Adv. Math. Fin. App., 2024, **9**(3), P. 817-839

Advances in Mathematical Finance & Applications
www.amfa.iau-arak.ac.ir
Print ISSN: 2538-5569
Online ISSN: 2645-4610
Doi: 10.22034/amfa.2023.1959465.1755

Original Research

Computing The Efficiency of Bank Branches With Financial Indexes, an Application of Data Envelopment Analysis (DEA) and Big Data

Fahimeh Jabari Moghadam[a], Farhad Hosseinzadeh Lotfi[a,] *, Mohsen Rostamy Malkhalifeh[a], Masood Sanei[b], Bijan Rahmani Parchkolaei[c].

[a] *Department of Mathematics, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[b] *Department of Mathematics, Central Tehran Branch, Islamic Azad University, Iran*
[c] *Department of Mathematics, Nour Branch, Islamic Azad University, Nour, Iran*

ARTICLE INFO

Article history:
Received 2022-05-28
Accepted 2023-01-01

Keywords:
Efficiency (Performance)
Clustering
Data Envelopment Analysis (DEA)
Data Mining
Big Data

ABSTRACT

In traditional Data Envelopment Analysis (DEA) techniques, in order to calculate the efficiency or performance score, for each decision-making unit (DMU), specific and individual DEA models are designed and resolved. When the number of DMUs are immense, due to an increase in complications, the skewed or outdated, calculating methods to compute efficiency, ranking and …. may not prove to be economical. The key objective of the proposed algorithm is to segregate the efficient units from that of the other units. In order to gain access to this objective, effectual indexes were created; and taken to assist, in regards the DEA concepts and the type of business (under study), to survey the indexes, which were relatively operative. Subsequently, with the help of one of the clustering techniques and the 'concept of dominance', the efficient units were absolved from the inefficient ones and a DEA model was developed from an aggregate of the efficient units. By eliminating the inefficient units, the number of units which played a role in the construction of a DEA model, diminished. As a result, the speed of the computational process of the scores related to the efficient units increased. The algorithm designed to measure the various branches of one of the mercantile banks of Iran with financial indexes was implemented; resulting in the fact that, the algorithm has the capacity of gaining expansion towards big data.
.

1 Introduction

Today, computing efficiency in various organizations and industries, is one of the essential procedures carried out, for the purpose of comparing, the amount of competitiveness credence, in the domestic and foreign scenarios of a country. Banks are not an exception in this respect. Hence, calculating the efficiency of banks and identifying the effective factors is extremely crucial. In DEA, abundant research has been performed to calculate the efficiency or performance of banks. DEA was initially utilized by Sherman [1] to compute the efficiency of banks. Charnes [2] proposed a model to

* Corresponding author. Tel.: 09123034649
E-mail address: farhad@hosseinzadeh.ir

attain a set of DMUs with multiple inputs and outputs. DEA classifies units into two groups, namely, the 'efficient and inefficient' units; and measures the efficiency score of each unit [3]. Cook [4] solved the conventional DEA models easily, with linear programming approaches. Complications relating to real-life problems, has given rise to DEA issues, with greater databases. When DEA, is big data related, two issues occur; 1- The number of inputs/outputs is high and 2- The number of DMUs are numerous. Computing efficiency, can be manipulated by using diverse models, which, based on the measured efficiency, can determine the efficient and inefficient DMUs. In the CCR model, m + s is present in limitation and n+1 exists as a variable; whereas, in the multiplier CCR model, n+1 has restrictions and m + s exists as a variable [2]. This results in an increment in the number of units or the number of variables, thus, increasing the complexities of calculations, leads to the query, as to the manner in which the models are implemented, when the sizes of m, s and n are exceedingly large. When the number of units increase, a significant volume of the linear programming models must be resolved. Processing big data by employing traditional methods is problematic or unfeasible. Big data emerged in the 1980's; and it can be reflected upon as, incrementing the volume of data, such that, it is difficult to analyze, process and store, by means of skewed database technologies. The 'big data' terminology correlates to data which are tremendously bulky, rapid or complex. Currently, big data is within the focus of modern scientific and commercial centers. In the DEA sphere, big data has created numerous problems for researchers. For example, when the DMUs are in extremely large numbers, due to an increment in computational complexities, a calculation of the efficiency of the DMUs is beyond the capacity of conventional methods. Given the application of DEA in the varied arenas of managerial and industrial aspects, including the growing acceptance of this technique, to evaluate the units under evaluation in big organizations, a rendering of methods which are capable of reducing computational issues and as a result, decreasing the calculation period, seems to be essential and beneficial. Numerous studies on the grounds of DEA, have been suggested for a large number of units, with the objectives of reducing, the computational period, for resolving linear programming models.,[5-6] similarly, in 1997.,[7] also proposed a hierarchical analysis approach, to reduce the implementation time, of big- sized DEA problems. Corhon and Sitari, rendered a lexicographic parametric programming method also, to decrease computational costs, when the efficient units are identified. , [8-15] were other studies in this domain. The key idea of these methods was to divide the units into smaller sets with the objectives of seeking an aggregate of efficient units. Subsequently, the efficiency of the inefficient units is calculated.

Zhu et.al.,[13] and Dellnitz.,[15], categorized the efficient units by classifying the n decision-making units (DMUs) into k+1 classifications then, by calculating the efficiency of each set, conducted the segregation of the efficient units. Dellnitz.,[15] developed the algorithm presented by Zhu et.al.,[13]. In this paper the same approach utilized by Zhu et.al, in relevance with classifying the data into smaller sets was followed. In order to seek the efficient units, Khezrimotlagh et.al.,[15], utilized and specified the criteria, where, units with the minimum input and the maximum output values, were concerned.In order to decrease the computational complexities, a new method has been rendered to segregate the efficient units from the inefficient ones, in this paper. Thereby, with the assistance of DEA concepts; and with due attention to the type of business, (in the current paper, being the bank), the relatively effectual indexes have been configured. Next, with one of the (two-step) clustering methods, the units are classified into efficient and inefficient classifications; and the DEA model is constructed with the efficient category. To assure the identification of the entire efficient units, the 'definition of dominance' is utilized and the efficient units which have remained, are specified and supplemented to the efficient classification of units. Thus, the model is updated. it is in this phase that; the efficiency of the inefficient

units is calculated. In actual fact, by segregating the efficient units from the inefficient, the number of units with which the DEA model is constructed, reduces. This results in a decrease in computational complications. To evaluate the validity of the proposed algorithm and determine the number of appropriate aggregates or sets for clustering, samples in sizes ranging from 100 to 9000 units were surveyed. The results illustrated that, as the amount of datum increments, the time-period for implementing the proposed algorithm decreases, then that of the time taken, when it is executed by the CCR model with an identical number of data. But, when the amount of data is meager, the execution time with the CCR model is less, or shows little or no variance, with the time taken, when implemented by the proposed model. In actuality, by increasing the number of units, a shorter time-period is required, to calculate the efficiency with the proposed algorithm. The Gams 23.4 Software has been taken advantage of, so as to calculate, the efficiency of the units and to employ the definition of dominance. The IBM **SPSS Modeler** 18 Software is utilized for clustering; whereas, the system adopted to implement is the Intel (R) Core (TM) i7-2670QM CPU @ 2.20GHz 2.20 GHz. RAM: 8.00 GB

In brief, the steps of the proposed method comprise of:

1. The data is classified into two sets, that is, efficient and inefficient data, by manipulating indexes, involving the two-stage clustering method

2. Specifying the entire efficient units, with the assistance of the concept of dominance; and supplementing them, to the efficient category or set

3. Calculating the efficiency of all the units, by resolving the problems of models with lower computational complexities, in comparison with classical models.

