Available online at http://ijdea.srbiau.ac.ir

Int. J. Data Envelopment Analysis (ISSN 2345-458X)

Vol. 10, No. 4, Year 2022, Article ID IJDEA-00422, Pages 23-34
Research Article

International Journal of Data Envelopment Analysis

Science and Research Branch (IAU)

# Machine learning clustering algorithms based on Data Envelopment Analysis in the presence of uncertainty

**R. Ghasempour Feremi, M. Rostamy-Malkhalifeh**[*]

Department of Mathematics, Science and Research Branch, Islamic Azad University, Tehran, Iran

Received 6 August 2022, Accepted 17 October 2022

**Abstract**

This study combines Data Envelopment Analysis (DEA) with machine learning clustering method in datamining for finding the most efficient Decision-Making Unit (DMU) and the best clustering algorithm, respectively. The problem of assessment of units by using DEA may not be straightforward due to the data uncertainty. Several scholars have been attracted to develop methods which incorporate uncertainty into input/output values in the DEA literature. On the other hand, in many real-world applications, the data is reported in the form of intervals. This means that each input/output value is selected from a symmetric box. In the DEA literature, this type of uncertainty has been addressed as Interval DEA approaches. The main goal of this study is to evaluate the efficiency of banks in the case of data uncertainty with cross-efficiency method in the DEA literature. For this purpose, we consider the BCC-CCR and CCR-BCC models in the presence of uncertain data to find the superior model. After applying the optimization models, in machine learning step, clustering method is applied. Clustering is a procedure for grouping similar items together which this group is called the cluster. Also, the different clustering algorithms can be used according to the behavior of data. In this study, we apply the farthest first and expectation maximization algorithms and show that, in the case of data uncertainty, the BCC-CCR and farthest first algorithms are as a superior optimization model and machine learning algorithm, respectively.

**Keywords:** Data Envelopment Analysis; Machine learning; Data Mining; Clustering; Cross-efficiency; Banking system.

[*] Corresponding author: Email: Mohsen_Rostamy@yahoo.com

# 1. Introduction

The traditional DEA model, CCR model, introduced by Charnes is a linear programming model which deals with the precise data where inputs and outputs values are deterministic and exactly known. However, in many real-world applications, there exists imprecise data due to the incomplete or non-attainable information, errors in measurements, unquantifiable variables, or any other source of reason. Some challenges in applying the DEA technique may arise due to the existence of imprecise data, for example, most of the models formulated in these situations are non-linear programming models [1].

The problem of the evaluation of units with imprecise data has attracted attentions of several scholars. For example, Cooper developed Imprecise Data Envelopment Analysis (IDEA) method. Their method can be applied in the situation where there exist both imprecisely and exactly-known data in which the IDEA models are transformed into linear programming problems [2]. Kim proposed a procedure to incorporate partial data into DEA. Their original model was a complicated non-linear model that was transformed into a linear programming problem by applying a linear scale transformation and the variable change technique [3].

In summary, there are three different types of approaches to model the imprecise data in DEA, i.e., fuzzy approaches, stochastic methods, and robust optimization-based techniques. For more information about fuzzy DEA, readers can refer to [4, 5, 6, 7]. Also input/output variables can be considered as random variables, which results in stochastic DEA models [8, 9]. Olesen and Petersen provided a review on stochastic DEA methods and Peykani presented a review on robust DEA methods. Robust Optimization (RO) is a technique to model optimization problems with uncertain data which aims to determine an optimal solution which is the best for all or the most possible realizations of the uncertain parameters [8, 10]. Ben-Tal and Nemirovski and Bertsimas and Sim investigated uncertainty in data and proposed different RO approaches to obtain the optimal solution [11-14].

Wang applied Ben-Tal and Nemirovski's approach in DEA to develop two robust formulations for the multiplier form of the CCR model in the presence of uncertain data. They considered the perturbations on inputs/outputs for the different uncertainty levels and computed the efficiency score of units and provided a ranking for them. Sadjadi and Omrani proposed the robust formulation of the multiplier form of CCR model based on the RO technique presented by Ben-Tal and Nemirovski. Their proposed model is a non-linear programming problem which shows the drawback of their approach of applying an inappropriate RO technique. They also presented the robust formulation of the multiplier form of the CCR model based on the RO technique proposed by Bertsimas. Unlike, the first model, the second is a linear programming model [15].

Sadjadi and Omrani proposed a bootstrapped robust model for the multiplier form of the CCR model, based on the approach of Bertsimas, to solve the perturbation and sampling error problems [16]. Sadjadi proposed an interactive robust model based on Bertsimas et al.'s approach (2004) to find the targets of units according to the DM's preferences [17]. Omrani (2013) proposed a RO technique, based on the robust approach of Bertsimas et al. (2004), to find the common set of weights in DEA by using the goal programming technique. Their method can be applied to evaluate the absolute efficiency score of units for different values of robustness levels in order to rank them [18]. Ehrgott (2018) used the framework of RO to propose a DEA model in case of data uncertainty. They provided a first-order algorithm to solve their model and showed that the optimal solution of it determined the maximum possible efficiency score of a unit [19].

