high degree of heterogeneity in the productive units to be analyzed, some of the quality of the proposed analysis is lost, since the commutations were being made between widely different items. Taking this fact into consideration, the present work proposes to combine another technique, based on data mining, with DEA, to obtain more homogeneous groups among the productive units. This hybrid solution is inspired by some recent work concerning multi-criteria decision methods bagging [3], [7]. In another

*Corresponding author. Tel.: +55-21-2416-1473.
Email address: lorran.rodrigues@ime.eb.br (Lorran Santos Rodrigues)

10.1016/j.procs.2022.01.085

aspect, the current work seeks to apply the combined techniques in the context of school enrollments in a school network, which has units spread around the State of Rio de Janeiro. Therefore, by identifying the most efficient schools in capturing enrollment in each group, it is possible to map best practices and understand behaviors, taking into account the socio-spatial reality in which each one is. This approach is relevant, especially in an atypical year like 2020, which had its socioeconomic dynamics severely shaken due to the COVID-19 pandemic. This adverse scenario was well discussed in several publications in recent time [8] . It is quite noticeable that the global pandemic disjointed industry efficiency comparison methodologies that take, for example, the performance history into consideration.

2. Theoretical Framework

2.1. Data Envelopment Analysis (DEA)

Data Envelopment analysis was introduced by Charnes, Copper and Rhodes extending Farrell's concept [1] to estimate efficiency by comparing each production unit with the efficient production frontier [6]. In this context, the DEA consists of a non-parametric method that determines the efficiency curve through linear programming, not requiring the specification of any functional relationship between the inputs and outputs. However, since it is deterministic, this technique is susceptible, to the prevalence of extreme observations (outliers), and errors of measurements.

Generally speaking, the approach applied to the efficiency measure is guided by the definition of Pareto-Koopmans, which elucidates: a) no output (goods and/or services) can be increased without any other output being reduced or some input (input) is increased, or b) none of the inputs can be reduced without any other input being increased or some output is reduced [9].

At first, the model proposed in 1978 by Charnes, Copper and Rhodes, known as CCR, sought to adapt analysis with constant scale returns. The model was then extended by Banker, Charnes, and Cooper in 1984 to include variable returns at scale. As a result, one has the VRS model [10]. Both models can be oriented in two ways to maximize efficiency:

- Reduce the consumption of inputs, maintaining the level of production
- Increase outputs by preserving input levels

The greatest limitation of the mathematical structure of classical models stems from the possibility of generating null weights for important variables, thus making the model incongruous with observed reality.

2.2. DEA CCR Mathematical Formulation

Suposing that are n business units (BU) where each BU has m inputs and s outputs. The relative efficiency of BU_o (in which o in the interval 1,2,...,n) is calculated by solving the fractional programming problem below.

$$\max w_0 = \frac{u_1 y_{1o} + u_2 y_{2o} + ... + u_s y_{so}}{v_1 x_{1o} + v_2 x_{2o} + ... + v_m x_{mo}} \tag{1}$$

Subject to:

$$\frac{u_1 y_{1j} + u_2 y_{2j} + ... + u_s y_{sj}}{v_1 x_{1j} + v_2 x_{2j} + ... + v_m x_{mj}} \leq 1 (j = 1, 2, ..., n) \tag{2}$$

$$u_1, u_2, ...u_s \geq 0 \tag{3}$$

$$v_1, v_2, ...v_s \geq 0 \tag{4}$$

In which j is the BU index, where j=1,2,...n.

This model can be converted to a linear programming problem as following.

$$max w_o = u_1 y_{1o} + ... + u_s y_{so} \tag{5}$$

Subjected to:

$$v_1 x_{1o} + ... + v_m x_{mo} = 1 \tag{6}$$

$$u_1 y_{1j} + ... + u_{sj} \leq v_1 x_{1j} + ... + v_m x_{mj} (j = 1, ..., n) \quad (7)$$

$$u_1, u_2, ..., u_s \geq 0 \quad (8)$$

$$v_1, v_2, ..., v_m \geq 0 \quad (9)$$

If $w_o = 1$ that means that UP_o is efficient compared to others BU. Otherwise if $w_o \leq 1$ then BU_o is inefficient. This model is known as input-oriented CCR.