This paper is organized as follows:

In Section 2: A synopsis of the Subject Literature

In Section 3: The Proposed Method

In Section 4: An Empirical Example

In Section 5: Conclusion and suggestions for future undertakings.

1 Subject Literature

1.1 Clustering and Statistical Concepts

Two-Stage Clustering Method: In general, ‘clustering’ is a statistical method, for grouping observations, within a framework of sub-divisions which are similar to each other and are based on one or several characteristics. In the contemporary paper, a two-step clustering method has been employed. Clustering comprises of a large family of methods and algorithms; and the two-step method is extensively used in media marketing studies. In other words, the study, segmenting and gaining an overall profile of the samples under investigation, is the key purpose of this methodology. The two-stage method is applicable where big data and the hybrid use of sequential and quantitative data is concerned. This point forms its diversity with all the other clustering methods. In fact, it is the only

algorithm that can offer both quantitative and qualitative data. In the initial step, the observations are within a framework of primary clusters and these pre-clusters, are taken to be as a primary nucleus and are also considered as an observation. In the second step, the hierarchical method for categorizing these nuclei, which form the output of the prior step are utilized, comprising of similar observations that are within a nucleus or core.

Outliers and extreme data: The identification and elimination of outlier data is a process to eradicate and exterminate data, in other words is an operation which enhances the qualitative aspects of datum. One of the foremost reasons for effective data mining is an efficacious analysis and a beneficial aggregate from an assortment of observations with due attention to their behavior., [16-17].
The analysis of outlier or aberrant data was studied from the beginning of the 19th century, in the statistical society. *'Outlier'* data signifies to data which is out of range. At times outlier data can be an aggravation and hassle; and sometimes the problem itself, could refer to the detection of outlier(s). Outliers are not erroneous data. These are datum, which are only, at a distance, from the distribution of records. One of the approaches for classifying outlier data, is the distribution-based method. In these methods, statistical models are developed for the data-sets; and then a statistical test is imposed, to determine, as to whether, a data conforms and fits in with this model, or not. Data with a low probability, of belonging to the statistical model are declared as 'outliers '. This approach is divided into two classifications, known as, parametric and non-parametric. In the parametric methods, it is presumed that, a principled distribution is pursued, such as, a normal distribution; whereas, in the non-parametric methods, this assumption does not exist. In this paper, in order to identify the outlier data, a parametric and a univariate method have been used. In the univariate approach, a single variable is surveyed at a time to detect outlier data. The method to identify, is such that, a normal distribution is taken into consideration for the data. Outlier data in this case, refers to the observations, which are of an immense distance, in relative to the mean central point, in context with the standard deviation index. This conception is denoted from the properties of a normal distribution; 99.7 percent of the data is between $\mu + 3\sigma$ و $\mu - 3\sigma$. In order to describe another label for outliers and excessively outdated data, $\mu + 5\sigma$ و $\mu - 5\sigma$ is taken into consideration; and in this paper, for the extreme data, the Modeller spss 18 Software has been utilized.

1.2 Data Envelopment Analysis (DEA) Concepts

- **The CCR Model:** Let us assume that we have "n" DMUs. Each DMU_j ($j = 1, \dots, n$) has the $X_j \geq 0$, X_j (x_{1j}, …, x_{mj}) input vector and utilizes it to produce the output vector $Y_j(y_{1j}, \dots, y_{sj}) Y_j \geq 0$. The relative efficiency of DMU_o comes to hand from the model hereunder:

$$
\begin{aligned}
&\min \theta_o \\
&s.t \ \sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_o x_{io} \qquad i = 1,\dots,m \\
&\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro} \qquad j=1,\dots,s \\
&\lambda_j \geq 0 \qquad j=1,\dots,n \\
&\theta_o \qquad \text{free}
\end{aligned}
\tag{1}
$$

$pps = \{(x, y) | x \ \ can\ produce\ \text{y}\}$ (possibility production set)

By observing the subjectivity comprising of constant returns to scale (CRS), convexity, feasibility, minimal interpolation of set, the possibility of production is as given below. [2]

$$\text{T}_{CRS} = \left\{ (x, y) \middle| x \geq \sum_{j=1}^{n} \lambda_j x_j, y \leq \sum_{j=1}^{n} \lambda_j y_j, \lambda \geq 0 \right\}$$

DMU_o is efficient, if a better or more enhanced point, than this, does not exist in the constructed pps.

DMU_o is inefficient, if and only if a better point than that is found.

Theorem 1: If DMU_t is inefficient, then in every optimal response (1)$\lambda_t^* = 0$

(1) can be constructed without taking any inefficient t unit under consideration.

$$\begin{aligned} &\min \theta_o \\ &s.t \ \sum_{\substack{j=1 \\ j \neq t}}^{n} \lambda_j x_{ij} \leq \theta_o x_{io} && \text{i=1,...,m} \\ &\sum_{\substack{j=1 \\ j \neq t}}^{n} \lambda_j y_{rj} \geq y_{ro} && \text{j=1,...,s} \\ &\lambda_j \geq 0 && \text{j=1,...,n } j \neq t \\ &\theta_o && \text{free} \end{aligned} \tag{2}$$

- **Definition of dominance**

Definition 1: $DMU_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ dominates $\text{DMU}_2 = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$ if and only if:

- ✓ DMU_1 is not worse than DMU_2 in the entire components
- ✓ It should be significantly better than DMU_2, in one component, to the minimum. That is:

$$\begin{pmatrix} -x_1 \\ y_1 \end{pmatrix} \begin{matrix} \geq \\ \neq \end{matrix} \begin{pmatrix} -x_2 \\ y_2 \end{pmatrix} \tag{3}$$

2 Proposed Model

If we are confronted with a restricted amount of data in DEA, the efficiency of units is computed according to the traditional method. But when big data is involved, due to the large volume of data, computational complexities increase. The expression of 'big data' designates to data which is extremely

colossal, rapid or complex. It is difficult or infeasible to process big data with conventional techniques. The day-to-day information in the commercial world is in a growing status; and thus, brings about the presence of a large volume of data. So as to analyze the socio-economic goals, scrutinizing such a great volume of data is a requirement. Hence, it is essential to render methods and algorithms which can assist in reducing these complications in relative to calculations. In general, we conduct such tasks to economize on costs. For instance, to compute an immense or big volume of data efficiency, it is essential to adapt to methods, which are expeditious, in this ever-changing world. In the current paper, an algorithm to reduce the computational complexity is suggested; where, by denoting the effective indexes, data is classified into two, consisting of efficient and inefficient units. (Henceforth, instead of using the term of a category comprising of efficient units; the term, a category of *"efficient units"* will be utilized. By this, we signify that, in the sets or classifications of efficient units, there is a probability that, in these sets, inefficient units also exist); and consequently, the DEA model is only implemented on the efficient set of units. On the basis of Theorem 1, in Section (2.2), the DEA model can be configured without taking the inefficient units into consideration. By eliminating the inefficient units, the amount -of units decreases outstandingly; and as a result, computational complexities decline.