On the other hand, although the banking industry in developing countries has had unprecedented growth, however, research on performance and the efficiency of this industry is almost challenging. Tsolas and Charles (2015) pointed out that the banks are crucial foundations in a country's budget and economy [20]. Given that the importance of economic institutes, many scholars have been attracted to evaluate the performance of banks in the various countries, for example see [21-29].

The various DEA models are commonly used in the different studies to rank and evaluate the efficiency of banks. Hence, we can obtain a comprehensive comparison of several efficiencies' insight into the bank's efficiency in the case of data uncertainty. This comparison is very useful for the bank practitioners who desire to assess efficiency at a proper step of its progression. This study applies four models in cross-efficiency in the presence of uncertain data, which eventually results in comparing several efficient and inefficient DMUs with interval data. Finally, we find the superior model which provides the valuable information for the bank managers to select the best model.

Given that the importance of the bank's efficiency assessment, this paper aims to measure the efficiency of banks in the case of data uncertainty by combining DEA with machine learning. For this purpose, we present the traditional DEA models, CCR-BCC and BCC-CCR, in the presence of interval data and develop the cross-efficiency model in the case of data uncertainty. In the next step, i.e., in machine learning step, we apply the different clustering algorithms, such as the farthest first and expectation maximization algorithms, with respect to the behaviour of data. Finally, we find the superior optimization model and the machine learning algorithm.

The rest of this paper unfolds as follows. Section 2 provides the preliminaries and basic definitions. Section 3 proposes the DEA models in the case of data uncertainty. The results are illustrated by a numerical example in Section 4. Finally, Section 5 concludes the paper and provides direct for future research.

## 2. Preliminaries and basic definitions

Consider a system of $n$ DMUs, denoted by $DMU_j, j = 1, \dots, n,$ where each unit consumes $m$ different inputs to generate $s$ different outputs. The $i^{th}$ input and $r^{th}$ output for $DMU_j$ are denoted by $x_{ij}$ and $y_{rj}$, respectively, for $i = 1, \dots, m$ and $r = 1, \dots, s$. Assume that $DMU_o$ is the unit under evaluation.

### 2.1 CCR-BCC model

Now, consider the manufacturing technology where if it produces $X_o$ and $Y_o$ then $\lambda X_o$ can produce $\lambda Y_o$ only when $\lambda \leq 1$. So, according to the include observations, convexity and feasibility axioms, the production possibility set (PPS) can be written as follows:

$$T_{CCR-BCC} = T_{NI} = \left\{(x, y) \mid \sum_{j=1}^{n} \lambda_j X_j \leq X, \sum_{j=1}^{n} \lambda_j Y_j \geq Y, \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \lambda_j \leq 1, \lambda \geq 0\right\} \quad (1)$$

The main goal of the input-oriented models in the DEA literature is to find a virtual unit in which the input $\theta x_o$ is not more than $x_o$, and the minimum production should be $y_o$, In the other word, we have the following model:

$$\begin{aligned} &\min \theta \\ &s.t. \\ &(\theta X_o, Y_o) \in T_{NI} \end{aligned} \quad (2)$$

Based on the structure of $T_{NI}$, model CCR-BCC is formulated as follows:

$$\begin{aligned} &\min \theta && (3)\\ &s.t.\\ &\quad \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_p && i=1,\dots,m,\\ &\quad \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rp} && r=1,\dots,s,\\ &\quad \sum_{j=1}^{n} \lambda_j \leq 1,\\ &\quad \lambda_j \geq 0, && j=1,\dots,n. \end{aligned}$$

**2.2 BCC-CCR model**

In this section, consider the manufacturing technology where if it produces $X_o$ and $Y_o$ then $\lambda X_o$ can produce $\lambda Y_o$ only when $\lambda \geq 1$. So, according to the include observations, convexity and feasibility axioms, the production possibility set (PPS) can be written as follows:

$$T_{BCC-CCR} = T_{ND} = \left\{(x,y) \mid \sum_{j=1}^{n} \lambda_j X_j \leq X, \sum_{j=1}^{n} \lambda_j Y_j \geq Y, \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \lambda_j \geq 1, \lambda \geq 0\right\} \quad (4)$$

As we said before, the goal of the input-oriented models is to find a virtual unit in which the input $\theta x_o$ is not more than $x_o$, and at least produce $y_o$, In the other word, we have the following model:

$$\begin{aligned} &&& (5)\\ &\min \theta\\ &s.t.\\ &(\theta X_o, Y_o) \in T_{ND} \end{aligned}$$