2.3. K-means Clustering

Clustering algorithms have extensive application in the data mining process [11]. They are used to classify individuals who have similar characteristics as belonging to the same group. Some grouping algorithms allow an individual to belong to more than one group. Others associate individuals with groups in a probabilistic rather than categorical way. There are still algorithms that aggregate individuals hierarchically. In this sense, the K-Means algorithm is a computational intelligence technique that groups and classifies data according to some distance measurement, being the method proposed by Macqueen in 1967 [12]. Generally speaking, the algorithm begins from the idea that each point in a group should be near the center of that group.

2.3.1. Method Description

First, we start choosing the k number of clusters that we want to find in a given dataset. Next k barycenter points are randomly selected. Then, for each barycenter, the nearest data points are assigned, as belonging to the group that orbits it. Therefore, a newly displaced barycenter is performed, now based on the distances between the points of the group. In sequence, the data points are once again reassigned as belonging to the nearest barycenter, finally, the method is repeated until convergence, as described in Fig 1.

2.4. Principal Component Analysis

According to Wold [13] principal component analysis (PCA) is the basis for multivariate data analysis. In short, the PCA provides an approximation of a data table, a data matrix, X, in terms of the product between two T and P' matrices. Being that, T and P', capture the essential patterns of X.

Wold [13] also defines the objectives of the PCA, among them stand out: simplification, data reduction, modeling, detection of extreme points, selection of variables, classification, prediction and separation of data.

It is noteworthy that the result of the analysis depends on the variables being properly normalized so that a unitary variance can be obtained.

3. Solution Proposal

3.1. Data Collection

Data from 19 schools distributed around the city of Rio de Janeiro were collected. The information retrieved refers to the period preceding the beginning of the school year, a phase in which the acquisition and renewal of students take place, also known as the enrollment cycle, due to its seasonal nature. The schools in question have as a new student acquisition business strategy to offer scholarships that are awarded based on the performance of the candidate in a multi-choice examination of Portuguese language and math. This performance is measured by the number of correct answers in each question that was collected and quantified in this research ($t_{average}$). Furthermore, another feature collected concerns the salary estimate (s_{idx}) for the respective candidate's financial officer based on registration form data together with the crossing of public and private databases. In another aspect, we also collected the estimated student amounts (c_{amount}) and capacities of the classes ($c_{capacieties}$) offered by each school, concomitantly with the school segment, which was subdivided into 5 categories: early childhood (ei), elementary 1 ($ef1$), elementary 2($ef2$), high school (em), and post-high school (cl). To model the efficiency of each school in enrolling students, the rate of candidates attending the test(att_{idx}) and the rate of students enrolled (enr_{idx}) until the period of data collection (17/11/2020) were calculated. distribution of each of the described values are shown in Table 1.

It is noted that the average attendance(att_{idx}) is close to 50% which is perceived as low and can be understood as an observable effect of the COVID-19 pandemic that had a strong impact on schools throughout the year of the