2.1 Steps of the Algorithm

Step 1: Normalizing data with an average norm

Our main objective is to segregate the efficient units from the inefficient ones; and our approach for this purpose is to utilize one of the clustering methods. In data mining processes, such as, classifying and clustering, it is essential to prepare the data for the algorithm. As normally, unprocessed data cannot be injected, into data mining and machine learning algorithms. This calls for the necessity to formulate data, with due attention to the conditions and the problem, for which, appropriate data is infused into the data mining algorithm. In order to prepare the data, we need to extract it from its original form and condition, adjusting it, according to the configuration which is suitable for the algorithm. So that clustering algorithms are implemented in a fair manner as to the data, initially, we normalize the data. In performing the normalization, the various dimensions are surveyed in an unbiased and fair way by the algorithm and the impact of one, does not overpower the remaining. In this paper, an average norm is employed to normalize the data. This is done in such a mode, that, for each factor (every column), the mean data is calculated primarily. Subsequently, the value of each observation which corresponds to that factor, is then divided by the average.

$$\frac{v_i}{\frac{\sum_{i=1}^{n} v_i}{n}} \qquad \text{i=1,...,n} \tag{4}$$

Step 2: Defining the relativity of the effectual indexes, given the evaluation objectives and type of business

After the performance of normalization, there is a necessity to define the relativity of the effective indexes, by taking advantage of the evaluating concepts. Since, in this paper, the objectives of clustering, is to separate the efficient units, in such a manner that, with the minimum of error, the efficient units are specified. (The presence of an inefficient unit or inefficient units, in a category or set

of efficient units does not pose problems for an algorithm). Hence, in defining the indexes, two issues are taken into consideration:

1. In accordance with the type of business, with the assistance of experts, input and output factors, which have an impact on units are identified
2. DEA concepts and the definition of efficiency in defining indexes are utilized. This is due to the fact that, just a mere use of input and output variables cannot be an appropriate criterion to denote the efficiency and inefficiency of units. The defining of several indexes given the business and DEA concepts, will assist greatly in detecting and segregating the efficient units. So as to get a better understanding of the requirement to construct indexes, we will take Unit o in Fig. (1) into consideration. This unit has an extremely high output, though, on the basis of the DEA definition it is not an efficient unit; as by consuming an extremely high input, it has attained such an output; that is, although, the output of Unit o is high, the ratio of output to input is inacceptable. Thus, Unit 0 is inefficient. As a result, to increase the capacity of differentiation, between the efficient and inefficient units, after normalizing the data in the ratio of each output to each input, is utilized in the definition of indexes. Let us assume that n observations with m input factors and s output factors are present. The indexes using the equation

$$\forall j,\forall i,\ \forall r \quad \frac{y_{rj}}{x_{ij}} \qquad r=1,...,s \quad i=1,...,m \quad j=1,...,n \tag{5}$$

are constructed. With due attention to the amount of m inputs and s outputs, the *m*n* index is configured with equation (5). From amidst these indexes, according to the type of data and by taking the opinions of the experts into view, regarding business, indexes, which have a greater influence in terms of the efficiency of DMUs, can be recognized and these indexes can be taken into consideration for performing clustering.

Step 3: The classification of units into two categories of efficient and inefficient with the indexes of Step 1 and one of the clustering approaches.

Even though, in major, the algorithms in relevance with clustering are of a uniform basis; but there are diversities as to the manner in which, comparisons (similarities) and or the distances are measured and likewise, the selection of labels for the items of each cluster, are in these methods. In the contemporary paper, the two-step clustering approach has been utilized, (1.2) is the two-step algorithm method, which is able to provide the qualitative and quantitative data concurrently; and in other words, are algorithms which are applicable for big data. As the main objective for our undertaking, is in relative with big data, we preferred to exploit the two-step clustering method in this paper. In order to perform clustering and identify the efficient units, it is essential that we distinguish the number of clusters. It is certain that, as the number of units increase, the number of clusters required, so as to segregate the efficient units will modify. After detecting K, appropriate to the two-step approach, the clustering of units will be dealt with and ultimately, the units will be divided into two categories, that is, the efficient and inefficient units. These are known as J, the overall data; *I.e.*, the efficient $\overline{J}$ and inefficient $\hat{J}$ data sets.

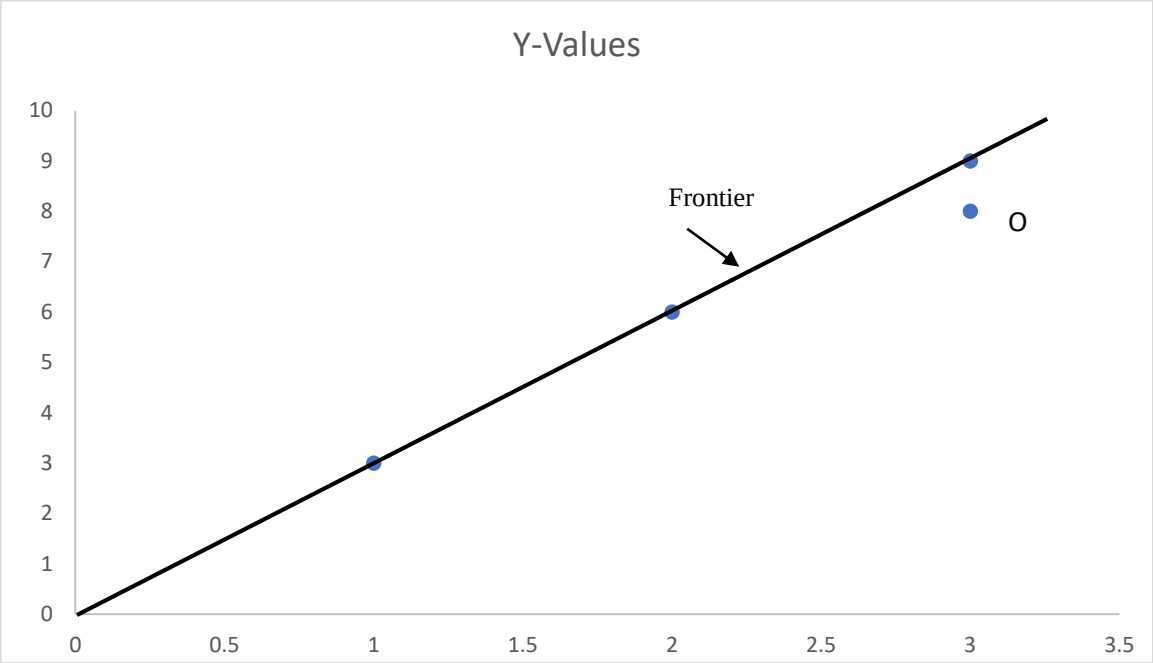


Fig.1: frontier

Step 4: After specifying the efficient units, the DEA model is configured with the efficient units, such that, in writing the DEA model, the efficient set of units are only utilized. If $\bar{J}$ is the efficient set, the CCR model for computing the efficiency of DMU_p will be in the form of (6).

$$\begin{aligned} &\min \ \theta \\ &\text{s.t} \quad \sum_{j\in\bar{J}} \lambda_j x_{ij} \le \theta x_p \qquad \text{i=1,...,m} \\ &\qquad \sum_{j\in\bar{J}} \lambda_j y_{rj} \ge y_p \qquad \text{j=1,...,r} \\ &\qquad \lambda_j \ge 0 \qquad\qquad \text{j}\in \bar{J} \end{aligned} \tag{6}$$

Step 5: Clustering was performed in Step 3, though, there is a possibility that, with the performance of clustering, some of the efficient units are not identified; so consequently, in order to specify the probable remaining efficient units, the concept of 'dominance' in DEA, is taken advantage of and the efficient units, which have remained behind, are identified.