Based on the structure of $T_{ND}$, model BCC-CCR is formulated as follows:

$$\begin{aligned} &\min \theta && (6)\\ &s.t.\\ &\quad \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_p && i=1,\dots,m,\\ &\quad \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rp} && r=1,\dots,s,\\ &\quad \sum_{j=1}^{n} \lambda_j \geq 1,\\ &\quad \lambda_j \geq 0, && j=1,\dots,n. \end{aligned}$$

**3. The efficiency evaluation in the case of data uncertainty**

Consider a system of $n$ DMUs, denoted by $DMU_j, j=1,\dots,n,$ where each unit consumes $m$ different inputs to generate $s$ different outputs. The $i^{th}$ input and $r^{th}$ output for $\text{DMU}_\text{j}$ are denoted by $x_{ij}$ and $y_{rj}$, respectively, for $i=1,\dots,m$ and $r=1,\dots,s$. Also, suppose that input and output values are not deterministic for all units and $x_{ij} \in [x_{ij}^L, x_{ij}^U]$ and $y_{rj} \in [y_{rj}^L, y_{rj}^U]$, where the lower and upper bounds are positive and finite values. Assume that $DMU_o$ is the unit under evaluation.

### 3.1 CCR-BCC in the case of data uncertainty

The input-oriented of $CCR_{Io} - BCC_{Io}$ model in the presence of uncertain data is as follows:

$$\begin{aligned} &\min \theta && (7)\\ &s.t.\\ &\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_p, && i=1,\dots,m, x_{ij}^L \leq x_{ij} \leq x_{ij}^U,\\ &\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rp}, && r=1,\dots,s, y_{rj}^L \leq y_{rj} \leq y_{rj}^U,\\ &\sum_{j=1}^{n} \lambda_j \leq 1,\\ &\lambda_j \geq 0 && j=1,\dots,n \end{aligned}$$

The robust counterpart of model (7) in the sense of Ben-Tal and Nemirovski (2000) is as follows:

$$\begin{aligned} &\min \theta && (8)\\ &s.t.\\ &(\theta - \lambda_o)x_{io}^L - \sum_{\substack{j=1\\ j\neq o}}^{n} \lambda_j x_{ij}^U \geq 0, && i=1,\dots,m,\\ &\sum_{\substack{j=1\\ j\neq o}}^{n} \lambda_j y_{rj}^L + (\lambda_o - 1) y_{ro}^U \geq 0, && r=1,\dots,s,\\ &\sum_{j=1}^{n} \lambda_j \leq 1,\\ &\quad \lambda_j \geq 0 && j=1,\dots,n \end{aligned}$$

The dual of the robust counterpart of $CCR_{Io} - BCC_{Io}$ is as follows:

$$\begin{aligned} &\max \sum_{r=1}^{s} u_r y_{ro}^U + u_0 && (9)\\ &\text{s.t.}\\ &\sum_{r=1}^{s} u_r y_{rj}^L - \sum_{i=1}^{m} v_i x_{ij}^U + u_0 \leq 0, && j \neq o,\\ &\sum_{r=1}^{s} u_r y_{ro}^U - \sum_{i=1}^{m} v_i x_{io}^L + u_0 \leq 0,\\ &\sum_{i=1}^{m} v_i x_{io}^L = 1,\\ &\quad u_r \geq 0, && r=1,\dots,s,\\ &\quad v_i \geq 0, && i=1,\dots,m.\\ &\quad u_0 \leq 0. \end{aligned}$$

**3.1 BCC-CCR in the case of data uncertainty**

The input-oriented of $BCC_{Io} - CCR_{Io}$ model in the presence of uncertain data is as follows:

$$\begin{aligned} &\min \theta && (10)\\ &s.t. \\ &\textstyle\sum_{j=1}^{n} \lambda_j x_{ij} \le \theta_p, \quad i = 1, \dots, m, x_{ij}^L \le x_{ij} \le x_{ij}^U, \\ &\textstyle\sum_{j=1}^{n} \lambda_j y_{rj} \ge y_{rp}, \quad r = 1, \dots, s, y_{rj}^L \le y_{rj} \le y_{rj}^U, \\ &\textstyle\sum_{j=1}^{n} \lambda_j \ge 1, \\ &\quad \lambda_j \ge 0 \qquad j = 1, \dots, n \end{aligned}$$

The robust counterpart of model (10) in the sense of Ben-Tal and Nemirovski (2000) is as follows:

$$\begin{aligned} &\min \theta && (11)\\ &s.t. \\ &(\theta - \lambda_o) x_{io}^L - \textstyle\sum_{j=1, j\neq o}^{n} \lambda_j x_{ij}^U \ge 0, \quad i = 1, \dots, m, \\ &\textstyle\sum_{j=1, j\neq o}^{n} \lambda_j y_{rj}^L + (\lambda_o - 1) y_{ro}^U \ge 0, \quad r = 1, \dots, s, \\ &\textstyle\sum_{j=1}^{n} \lambda_j \ge 1, \\ &\quad \lambda_j \ge 0 \qquad j = 1, \dots, n \end{aligned}$$