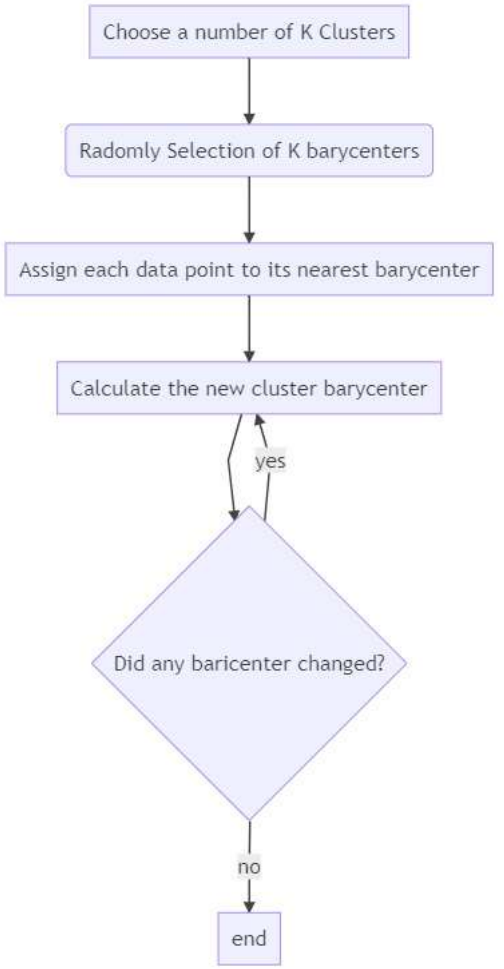

Fig. 1. K-means Diagram

present study. Another relevant aspect is the high standard deviation between schools, showing high variability between the ability to convert candidates to students. The number of classes also presents a high variance, to the point that there are schools with only two classes in contrast to the general average of 15 per school.

3.2. Clustering Schools

With the data summation by schools, normalization was performed between the extreme values of each vector in the school matrix, according to the formula below:

$$v = \frac{a_{ij} - \min a_{ij}}{\max a_{ij} - \min a_{ij}} \tag{10}$$

Using the principal component analysis technique, the dimensionality of the problem was reduced to two main components (Figure 2). At first, it is possible to identify some degree of separation between groups of schools. To perform the clustering and decrease group variability, the k-means algorithm was utilized. Also, to find a sub-optimal number of clusters the ”Elbow Method” was performed, in which, the sum of the squares of the distances of the groups up to their respective barycenter is calculated (Also known as the inertia of the clusters), and plotted for different numbers of clusters (Figure 3). Next, one searches for the point before the inertia starts to decrease

Table 1. Distribution of Interest Variables

att_{idx}	enr_{idx}	$t_{average}$	s_{idx}	$c_{capacity}$	cl	ei	ef1	ef2	em	c_{amount}	
count	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00
mean	0.52	0.47	8.40	2.04	1309.89	0.30	0.00	0.11	0.22	0.37	15.58
std	0.05	0.09	1.00	0.53	451.81	0.22	0.02	0.17	0.13	0.17	5.09
min	0.45	0.35	6.71	1.42	270.00	0.00	0.00	0.00	0.00	0.00	2.00
25%	0.49	0.39	7.85	1.68	1085.00	0.19	0.00	0.00	0.09	0.32	13.50
50%	0.51	0.46	8.25	1.90	1403.00	0.28	0.00	0.00	0.26	0.42	18.00
75%	0.55	0.52	8.73	2.32	1557.00	0.36	0.00	0.24	0.28	0.44	18.50
max	0.66	0.69	11.01	3.54	2376.00	1.00	0.08	0.56	0.44	0.62	23.00