As stated in Section (2.2), in accordance with the definition of dominance, if unit t is a member of the efficient set of units; and unit p is of the inefficient set, when, the equation $\begin{pmatrix} -X_t \\ Y_t \end{pmatrix} \gtreqless \begin{pmatrix} -X_p \\ Y_p \end{pmatrix}$ is stipulated, we proclaim that, DMU_t is *dominant* over DMU_p, or in other words, DMU_p is dominated by DMU_t. The entire members of the inefficient set $\hat{J}$, which has been dominated, in the minimal, by one member of the efficient $\bar{J}$ set of units, is definitely recognized as an inefficient unit. We ascribe, $\hat{J}_{1.}$ to the members comprising of such an aggregate. But there is a probability, that, (in the case), of members of the inefficient set, which have not been dominated by any of the members of the efficient set, are efficient units; as it could be feasible, that they may be dominated by a combination or hybrid of efficient

units. We signify an aggregate of such members as, $\hat{J}_{2.}$ So as to detect whether, these members are efficient or not, their efficiency is computed by *Step 4* of the DEA model. Four diverse conditions could occur for units under evaluation.

1. $A_1 = \left\{ J \middle| DMU_j \text{ in model (6)which is inefficient } (\theta \prec 1) \right\}$
2. $A_2 = \left\{ J \middle| DMU_j \text{ in model (6) which is efficient}(\theta=1) \right\}$
3. $A_3 = \left\{ J \middle| DMU_j \text{ in model (6) which is super-efficient}(\theta \succ 1) \right\}$
4. $A_4 = \left\{ J \middle| DMU_j \text{ in model (6) which is infeasible} \right\}$

As a result, our new classification, with due attention to the results attained from the dominance and implementation of Step 4, (a model which was configured, with an efficient set (of units), gained from the clustering constructed); modifies as follows:

$$\overline{J_{NEW}} = \bar{J} \cup A_2 \cup A_3 \cup A_4$$
$$\widehat{J_{NEW}} = \hat{J}_2 \cup A_1$$

Step 6: Writing a DEA model with a new set of efficient units till all the efficient units are specified.

Step 7: Computing the efficiency of a set of inefficient units with the final DEA model of Step 6.

The steps of the algorithm are summarized as given below:

Start

Step1: Normalizing the data

Sep 2: Defining the effectual indexes, with due attention to DEA concepts and type of business

Step 3: Classification of data into two aggregates of "efficient and inefficient"; with the indexes of Step 2 and clustering methods

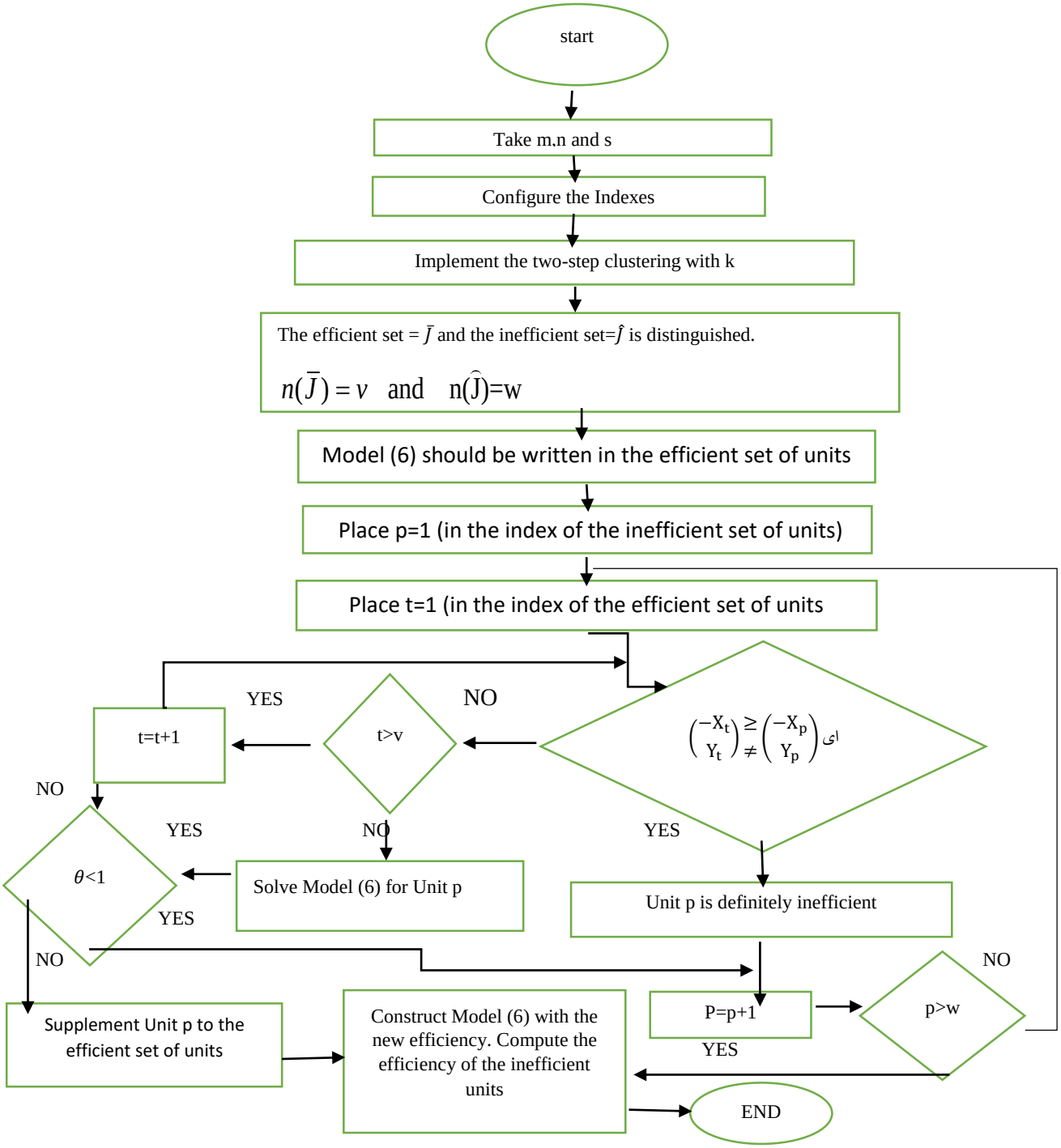
Step 4: The configuration of a DEA model with efficient units

Step 5: Utilizing the definition of dominance to specify the efficient units which have remained behind in the set(s) of inefficient units

Step 6: Restoration of the 4th Step of the DEA model with the new and efficient members of the 5th Step, till the entire efficient units are identified

Step 7: Calculating the efficiency of the inefficient units with the final model of the 6th Step.