The dual of the robust counterpart of $CCR_{Io} - BCC_{Io}$ is as follows:

$$\begin{aligned} &\max \textstyle\sum_{r=1}^{s} u_r y_{ro}^U + u_0 && (12)\\ &\text{s.t.} \\ &\textstyle\sum_{r=1}^{s} u_r y_{rj}^L - \sum_{i=1}^{m} v_i x_{ij}^U + u_0 \le 0, \quad j \neq o, \\ &\textstyle\sum_{r=1}^{s} u_r y_{ro}^U - \sum_{i=1}^{m} v_i x_{io}^L + u_0 \le 0, \\ &\textstyle\sum_{i=1}^{m} v_i x_{io}^L = 1, \\ &\quad u_r \ge 0, \qquad r = 1, \dots, s, \\ &\quad v_i \ge 0, \qquad i = 1, \dots, m. \\ &\quad u_0 \ge 0. \end{aligned}$$

In the next section, we use a dataset of Iranian banks to combine the DEA models, presented in the previous section, with the clustering algorithms.

**4. Evaluation in clustering**

This section considers a dataset which includes 12 banks in Iran as the DMUs with three inputs, number of staffs $(x_1)$, computer terminals $(x_2)$ and space $(x_3)$ to produce three outputs, deposits $(y_1)$, loans $(y_2)$ and charge $(y_3)$. The data is summarized in Table 1.

we apply the cross-efficiency method by using models (9) and (12) and the cross-efficiency matrices are reported in Table 2 and Table 3, respectively.

According to Table 2, the following results are obtained:

- $DMU_{12}$ has the first and the largest mean of the efficiency in model $CCR_{Io} - BCC_{Io}$.
- $DMU_7$ has the second efficiency score in model $CCR_{Io} - BCC_{Io}$.
- $DMU_5$ has the third efficiency score in model $CCR_{Io} - BCC_{Io}$.

Table 1. The data of Iranian banks.

| DMU | $x_1^L$ | $x_1^U$ | $x_2^L$ | $x_2^U$ | $x_3^L$ | $x_3^U$ | $y_1^L$ | $y_1^U$ | $y_2^L$ | $y_2^U$ | $y_3^L$ | $y_3^U$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.7602 | 1.1404 | 0.56 | 0.84 | 0.1240 | 0.1860 | 0.1520 | 0.2280 | 0.4171 | 0.6257 | 0.2341 | 0.3511 |
| 2 | 0.6370 | 0.9554 | 0.48 | 0.72 | 0.8000 | 1.2000 | 0.1813 | 0.2719 | 0.5019 | 0.7529 | 0.3699 | 0.5549 |
| 3 | 0.6386 | 0.9578 | 0.60 | 0.90 | 0.4100 | 0.6150 | 0.1826 | 0.2740 | 0.7762 | 1.1644 | 0.2085 | 0.3127 |
| 4 | 0.6921 | 1.0381 | 0.44 | 0.66 | 0.1680 | 0.2520 | 0.1542 | 0.2312 | 0.5059 | 0.7589 | 0.8000 | 1.2000 |
| 5 | 0.6521 | 0.9781 | 0.68 | 1.02 | 0.2140 | 0.3210 | 0.1866 | 0.2800 | 0.5777 | 0.8665 | 0.1970 | 0.2956 |
| 6 | 0.6733 | 1.0099 | 0.52 | 0.78 | 0.4000 | 0.6000 | 0.1655 | 0.2483 | 0.4820 | 0.7230 | 0.4551 | 0.6827 |
| 7 | 0.5751 | 0.8627 | 0.48 | 0.72 | 0.2800 | 0.4200 | 0.1459 | 0.2189 | 0.7200 | 1.0800 | 0.5726 | 0.8590 |
| 8 | 0.6282 | 0.9424 | 0.60 | 0.90 | 0.0960 | 0.1440 | 0.1000 | 0.1500 | 0.1872 | 0.2808 | 0.2382 | 0.3572 |
| 9 | 0.3805 | 0.5707 | 0.48 | 0.72 | 0.1080 | 0.1620 | 0.0641 | 0.0961 | 0.2914 | 0.4372 | 0.1951 | 0.2927 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 0.5426 | 0.8138 | 0.44 | 0.66 | 0.4080 | 0.6120 | 0.0654 | 0.0982 | 0.1468 | 0.2202 | 0.0389 | 0.0583 |
| 11 | 0.5690 | 0.8534 | 0.80 | 1.20 | 0.2440 | 0.3660 | 0.1694 | 0.2540 | 0.2543 | 0.3815 | 0.3225 | 0.4837 |
| 12 | 0.6490 | 0.9736 | 0.52 | 0.78 | 0.2040 | 0.3060 | 0.0982 | 0.1472 | 0.7380 | 1.1070 | 0.5023 | 0.7535 |