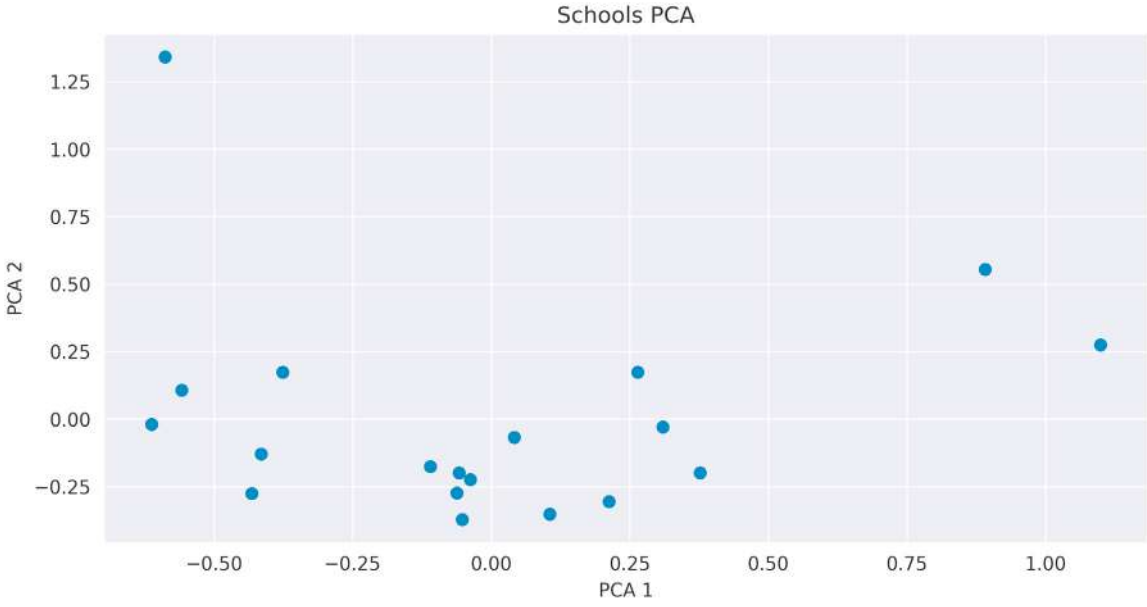

Fig. 2. Schools PCA

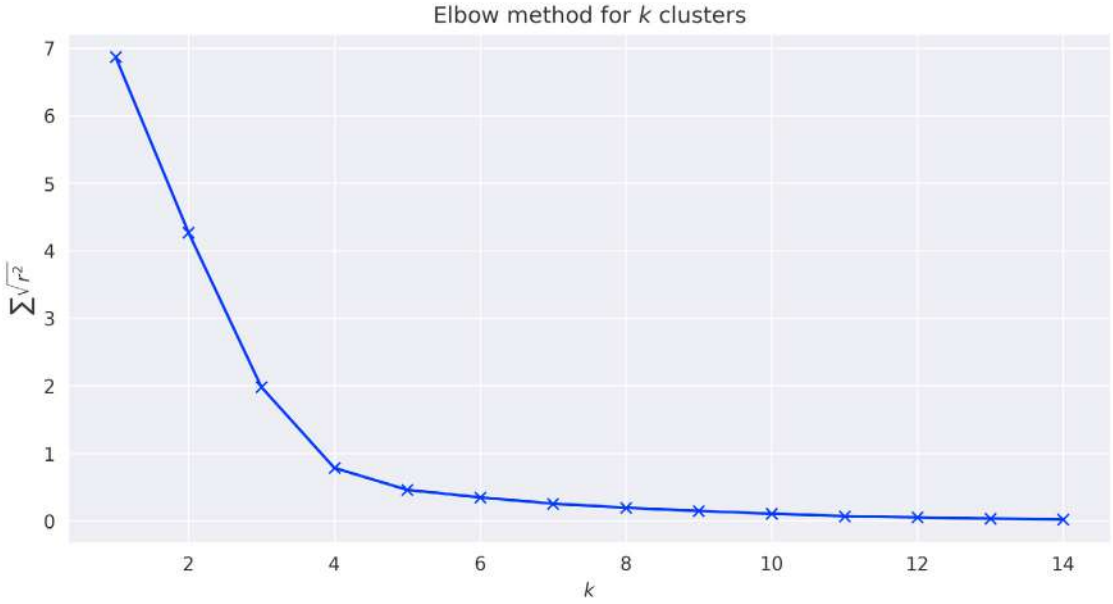

Fig. 3. Elbow Method

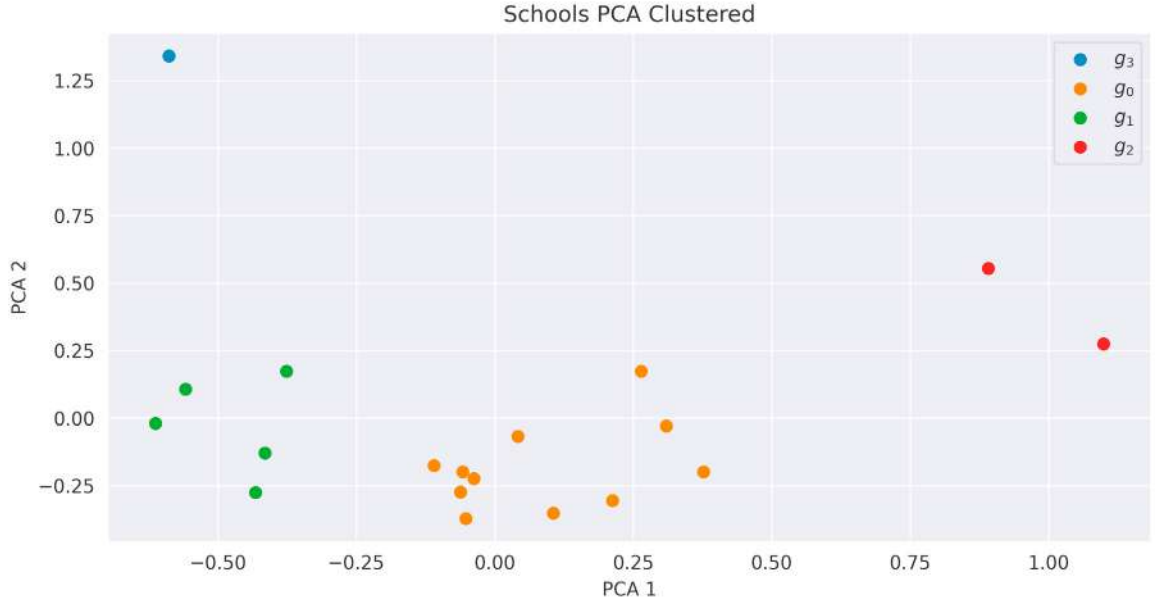

Fig. 4. Schools Clustered

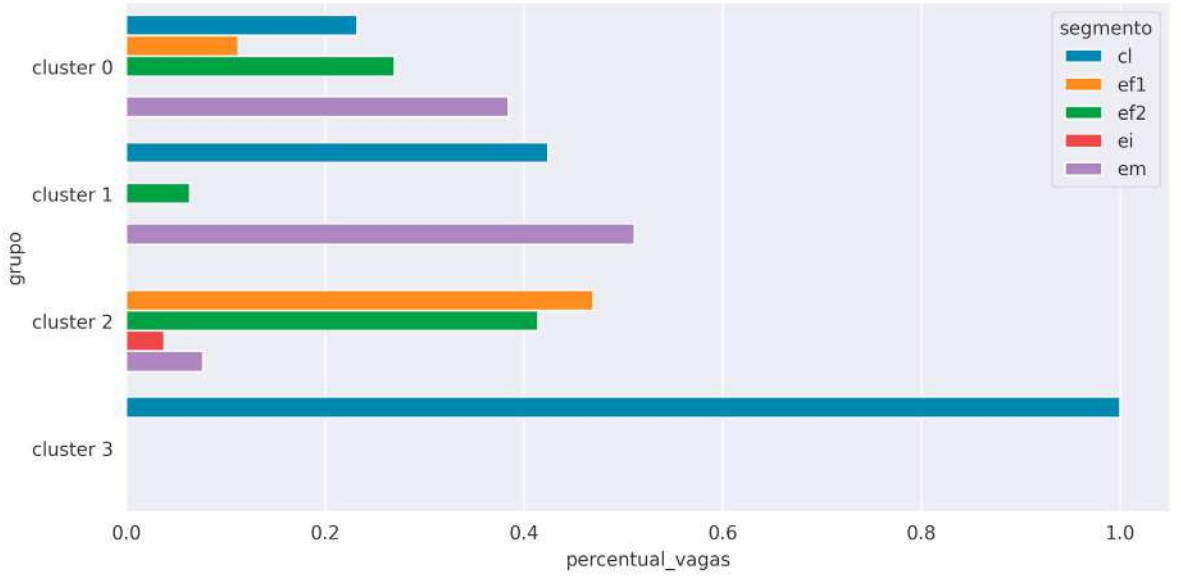

Fig. 5. Rate of class capacity per Segment per Cluster

slowly. In this dataset, this point was identified as 4 Clusters. Finally, the groups represented in the figure were obtained as a result (Fig 3).