END



2.2 Stages to Seek the Number of The (K) Clusters and A Survey to Validate the Algorithm

In order to gain assurance of the presence of efficient units, obtained from clustering and to specify the number of clusters with which, the maximum efficient units can be identified; and likewise, locating the rate of (percentage), identified for the efficient units, as well as calculating the impact of the

proposed algorithm, as to the computation time-period, Stages 1 to 12 that have been rendered below, from each one of the intervals of [250, 350], [100, 200], [1400, 1500], [700, 800], [3000, 4000] and [4000,…] were executed. One of the cases of these samples, was in relevance with the elimination of extreme data, of the samples in the sizes of n=1459, n=729, n=291 and n=145. In actuality, by implementing the Stages 1 to 12, the validity and accuracy of the algorithm is investigated.

In the entire samples, the following steps were performed to find the appropriate number of clusters and to survey the verification of the algorithm:

1. The efficiency of the units was computed by the CCR model and the efficient units were identified. The 'genuine efficiency' and the implementation time-period was recorded. (It should be noted that, during the stages of implementing the proposed algorithm, information regarding the efficiency scores of units; and as to whether, which of the units are efficient, is not at hand. At this stage, it is only to investigate the validity of the algorithm, that the efficiency score is calculated)
2. After normalizing the data, the effective relativity indexes are configured. (Based on the first step of the proposed algorithm)
3. Two-step clustering method was implemented on the constructed indexes, commencing from k=2. In stages 3 and 4, the objective is to find a k which is suitable for clustering. (According to step 2 in the proposed algorithm)
4. The percentage of the efficient units present, in the configured sets were computed from clustering. Two termination conditions (termination conditions are utilized only for attaining the k number of clusters and a percentage for identifying efficient units. In implementing the stages of the algorithm, we have no information as to the efficient units, as a result, employing the termination condition, is pointless). For the number of denoted clusters; and with due attention to the specified percentage of the efficient units, in varied circumstances, a number of appropriate (k) clusters were identified for clustering financial data. (In this paper, k=4, was referred to, as the number of clusters, which were appropriate).
5. Constructing a DEA model with a set of efficient units attained from clustering
6. Utilizing the definition and concept of dominancy, for the specification of the presence of probable efficient units, in the set of inefficient units. If an efficient unit is found, it is supplemented to the efficient set(s) of units
7. With the efficient set(s) of Step 6, a new model of DEA comes to hand
8. The efficiency computation of the inefficient units with the model of Step 7 (estimation of efficiency)
9. Comparing the efficiency scores of inefficient units obtained from Steps 1 to 7, implies the comparison of the actual efficiency scores, with that of the efficiency estimates
10. Calculating the time-period of computations with the proposed algorithm
11. Formulating a time-table to compare the time-period for computing efficiency with the CCR model and the time-period for the proposed algorithm
12. Conclusion.

3 Example: Utilizing the Data of One of the Mercantile Banks of Iran in Varied Sizes to Algorithm

As mentioned previously, so as to achieve a better result and to gain access to a suitable (k), more than 40 samples were investigated. The results of the algorithm executed on 27 samples, has been illustrated in Tables 1 to 15. For 4 examples (n=145, n=291, n=729 and n=1459), was implemented in two cases, without eliminating the observations, the algorithm was executed, by safeguarding the extremely outlier data, descriptions of the very outmoded data have been given in section 2. To specify the extremely outlier data, we shall initially execute the Step 1,2 of the algorithm, that is, indexes are configured and consequently, for identifying the extreme data, we employ the sixth index. In Section 2.1 explanations have been rendered that, the identification of extreme data has taken place on the distribution-based method and the univariate approach. Results for implementing the steps of the algorithm have been given in Tables 1 to 15. To perform the stages for validating the algorithm, initially all the efficient samples of units were computed by the CCR method and the efficient units are specified in each sample. Besides this, after normalizing the data on an average norm, the relativity indexes are configured, these have been explained in Step 2 of the proposed algorithm, which defines the relativity and effectual indexes, with objectives of evaluation and the type of business performed. With due attention to the fact that, the data within our access is the data relevant to one of the mercantile banks of Iran, indexes suitable to the bank data were introduced. This data consisted of 2 input factors and 1 undesirable output factor, (which we considered as an input) and are the 5 output factors were verified. The input factors comprise of 1) Staff Privilege, 2) the Interest paid, 3) deferred receivables/ Demand (undesirable output) and output factors encompassing 1) Facilities, 2) Sum of Deposits, 3) Interest earned, 4) Bank charges or Fee received and 5) other sources.

Privilege of Staff: Consists of a hybrid of the digit, training, executive positions and experience of personnel or staff of a branch.

Interest Paid: This is the 'Interest' which the bank pays in exchange for the deposits of clientele.

Deferred Receivables/Demand: Non-current demands, which, in banking literature is reputedly known as 'delinquent receivables' and are the result of the occurrence of credit risks. The non-current demand is in actuality, the demands which have failed to be paid at the appointed time. This lack of timely payment, jeopardizes the profitability of the bank in numerous ways.

Facilities: Types of long-term and short-term loans and types of credit cards.

Sum of Deposits:

1. Quarterly short-term investment deposits (or three-monthly deposits)
2. Short-term six-monthly investment deposits
3. Long-term investment deposits (for 1 - year duration)
4. Long-term investment deposits (for 2-year duration)

Profit Earned: Profit that the bank receives from customers in exchange for the facilities provided.

Fee or Bank Charges Received: In 'Gharz-al-Hassaneh' banks, facilities are granted without interest and are only in return for a fee or bank charges range from 0 to 4 percent. The fee is utilized to cover the operating costs of such banks. These (expenses) include, the security and maintenance of the

technical and software infrastructure, current maintenance costs of the branches are allotted to costs such as, manpower, rentals, telephone, electricity, gas and…. payments.

Other Sources:

1. Tax Revenues
2. Incomes achieved from government monopolies and taxes
3. Incomes obtained from services and sale of goods
4. Gas and Hydrocarbon (oil) revenues
5. Insurance Premiums
6. Grants
7. Transfer-related and miscellaneous incomes.

As explained, the utilization of input and output factors, cannot be an appropriate criterion to identify the efficient units exclusively. The input and output values alone, are incapable of rendering sufficient information, as to the units being efficient or inefficient. But by using the relativity or ratio of outputs as to the inputs, a more absolute information is attained to evaluate the units obtained. As a result, defining indexes that are effective in the business under study, using the evaluating concepts in DEA and the help of experts, in the business within survey, can be effective and beneficial to increase the dissimilarity between the efficient and inefficient units, after normalization of data in relevance with each input, to that of each output. If, in lieu of each input and each output the ratio or relativities of every output to each input is constructed, with 3 inputs and 5 outputs, 15 indexes come to hand. Solving the problem with 15 indexes also renders a wholesome result; but, according to the prior discussion, the relativities employed, are those, which have a greater impact as to the efficiency of banking units. (Our objective is to configure effective indexes and likewise a reduction in computational complexities). On all the samples, the varied modes of $\frac{y_r}{x_i}$ were investigated; and finally, the relativities which had insignificant impact on detecting the efficient units were eliminated and we selected the effectual relativities or ratios as indexes to be utilized in the classification. Five indexes given hereunder are defined:

$$1.\frac{Deferred\ \ Receivables}{Sum\ of\ \ Deposits}$$

$$2.\frac{Interest\ \ Paid}{Sum\ of\ \ Deposits}$$

$$3.\frac{Deferred\ \ Receivables}{Facilities}$$

$$4.\frac{Sum\ of\ \ Deposits}{Staff\ \ Privilege}$$

$$5.\frac{Interest\ or\ Profit\ Earned}{Staff\ \ Privilege}$$

Whereas, defining the efficiency in DEA, as $E = \frac{u_1 y_1 + \cdots + u_s y_s}{v_1 x_1 + \cdots + v_m x_m}$ (when we have s outputs and m inputs); and utilizing the $\frac{total\ outputs}{total\ inputs}$ of 6 was rendered with normalized data; in such a manner that, after normalizing of data with the average norm, the sum of the input factors and that of the output factors were brought to hand and then the ratio of the total output to the total input was formed; and this index was known to be the sixth index for classification purposes. It should be noted that, prior to the introduction of the indexes, all the factors are normalized with the average norm. After performing the normalization, the sixth indexed that is introduced, is configured.