Table 2. The cross-efficiency matrix by using model (9).

| Efficiency | 0.94 | 0.80 | 0.91 | 0.96 | 1 | 0.88 | 1 | 0.76 | 0.85 | 0.79 | 0.97 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 0.90 | 0.85 | 0.897 | 0.92 | 0.95 | 0.87 | 0.96 | 0.80 | 0.86 | 0.81 | 0.93 | 0.97 |
| Rank | 6 | 10 | 7 | 5 | 3 | 8 | 2 | 12 | 9 | 11 | 4 | 1 |
| DMU | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 1 | 0.74 | 0.86 | 0.91 | 0.97 | 0.88 | 0.91 | 0.79 | 0.86 | 0.71 | 0.98 | 0.99 |
| 2 | 0.86 | 1 | 0.88 | 0.91 | 0.92 | 0.90 | 0.98 | 0.75 | 0.83 | 0.71 | 0.91 | 0.99 |
| 3 | 0.91 | 0.85 | 1 | 0.90 | 0.90 | 0.83 | 0.98 | 0.84 | 0.88 | 0.79 | 0.90 | 0.90 |
| 4 | 0.88 | 0.85 | 0.88 | 1 | 0.96 | 0.85 | 0.98 | 0.78 | 0.89 | 0.83 | 0.90 | 0.98 |
| 5 | 0.87 | 0.81 | 0.87 | 0.91 | 1 | 0.87 | 0.92 | 0.78 | 0.88 | 0.81 | 0.90 | 0.97 |
| 6 | 0.89 | 0.82 | 0.91 | 0.92 | 0.97 | 1 | 0.95 | 0.76 | 0.81 | 0.78 | 0.93 | 0.97 |
| 7 | 0.91 | 0.81 | 0.90 | 0.89 | 0.97 | 0.88 | 1 | 0.80 | 0.86 | 0.76 | 0.99 | 0.95 |
| 8 | 0.91 | 0.89 | 0.91 | 0.91 | 0.93 | 0.88 | 0.95 | 1 | 0.85 | 0.82 | 0.96 | 0.91 |
| 9 | 0.90 | 0.87 | 0.88 | 0.95 | 0.91 | 0.91 | 0.96 | 0.81 | 1 | 0.81 | 0.91 | 0.98 |
| 10 | 0.91 | 0.88 | 0.87 | 0.92 | 0.95 | 0.81 | 0.97 | 0.78 | 0.85 | 1 | 0.91 | 0.97 |
| 11 | 0.89 | 0.89 | 0.89 | 0.90 | 0.96 | 0.76 | 0.97 | 0.82 | 0.84 | 0.79 | 1 | 0.98 |
| 12 | 0.91 | 0.81 | 0.90 | 0.90 | 0.98 | 0.92 | 0.99 | 0.80 | 0.89 | 0.79 | 0.92 | 1 |

Table 3. The cross-efficiency matrix by using model (12).

| Efficiency | 0.95 | 0.81 | 0.92 | 0.97 | 1 | 0.89 | 1 | 0.77 | 0.86 | 0.80 | 0.98 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 0.91 | 0.86 | 0.90 | 0.93 | 0.96 | 0.88 | 0.97 | 0.81 | 0.87 | 0.82 | 0.94 | 0.98 |
| Rank | 6 | 10 | 7 | 5 | 3 | 8 | 2 | 12 | 9 | 11 | 4 | 1 |
| DMU | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 1 | 0.75 | 0.87 | 0.92 | 0.98 | 0.89 | 0.92 | 0.80 | 0.87 | 0.72 | 0.99 | 1 |
| 2 | 0.87 | 1 | 0.89 | 0.92 | 0.93 | 0.91 | 0.99 | 0.76 | 0.84 | 0.72 | 0.92 | 1 |
| 3 | 0.92 | 0.86 | 1 | 0.91 | 0.91 | 0.84 | 0.99 | 0.85 | 0.89 | 0.80 | 0.91 | 0.91 |
| 4 | 0.89 | 0.86 | 0.89 | 1 | 0.97 | 0.86 | 0.99 | 0.79 | 0.90 | 0.84 | 0.91 | 0.99 |
| 5 | 0.88 | 0.82 | 0.88 | 0.92 | 1 | 0.88 | 0.93 | 0.79 | 0.89 | 0.82 | 0.91 | 0.98 |
| 6 | 0.90 | 0.83 | 0.92 | 0.93 | 0.98 | 1 | 0.96 | 0.77 | 0.82 | 0.79 | 0.94 | 0.98 |
| 7 | 0.92 | 0.82 | 0.91 | 0.90 | 0.98 | 0.89 | 1 | 0.81 | 0.87 | 0.77 | 1 | 0.96 |
| 8 | 0.92 | 0.90 | 0.92 | 0.92 | 0.94 | 0.89 | 0.96 | 1 | 0.86 | 0.83 | 0.97 | 0.92 |
| 9 | 0.91 | 0.88 | 0.89 | 0.96 | 0.92 | 0.92 | 0.97 | 0.81 | 1 | 0.82 | 0.92 | 1 |
| 10 | 0.92 | 0.89 | 0.88 | 0.93 | 0.96 | 0.82 | 0.98 | 0.79 | 0.86 | 1 | 0.92 | 0.98 |
| 11 | 0.90 | 0.90 | 0.90 | 0.91 | 0.97 | 0.77 | 0.98 | 0.83 | 0.85 | 0.80 | 1 | 0.99 |
| 12 | 0.92 | 0.82 | 0.91 | 0.91 | 0.99 | 0.93 | 1 | 0.81 | 0.90 | 0.80 | 0.93 | 1 |