The g_0 stands out for having the lowest average of correct answers, along with parents with the second-worst salary indicator. The classes of this group also present themselves as the largest. In terms of the distribution of class vacancies, it presents itself with a diverse profile of class offers, however, with no classes in the children's segment (Figure 5).

g_1 has the best attendance rate among the groups and a median utilization rate. The average of grades is the lowest, compared to the others. However, the salary indicator is the second highest. Although it is not the group that has, on average, the highest number of classes, it is the group that has the highest average student capacity. This coincides with the fact that it is a group that presents classes, almost in their entirety, with high school and post-high school classes, where the higher degree of independence between students and teachers, allows the presence of larger classes.

In further analysis, g_2 is the group with the worst utilization rate, although the performance of students in the cluster is close to the first two groups, and the salary indicator as well. In this case, it is worth mentioning that the class's capacity is small, both because there are few classes, as well as because it is a group that contains only elementary school classes. Classes that due to the pandemic and insecurity, and often the malaise generated by remote education, made those responsible more afraid about the prospect of studies in the following year.

Finally, one has g_3 which is, essentially, formed by a school that has a very different profile from the other ones. Where the estimated average salary, reaches 20% more than the previous groups, and where the focus is specific post-medium classes. Classes oriented to renowned public institutions entrance exams.

3.3. Clusters Data Envelopment Analysis

Since it is intended to compare the efficiencies of schools, in groups not so heterogeneous within the network, we chose not to perform the analysis for g_3. Since it is a group with only one school, it is an extreme value(outlier) in terms of the percentage of enrolled, concerning several others in the network.

To perform data envelopment analysis, we used a program written in the python programming language, that implements the algorithm discussed in section 2.2 using the PulP library [14] to solve the linear programming problem.

Table 2. Amount of Schools per Cluster

cluster	count
0	11
1	5
2	2
3	1

Table 3. g_0 Eficiency

city	neighberhood	school	efficiency
4	5	10	0.93
2	2	9	0.97
4	14	18	1.00
5	15	14	1.00
4	12	16	1.00
4	7	5	1.00
4	4	1	1.00
4	10	13	1.00
6	16	15	1.00
4	9	12	1.00
1	1	4	1.00

3.4. Results

For g_0 schools 10 and 9 were below the efficiency frontier as observed in Table 3.

For g_1 the school that is performing below the efficiency frontier is school 17 as shown in Table 5.

Through comparison, the data envelopment analysis was performed for the set of non-grouped data. It is notable that in Table 6, there are more inefficient units than presented in the first two tables (five concerning the two in g_0 and one for g_1). This happens because the groups have a degree of heterogeneity that triggers incoherent efficiency values, given the reality of that specific group. It was also observed that the same schools are operating below capacity. Within the groups, there are also convergent results with the calculated for the dataset without grouping.

The strategy proposed for the problem presented above mainly focus on business revenue aspects of attracting new students to the chosen group of private schools, but other context such as retention and student predicted performance can be used to obtain an overall understanding of the BU performance. In that sense, a multidisciplinary approach can benefit both further research on the subject, such as studied business units.

4. Conclusion

Data envelopment analysis is a robust and practical method for finding efficiency boundaries in a dataset. However, the premise of homogeneity between business units is not simple to be met. In this sense, the use of PCA and K-means techniques were presented as means of extracting relevant groups in an unsupervised way within

Table 4. g_1 Efficiency

city	neighborhood	school	efficiency
4	13	17	0.78
4	8	6	1.00
4	6	2	1.00
3	3	11	1.00
4	8	8	1.00

Table 5. Efficiency without grouping

city	neighborhood	school	efficiency
4	13	17	0.77
4	5	10	0.93
2	2	9	0.97
3	3	11	0.97
4	6	2	0.98
4	12	16	1.00
4	8	6	1.00
4	8	7	1.00
5	15	14	1.00
0	0	0	1.00
4	7	5	1.00
4	10	13	1.00
4	14	18	1.00
4	4	1	1.00
4	11	3	1.00
6	16	15	1.00
4	8	8	1.00
4	9	12	1.00
1	1	4	1.00

the proposed dataset. This solution converges with what has already been observed in the literature regarding the synergy between DEA and unsupervised clustering algorithms, such as K-means [15].