After which, with two-step approach and the sixth index, which has been presented, classification takes place. In order to halt the classification, we have taken two conditions under consideration, (it should be observed that, the two conditions for termination, are only for the stages to seek a number of appropriate clusters. In the initial stage, for the phases to find a number of suitable clusters and to survey the validity of an efficient algorithm, units were computed, until a number of suitable clusters are attained. In the steps of the proposed algorithm, (information regarding the efficient units is not available). By being attentive to the fact that, in the preliminary stage to survey the legitimacy and find the suitable number of clusters, the efficiency of the units was computed. Termination conditions are in the two following forms:

1. Identification of the entire efficient units
2. Case 1$\big||E_K| - |E_{K+1}|\big| \leq 1$

In case 1) the smallest number of (k) clusters which identifies the entire efficient units is desirable. In other words, incrementing the number of classifications, till the time that, the data of all the efficient units are specified. In case 2), $|E_K|$ is the number of efficient units identified by classification (k); and $|E_{K+1}|$ is the number of efficient units identified by the k+1 classification. Case 2) states that, if by increasing the number of classifications from k to k+1, a new efficient unit is not specified, the classification is paused. Following the performance of classification with the IBM **SPSS Modeler** 18 Software, and the two-step clustering method, the number of efficient units identified was surveyed with a varied number of classifications.

Table 1: Clustering by Two-Step Method (Without Deleting Observations)

Numbers of units	input	output	Number of clusters	Identification percentage	Number of efficient units	Identification number
145	3	5	2 3 4 5	%75 %87.5 %87.5 %100	8	6 7 7 8
291	3	5	2 3 4 5	%83.3 %83.3 %83.3 %100	12	10 10 10 12
729	3	5	2	%100	7	7
1459	3	5	2 3 4	%83.3 %83.3 %100	12	10 10 12

Our objective is in implementing the above-mentioned stages; and in contemplating upon the several sample sizes, for finding an appropriate k, for classifying units. The results of executing the algorithm on the samples is demonstrated in Tables 1 to 15. At this stage, seeking k (number of suitable clusters) is crucial. In Table (1), the results of clustering pertaining to 4 samples *i.e.*, 145, 291, 729 and 1459 have been shown. As can be observed, in all the samples, a cent per cent detection of the efficient units has occurred, but with a number of diverse clusters. In the sample with a size of 729, with a k=2 number of clusters, all the efficient units were specified, though, with the number of units of less or more than 729, the minimum number of clusters required is 4.

Table 2: Clustering by Deleting Extreme Data

Number of units	input	output	Number of clusters	Identification percentage	Number of extreme data	Number of efficient units	Identification number
144	3	5	2 3 4	%50 %68.5 %100	1	16	8 10 16
289	3	5	2 3 4	%70.5 %79.4 %100	2	17	12 13 17
722	3	5	2 3 4 5 6 7	%66.66 %73.3 %73.3 %73.3 %73.3 %100	7	15	6 11 11 11 11 15
1444	3	5	2 3 4 5 6	%47 %88 %88 %94 %100	15	17	8 15 15 16 17

Table 3: Condition $\left||E_K| - |E_{K+1}|\right| \leq 1$ without deleting any observations

Number of units	Number of clusters	Identification percentage
145	2 3	%75 %87.5
291	2	%83.3
729	2	%100
1459	2	%83.3

As a regulated tendency was absent in this Table, these four samples were investigated, whilst extreme data was eliminated from them. In Table (2), the results of classification of samples in sizes of 145, 291, 729 and 1459 from which, extreme data is eradicated has been collected. Results show that, by eliminating very skewed data for a number of 722 and 1444 data, for a number of clusters to identify the efficient units finds an increase. By imposing the second halting condition *i.e.*, $\left||E_K| - |E_{K+1}|\right| \leq 1$, Tables (3 and 4) come to hand, in two conditions being, with and without the elimination of extreme data.

Results express that, from the 4 Tables under discussion, it can be stipulated that, with k=4, *i.e.*, with the number of clusters equivalent to 4, in the minimal, 83 percent of the efficient units, are specified by the two-step approach by clustering.

Table 4: Condition $\left||E_K| - |E_{K+1}|\right| \leq 1$ by deleting extreme data

Number of units	Number of clusters	Identification percentage
144	2 3 4	%50 %68.5 %100
289	2 3 4	%70.5 %79.4 %100
729	2 3	%66.66 %73.3
1444	2 3	%47 %88

Table 5: samples at a distance [100,200]

Number of units	Number of clusters	Identification percentage	Number of efficient units	Identification number
132	2 3 4 5	%85 %85 %85 %85	7	6 6 6 6
156	2 3 4	%75 %75 %100	12	9 9 12
131	2 3 4 5	%66 %83 %83 %83	12	8 10 10 10
145	2 3 4 5	%75 %87.5 %87.5 %100	8	6 7 7 8
144	2 3 4	٪50 ٪68.5 ٪100	16	8 10 16

Table 6: samples at a distance [250,350]

Number of units	Number of clusters	Identification percentage	Number of efficient units	Identification number
292	2 3	%69.23 %100	13	9 13
280	2 3 4	%92 %92 %92	13	12 12 12
303	2 3 4	%87.5 %87.5 *%100	8	7 7 8
291	2 3 4 5	%83.3 %83.3 %83.3 %100	12	10 10 10 12
289	2 3 4	%70.5 %79.4 %100	17	12 13 17

Table 7: Samples at a Distance [700,800]

Number of units	Number of clusters	Identification percentage	Number of efficient units	Identification number
733	2	%100	12	12
715	2 3 4 5	%72 %72 %72 %72	11	8 8 8 8
731	2 3 4 5	%80 %80 %80 %80	10	8 8 8 8
729	2	%100		
722	2 3 4 5 6 7	%66.66 %73.3 %73.3 %73.3 %73.3 %100	15	6 11 11 11 11 15

Table 8: samples at a distance [1400,1500]

Number of units	Number of clusters	Identification percentage	Number of efficient units	Identification number
1446	2 3 4	%77.77 %77.77 %77.77	18	14 14 14
1464	2 3 4	%82.35 %82.35 %82.35	17	14 14 14
1414	2 3 4 5	%78.5 %78.5 %78.5 %100	14	11 11 11 14
1459	2 3 4	%83.3 %83.3 %100	12	10 10 12
1444	2 3 4 5	%47 %88 %88 %94 %100	17	8 15 15 16 17

Table 9: samples at a distance [3000,4000]

Number of units	Number of clusters	Identification percentage	Number of efficient units	Identification number
3038	2 3 4	%75 %75 %75	32	24 24 24
3104	2 3 4	%77 %77 %77	22	17 17 17
3030	2	%100	24	24
3000	2	%100	12	12