Table 4. The CCR-BCC cross-efficiency of the banks during 2015-2017.

| DMU | Eff. (2015) | Rank (2015) | Eff. (2016) | Rank (2016) | Eff. (2017) | Rank (2015) | Mean |
|---|---|---|---|---|---|---|---|
| 1 | 90.41 | 6 | 81.42 | 6 | 93.65 | 5 | 88.49 |
| 2 | 85.23 | 10 | 61.79 | 10 | 69.34 | 10 | 72.12 |
| 3 | 89.12 | 7 | 76.95 | 7 | 90.12 | 6 | 85.40 |
| 4 | 91.67 | 5 | 87.37 | 5 | 97.93 | 3 | 92.32 |
| 5 | 94.76 | 3 | 99.95 | 3 | 100 | 1 | 98.24 |
| 6 | 87.32 | 8 | 72.81 | 8 | 67.65 | 11 | 75.93 |
| 7 | 96.41 | 2 | 99.97 | 2 | 98.58 | 2 | 98.32 |
| 8 | 80.20 | 12 | 54.63 | 12 | 74.96 | 9 | 69.93 |
| 9 | 86.13 | 9 | 64.69 | 9 | 83.39 | 8 | 78.07 |
| 10 | 80.45 | 11 | 58.95 | 11 | 59.16 | 12 | 66.21 |
| 11 | 93.32 | 4 | 99.94 | 4 | 85.81 | 7 | 93.02 |
| 12 | 97.41 | 1 | 100 | 1 | 97.46 | 4 | 98.29 |
| Mean | 89.59 | | 79.87 | | 84.84 | | |

Table 5. The BCC-CCR cross-efficiency of the banks during 2015-2017.

| DMU | Eff. (2015) | Rank (2015) | Eff. (2016) | Rank (2016) | Eff. (2017) | Rank (2015) | Mean |
|---|---|---|---|---|---|---|---|
| 1 | 90.42 | 6 | 81.43 | 6 | 93.66 | 5 | 88.50 |
| 2 | 85.24 | 10 | 61.80 | 10 | 69.35 | 10 | 72.13 |
| 3 | 89.13 | 7 | 76.96 | 7 | 90.13 | 6 | 85.41 |
| 4 | 91.68 | 5 | 87.38 | 5 | 97.94 | 3 | 92.33 |
| 5 | 94.77 | 3 | 99.96 | 3 | 100 | 1 | 98.25 |
| 6 | 87.33 | 8 | 72.82 | 8 | 67.66 | 11 | 75.94 |
| 7 | 96.42 | 2 | 99.98 | 2 | 98.59 | 2 | 98.33 |
| 8 | 80.21 | 12 | 54.64 | 12 | 74.97 | 9 | 69.94 |
| 9 | 86.14 | 9 | 64.70 | 9 | 83.40 | 8 | 78.08 |
| 10 | 80.46 | 11 | 58.96 | 11 | 59.17 | 12 | 66.22 |
| 11 | 93.33 | 4 | 99.95 | 4 | 85.82 | 7 | 93.03 |
| 12 | 97.42 | 1 | 100 | 1 | 97.47 | 4 | 98.30 |
| Mean | 89.60 | | 79.88 | | 84.85 | | |

Similarly, according to Table 3, the following results are obtained:

- $DMU_{12}$ has the first and the largest mean of the efficiency in model $BCC_{Io} - CCR_{Io}$.
- $DMU_7$ has the second efficiency score in model $BCC_{Io} - CCR_{Io}$.
- $DMU_5$ has the third efficiency score in model $BCC_{Io} - CCR_{Io}$.