Thus, the use of the DEA allows finding the units that are in the worst performance, within the context of the group, in contrast to those that are better performing. Such analysis is useful for mapping good practices, creating key production indicators, and continuous process improvement.

Further research relies in utilizing other unsupervised clustering methods such as Hierarchical, Fuzzy clustering and Density Based clustering to contrast results and generalizing the proposed method for conducting DEA analysis.

References

[1] M. Farrell, The measurement of productive efficiency, 1957.

[2] A. S. Oliveira, C. F. S. Gomes, C. T. Clarkson, A. M. Sanseverino, M. R. S. Barcelos, I. P. A. Costa, M. Santos, Multiple criteria decision making and prospective scenarios model for selection of companies to be incubated, Algorithms 14 (4). doi:10.3390/a14040111.
URL https://www.mdpi.com/1999-4893/14/4/111

[3] M. Â. L. Moreira, C. F. S. Gomes, M. dos Santos, M. do Carmo Silva, J. V. G. A. Araujo, Promethee-sapevo-m1 a hybrid modeling proposal: Multicriteria evaluation of drones for use in naval warfare, in: A. M. T. Thomé, R. G. Barbastefano, L. F. Scavarda, J. C. G. dos Reis, M. P. C. Amorim (Eds.), Industrial Engineering and Operations Management, Springer International Publishing, Cham, 2020, pp. 381–393.

[4] M. L. Moreira, I. P. de Araújo Costa, M. T. Pereira, M. dos Santos, C. F. S. Gomes, F. M. Muradas, Promethee-sapevo-m1 a hybrid approach based on ordinal and cardinal inputs: Multi-criteria evaluation of helicopters to support brazilian navy operations, Algorithms 14 (5). doi:10.3390/a14050140.
URL https://www.mdpi.com/1999-4893/14/5/140

[5] C. F. S. A. Gomes, M. d. Santos, L. F. H. A. d. S. d. B. Teixeira, A. M. Sanseverino, M. R. d. S. Barcelos, SAPEVO-M: A GROUP MULTICRITERIA ORDINAL RANKING METHOD, Pesquisa Operacional 40.
URL http://www.scielo.br/scielo.php?script=sci$_a$rttextpid = S0101 − 74382020000100212nrm = iso

[6] W. Cooper, L. Seiford, K. Tone, Introduction to data envelopment analysis and its uses: With dea-solver software and references, 2005.

[7] F. M. Tenório, M. dos Santos, C. F. S. Gomes, J. de Carvalho Araujo, Navy warship selection and multicriteria analysis: The THOR method supporting decision making, in: Industrial Engineering and Operations Management, Springer International Publishing, 2020, pp. 27–39. doi:10.1007/978-3-030-56920-4$_3$.

[8] I. P. de Araújo Costa, A. Sanseverino, M. Barcelos, M. Belderrain, C. Gomes, M. Santos, Choosing flying hospitals in the fight against the covid-19 pandemic: structuring and modeling a complex problem using the vft and electre-mor methods, IEEE Latin America Transactions 100 (1e).

[9] A. Charnes, W. Cooper, B. Golany, L. Seiford, J. Stutz, Pareto-optimality, efficiency analysis and empirical production functions., 1983.

[10] W. Cooper, L. Seiford, K. Tone, Data envelopment analysis: A comprehensive text with models, applications, references and dea-solver software, 1999.

[11] D. Arthur, S. Vassilvitskii, K-means++: The advantages of careful seeding, SODA '07, Society for Industrial and Applied Mathematics, USA, 2007, p. 1027–1035.
[12] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA, 1967, pp. 281–297.
[13] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and intelligent laboratory systems 2 (1-3) (1987) 37–52.
[14] S. Mitchell, M. O'Sullivan, I. Dunning, Pulp : A linear programming toolkit for python, 2011.
[15] F. F. Razi, A hybrid dea-based k-means and invasive weed optimization for facility location problem, Journal of Industrial Engineering International 15 (3) (2019) 499–511.