Table 10: samples at a distance [4000, …]

Number of units	Number of clusters	Identification percentage	Number of efficient units	Identification number
4568	2 3 4	%81 %81 %81	37	30 30 30
4545	2 3	- %100	15 15	- 15
9000	2	%100	16	16

Table 11: Condition $\left||E_K| - |E_{K+1}|\right| \leq 1$ at a distance [100,200]

Number of units	Number of clusters	Identification percentage
132	2	%85
156	2	%75
131	2 3	%66 %83
145	2 3	%75 %87.5
144	2 3 4	%50 %68.5 %100

Table 12: Condition $\left||E_K| - |E_{K+1}|\right| \leq 1$ at a distance [250,350]

Number of units	Number of clusters	Identification percentage
292	2 3	%69.23 %100
280	2	%92
303	2	%87.5
291	2	%83.3
289	2 3 4	%70.5 %79.4 %100

Table 13: Condition $\left||E_K| - |E_{K+1}|\right| \leq 1$ at a distance [700,800]

Number of units	Number of clusters	Identification percentage
733	2	%100
715	2	%72
731	2	%80
729	2	%100
722	2 3	%66.66 %73.3

Due to the fact that, a distinct response cannot be achieved by investigating only 8 samples, three more samples, from each size must be selected and clustering with the configured indexes implemented on it, with each of the two halting conditions. As there was no access to data with sizes of more than 1,500 branches, samples relevant to sizes 3000 to 9000 were generated with the IBM **SPSS Modeler** 18 Software and according to the steps of the algorithm, these samples were surveyed. Samples related to sizes 3000 to 9000 are illustrated in Tables (9 and 10).

For example, in the sample consisting of a size of 3030, all the efficient units were identified with (k=2). Though, in the sample having a size of 3038, seventy-five percent of the efficient units were recognized with (k=3); (Table 9). Likewise, for a sample with a size of 9000, (k=2) identified the entire efficient units (Table 10). The results of all the samples are given in Tables (5 to 14).

Table 14: Condition $\left||E_K| - |E_{K+1}|\right| \leq 1$ at a distance [1400,1500]

Number of units	Number of clusters	Identification percentage
1446	2	%77.77
1464	2	%82.35
1414	2	%78.5
1459	2	%83.3
1444	2 3	%47 %88

To summarize the results of the Tables with different sizes and attaining a number of suitable clusters for classifying each sample with the smallest k, such that, it detects the maximum number in identifying the efficient units and this is given in Table (15). It can be said that, the smallest k for the various sizes ranges from k=2 to k=4. Hence, it can be claimed that, with k=4 classifications, we will probably have the maximum number of efficient units identified in the assorted samples. As a result, for the financial (banking) data, a number of k=4 has been introduced for clustering purposes.

Table15: Find the best K

Number of units	The smallest K	Highest detection percentage
[120,160]	4	%100
[250,310]	3	%100
[700,750]	2	%100
[1400,14100]	4	%100
[3000,3150]	2	%100
[4500,4600]	2	%100

In continuation, based on Step 5, the DEA model is constructed with the classified efficient units that have come to hand. Since, it is possible that with clustering all the efficient units are not recognized, in order to continue on this path, the 'definition of dominance' will be utilized. In other words, it could be possible that a specific sample is at our access and this sample is unable to detect all the efficient units with k=4. As a result, it is essential that, the units which have been impeded, are identified with an approach; resulting in the usage of the definition of dominance. For example, 131 units, from the first step is taken under study. After computing the efficiency score of all the units, 12 efficient units were located. (It should be observed that, in order to survey the validity and accuracy of the algorithm, the

efficiency of the units have been calculated and the efficient units identified). Six of the mentioned indexes have been configured and the clustering is performed with k=4, in a two-step approach. Executing the clustering with k=4, leads to the identification of 10 units out of the 12 efficient units. The DEA model is constructed out of a set of efficient units. So as to specify the other efficient units on the basis of Step (6) from the steps to verify the algorithm; and Step 4 of the proposed algorithm utilizes the definition of dominance. As stated previously, in this sample, 12 efficient units are present; and 10 units from this, have been detected by clustering. Units 16 and 105 have not been identified. In this sample, by classifying the data, the set comprising of efficient units has 31 members. (In section 3, it has been elucidated that, in the set comprising of efficient units, it could be possible that, inefficient units can also be present and this issue does not impose a void on the algorithm). The set of inefficient units consist of 100 members. Following the comparison of the data with the definition of dominance, 14 units from that of the inefficient classification were not dominated by any member of the efficient set. Thus, according to the steps of the algorithm, these 14 units were constructed with a DEA model and with the efficient classification or set (Step 5), which we then evaluate. The efficiency scores of which, are given below: DMU_{16}=2.33 DMU_{21}=0.33 DMU_{22}=1.20 $DMU_{23=}$0.16 DMU_{24}=0.43 DMU_{56}=0.83 DMU_{57}=0.28 DMU_{60}=0.12 DMU_{63}=0.37 DMU_{72}=0.48 DMU_{105}=2.92 DMU_{11}=0.24 DMU_{128}=0.35 DMU_{129}=0.10

Units 105, 22 and 16 have efficiency scores which are higher than 1, thereby, they are of a super-efficiency classification. As a result, these 3 units are conveyed from the inefficient to the efficient set. As observed, the two units 105 and 16, were units of similar performance, which were not specified by clustering. This results in the fact that, by performing a comparison, with the definition of dominancy or being dominant, efficient units, which had not been specified, were identified. Now, the new DEA model is constructed with the novel efficient set and the efficiency scores of the inefficient units are computed by this model. Results of the survey are collected in Tables (1 to 15). An overall conclusion which could be drawn, is that, with the clustering method and the definition of dominance, classifications comprising of efficient units can be identified. To study the accuracy of Steps 3 and 4 of the proposed algorithms, after specifying the efficient units, the DEA model is configured with the set consisting of efficient units and the efficiency scores of the inefficient units is calculated with this model; and these scores are compared with the efficiency scores, that have been achieved by solving the DEA model with the entire units. (Initially, when the algorithm is commenced, the efficiency of all the units was calculated with the CCR model). In actual fact, we traversed two paths:

Path 1: The utilization of n units in configuring a DEA model and the computation of n units

Path 2: Using the classification of efficient units in constructing the DEA model and calculating the efficiency of the n unit with the constructed model.

Hence, there are two efficiency scores present for each DMU. $E\epsilon R^n$ signifies the efficiency of the first path and $E^{/}\epsilon R^n$, is the efficiency for the second path. The elimination of efficient units, has no impact on the efficiency scores of units. (Refer to 2.2). As a result, $\left\|E - E^{/}\right\| \leq \varepsilon$ is always prevailing. This results in the fact that, the condition $\left\|E - E^{/}\right\| \leq \varepsilon$, indicates the algorithm's precision. We have designated that our algorithm has the capacity of detecting the efficient units. Though, in addition, it must be proven, that in the presence of big data, the time-period for implementing the proposed algorithm is less than that of the CCR model, for the calculation of the 5 samples, 279, 733, 1459, 4545 and 9000. Results have been collected in Table (16).