In the following, the mean of the CCR-BCC cross-efficiency score of the banks during 2015-2017 are summarized in Table 4.

In the following, the mean of the BCC-CCR cross-efficiency score of the banks during 2015-2017 are summarized in Table 5.

The following results are obtained by using Table 4 and Table 5.

- Models BCC-CCR and CCR-BCC determine the same ranks for all units.
- Model BCC-CCR shows the first efficiency mean during 2015-2017.
- Model CCR-BCC shows the second efficiency mean during 2015-2017

According to model CCR-BCC in Table 4, we have:

- The 7th bank shows the first or the largest efficiency mean equals to 98.32.
- The 12th and the 5th banks have the second and the third ranks among the units and their efficiency scores are 98.29 and 98.24, respectively.
- The efficiency score of the 10th Bank is 66.21 and is ranked 12th and shows the lowest efficiency score.
- The second and the 8th banks are ranked 11th and 12th and their efficiency scores are 72.12 and 69.93, respectively.

According to model BCC-CCR in Table 4, we have:

- The 7th bank shows the first or the largest efficiency mean equals to 98.33.
- The 12th and the 5th banks have the second and the third ranks among the units and their efficiency scores are 98.30 and 98.25, respectively.
- The efficiency score of the 10th Bank is 66.22 and is ranked 12th and shows the lowest efficiency score.
- The second and the 8th banks are ranked 11th and 12th and their efficiency scores are 72.13 and 69.94, respectively.

Regarding the above discussion, models BCC-CCR and CCR-BCC are located in the first and second places, respectively. Finally, the following relation are held for all units:

$BCC - CCR > CCR - BCC$

In the next step, we compare the clustering algorithms, the farthest first and the expectation maximization.

The farthest first algorithm is a modified of K-means that seats each cluster center in sequence at the point farthest from the exiting cluster center lying inside the data range. It is appropriate for large-scale data sets. Farthest first is a heuristic created process of clustering. It also selects centroid and allocates the items in clusters. This algorithm provides fast clustering in most of the cases since less relocation and modification is required. Expectation maximization algorithm is an iterative technique for the detection of maximum possibility in statistical models, and undetected hidden variables determine it. It provides a valuable result for the actual world data set.

Table 6. Accuracy comparison contained by clustering algorithms
(All numbers are in percent)

| Algorithms | CCR-BCC | BCC-CCR |
|---|---|---|
| Farthest First | 74.3425 | 81.8713 |
| Expectation Maximization | 71.5612 | 78.1346 |
| Average | 75.4519 | 80.0030 |

The EM iteration substitutes among acting an expectation (E) step, which produces a purpose intended for the expectation of the log-possibility assessed utilizing the existing evaluation for the factors, and a maximization (M) step, which calculates factors maximizing the expected log-possibility establish on the E step. These factor-estimates are then utilized to conclude the delivery of the hidden variables in the next E step.

According to Table 6, we have:

- The maximum of accuracy within two assessments is improved.
- The average accuracy within two algorithms, is augmented.
- The accuracy of two algorithms is increased.

Finally, after BCC-CCR and CCR-BCC models are applied in the first step, in the extraction of hidden rules of the second clustering step, BCC-CCR is the superior model.

## 5. Conclusion

This paper considered box-uncertainty in DEA models where each input/output variable varies in an interval. A robust optimization framework was proposed for performance measurement and ranking of DMUs with interval data. This study combined DEA with machine learning clustering method in datamining for finding the most efficient Decision-Making Unit (DMU) and the best clustering algorithm, respectively. For this purpose, we considered the BCC-CCR and CCR-BCC models in the presence of uncertain data to find the superior model. Then, we applied the different clustering method, such as the farthest first and expectation maximization algorithms and showed that, in the case of data uncertainty, the BCC-CCR and farthest first algorithms were as a superior optimization model and machine learning algorithm, respectively.

**References**


[1] Charnes, A., Cooper, W.W., Rodes, E., 1978. Measuring the efficiency of decision-making units, European Journal of Operational Research, 2, 444-429.

[2] Cooper, W.W., Park, K.S., Yu, G. (1999). IDEA and AR-IDEA: models for dealing with imprecise data in DEA, *Management Science*, 45 (4), 597-607.

[3] Kim, S.H., Park, C.K. and Park, K.S. (1999). An application of data envelopment analysis in telephone offices evaluation with partial data, *Computers & Operations Research*, 26 (1), 59-72.

[4] Peykani, P., Seyed Esmaeili, F. S., Rostamy-Malkhalifeh, M., & Hosseinzadeh Lotfi, F. (2018). Measuring productivity changes of hospitals in Tehran: the fuzzy Malmquist productivity index. *International Journal of Hospital Research*, *7*(3), 1-16.

[5] Rostamy-Malkhalifeh, M., & Mollaeian, E. (2012). Evaluating performance supply chain by a new non-radial network DEA model with fuzzy data. *Science*, *9*.

[6] Sanei, M., Rostami-Malkhalifeh, M., & Saleh, H. (2009). A new method for solving fuzzy DEA models. *International Journal of Industrial Mathematics*, *1*(4), 307-313.

[7] ROSTAMY, M. M., Sanei, M., & Saleh, H. (2009). A new method for solving fuzzy DEA models by trapezoidal approximation.

[8] Olesen, O. B., & Petersen, N. C. (2016). Stochastic data envelopment analysis—A review. *European Journal of Operational Research*, 251(1), 2-21.

[9] Peykani, P., Mohammadi, E., Sadjadi, S. J., & Rostamy-Malkhalifeh, M. (2018, May). A robust variant of radial measure for performance assessment of stock. In *3th International Conference on Intelligent Decision Science, Iran*.

[10] Peykani. P., E. Mohammadi, R. Farzipoor Saen, S. J. Sadjadi an M. Rostamy-Malkhalifeh (2020). Data envelopment analysis and robust optimization: A review, *Expert Systems*.

[11] Ben-Tal, A., Nemirovski, A. 1998. Robust convex optimization, *Mathematics of Operations Research*, 23 (4), 769-805.

[12] Ben-Tal, A., Nemirovski, A. 2000. Robust solutions of linear programming problems contaiminated wit uncertain data, *Mathematical Programming*, 88 (3), 411-421.

[13] Ben-Tal, A., Nemirovski, A. 1999. Robust solutions of uncertain linear programs, *Operations Research Letters*, 25 (1), 1-13.

[14] Bertsimas, D., Sim, M. 2004. The price of robustness, *Operations Research*, 52 (1), 35-53.

[15] Sadjadi, S.J., Omrani, H. (2008). Data envelopment analysis with uncertain data: An application for Iranian electricity distribution companies, *Energy Policy*, 36 (11), 4247-4254.

[16] Sadjadi, S.J., Omrani, H. (2010). A bootstapped robust data envelopment analysis model for efficiency estimating of telecommunication companies in Iran, *Telecommunication Policy*, 34 (4), 221-232.

[17] Sadjadi, S.J., Omrani, H., Makui, A., Shahanaghi, K. (2011). An interactive robust data envelopment analysis model for determining alternative targets in Iranian electricity distribution companies, *Expert Systems with Applications*, 38 (8), 9830-9839.

[18] Omrani, H. (2013). Common weights data envelopment analysis with uncertain data: A robust optimization approach, *Computers Industrial Engineering*, 66 (4), 1163-1170.

[19] Ehrgott, M., Holder, A., & Nohadani, O. (2018). Uncertain data envelopment analysis. *European Journal of Operational Research*, 268(1), 231-242.

[20] Tsolas, I. E., & Charles, V. (2015). Incorporating risk into bank efficiency: A satisficing DEA approach to assess the Greek banking crisis. *Expert systems with applications*, *42*(7), 3491-3500.

[21] Schure, P., Wagenvoort, R., & O'Brien, D. (2004). The efficiency and the conduct of European banks: Developments after 1992. *Review of Financial Economics*, *13*(4), 371-396.

[22] Řepková, I. (2014). Efficiency of the Czech banking sector employing the DEA window analysis approach. *Procedia Economics and Finance*, *12*, 587-596.

[23] Lin, T. T., Lee, C. C., & Chiu, T. F. (2009). Application of DEA in analyzing a bank's operating performance. *Expert systems with applications*, *36*(5), 8883-8891.

[24] Luo, Y., Bi, G., & Liang, L. (2012). Input/output indicator selection for DEA efficiency evaluation: A empirical study of Chinese commercial banks. *Expert Systems with Applications*, *39*(1), 1118-1123.

[25] Liu, S. T. (2010). Measuring and categorizing technical efficiency and productivity change of commercial banks in Taiwan. *Expert Systems with Applications*, *37*(4), 2783-2789.

[26] Sokic, A. (2015). Cost efficiency of the banking industry and unilateral euroisation: A stochastic frontier approach in Serbia and Montenegro. *Economic Systems*, *39*(3), 541-551.

[27] Kamarudin, F., Sufian, F., & Nassir, A. M. (2016). Does country governance foster revenue efficiency of Islamic and conventional banks in GCC countries? *EuroMed Journal of Business*.

[28] Kamarudin, F., Sufian, F., Loong, F. W., & Anwar, N. A. M. (2017). Assessing the domestic and foreign Islamic banks efficiency: Insights from selected Southeast Asian

countries. *Future Business Journal*, *3*(1), 33-46.

[29] Berger, A. N., & Humphrey, D. B. (1997). Efficiency of financial institutions: International survey and directions for future research. *European journal of operational research*, *98*(2), 175-212.