Table16: Runtime Comparison

Number of units	CCR	Proposed Method
279	00:00:36.132	00:00:40.015
733	00:01:44.163	00:01:43.552
1459	00:04:12.616	00:03:19.880
4545	00:19:33.248	00:13:07.529
9000	00:59:19.738	00:27:57.878

As is noted, the time duration for implementing the proposed algorithm is more for sample 279. However, by incrementing the size of the sample, the time taken to execute the proposed algorithm, witnesses gradual decrement, in relevance to solving with the CCR approach; such that, in the sample size of 9000 the implementation time for the proposed algorithm is even less than half of the time duration taken, when executed with the CCR model. The execution time of the proposed algorithm was 27: 57.878 and the implementation time of the CCR model with all units was 59: 19.738. This feature exists for the entire algorithms that are configured for big data. This is in such a manner, the algorithms for big data, are designed such that, with bulk data they perform extremely well, (rendering good results) and does not function fittingly for small amounts of data. The algorithm's implementation time-period comprises of the implementation period relative to clustering in the IBM **SPSS Modeler** 18 Software and the period to carry out the other steps in GAMS software 23.4. (Table 16) System used to run Intel (R) Core (TM) i7 - Intel(R) Core (TM) i7-2670QM CPU @ 2.20GHz 2.20 GHz.

By executing the proposed algorithm with diverse samples and from the results collected in the Tables, it can be stated that, the validity of the proposed algorithm and a reduction in the time for computation is endorsed. Due to the fact that, the algorithm has been experimented on data of several sizes from100 to 9000, as a result this algorithm has the capacity to generalize big data.Similar and corresponding to the method proposed, Zhu et.al.,[13], Dellnitz.,[15] and Kherzimotlagh et.al.,[14], constructed a DEA model by segregating the efficient units from the inefficient ones; and thereby, reducing the computational complexities. Its diversity, with that, of the model proposed in this paper, is the method in which, the efficient units are differentiated from the inefficient units.

Zhu et.al .,[13] and Dellnitz.,[15], divide the n data by the k+1 set, next, in each set, the efficiency of the units is computed and the efficient units of each set are identified and are placed in an aggregation. Subsequently, k=K+1 is established and a re-classification is performed. The efficient units are recognized. Thereafter, the efficient units are manipulated, so as to configure a DEA model and the efficiency of all the units are gauged by the devised model. Calculating the efficiency of the entire units of each set and a continuing procedure in relevance with k=k+1 requires the resolving of a significant number of optimization problems. To segregate the efficient units from the inefficient ones, Khezrimotlagh et.al., [14], set the maximum or highest output value and the minimum input value, relevant to each one of the components as being a criterion, for detecting or specifying the efficient units and their separation. In the proposed method, to discriminate the efficient units from those that are inefficient, indexes are created with due attention to the DEA concept and the type of business (the bank in this circumstance); and with the help of a two-stage clustering, the efficient units are segregated from the inefficient units; and the necessity does not arise to solve an optimization-related issue. The implementation of clustering with the IBM ***SPSS Modeler*** 18 and or *Python* (Software), has a high speed for computation. To quote an example, it takes 00:00:04 seconds to classify 9,000 data into k=4

sets with the IBM SPSS Modeler 18 Software. It is evident that, when the dimensions of the problem are high, (*i.e.*, m+n+s is vast), in which case, the solving of optimization issues will be elevated in a complex manner. Since the proposed approach does not involve the resolving of any optimization-related problem in the first stage and specifies the entire efficient units, therefore, by employing this method, the computation process can be decremented notably.

5 Conclusion

Banking productivities and activities experience rapid modification. Thus, research in the field of the efficiency or performance of the banking industry is extremely crucial. When the amount of DMUs is exceedingly immense, because of computational complexities, the conventional DEA models may prove to be inadequate or even infeasible. Discussion in relative with, analyzing big or large-scale data is involved. Scrutinizing big data explores a large amount of information to reveal latent patterns, correlations and other perceptions, as well as evaluation and survey. There is no existing solo technology which encompasses big data analysis. Thereby, this paper renders an algorithm to reduce the computational complexities during the manipulation of big data. This is performed in such a manner that, by configuring the indexes, which have an impact on the efficiency score and clustering methods, the data is divided into two classifications, namely, efficient and inefficient. Finally, the DEA model is constructed with an efficient set of units and the efficiency of the entire units are calculated with the constructed model. In the proposed approach, there is no requirement to solve the problem of optimization, in order to segregate the efficient from the inefficient units; and the entire efficient units, are separated from the inefficient ones with the assistance of the clustering method; and by taking advantage, of the concept of dominance. The utilization of clustering methods causes a decrease in computational complexities and as a result, a reduction in the time-period concerning calculations for big data. For example, in connection with the proposed method, for a sample consisting of 9,000 units, instead of solving the issue of optimization, a time-period of 00:27:57.878 is necessary to compute the efficiency of all the units, due to the utilization of clustering methods. But if this calculation is to be conducted by the CCR model and the efficiency calculated, a period of 00:59:19.738 will be required. It is apparent that when the problem is a highly sizeable one, solving the optimization problems will include high complexity. However, the proposed method identifies the entire efficient units in the initial stage without resorting to any optimization problem. Thence, by utilizing this approach the process for calculation can be reduced drastically. In view of later undertakings, an algorithm designed for segregating the efficient units of the BCC model is proposed. As another recommendation, it seems to be beneficial to survey and design an algorithm, when the set of efficient units is also considered as big data itself.

References

[1] Sherman, H.D. and Gold, F., Bank Branch Operating Efficiency. Evaluation with Data Envelopment Analysis, *Journal of Banking and Finance*, 1985; 9, 297-315. Doi:10.1016/0378-4266(85)90025-1

[2] Charnes, A., Cooper, W. W., & Rhodes, E., Measuring the efficiency of decision making units, *European journal of operational research*, 1978; 2(6), 429-444. Doi: 10.1016/0377-2217(78)90138-8

[3] Banker, R. D., Charnes, A., & Cooper, W. W., Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management science*; 1984; 30(9), 1078-1092. Doi: 10.1287/mnsc.30.9.1078

[4] Cook, W. D., & Seiford, L. M., Data envelopment analysis (DEA)–Thirty years on. *European journal of operational research*, 2009; 192(1), 1-17. Doi: 10.1016/j.ejor.2008.01.032

[5] Dulá, J. H., & Helgason, R. V., A new procedure for identifying the frame of the convex hull of a finite collection of points in multidimensional space, *European Journal of Operational Research*, 1996; 92(2), 352-367. Doi:10.1016/0377-2217(94)00366-1

[6] Dulá, J. H., Helgason, R. V., & Venugopal, N., An algorithm for identifying the frame of a pointed finite conical hul, *Informas Journal on Computing*, 1998; 10(3): 323-330. Doi:10.1287/ijoc.10.3.323

[7] Barr, R. S., & Durchholz, M. L, Parallel and hierarchical decomposition approaches for solving large-scale data envelopment analysis models, *Annals of Operations Research*, 1997; 73, 339-372. Doi:10.1023/a:1018941531019

[8] Sueyoshi, T., & Chang, Y. L., Efficient algorithm for additive and multiplicative models in data envelopment analysis, *Operations Research Letters*, 1989; 8(4), 205-213. Doi:10.1016/0167-6377(89)90062-x.

[9] Dulá, J. H., & Thrall, R. MA., computational framework for accelerating DEA, *Journal of Productivity Analysis*, 2001; 16(1), 63-78. Doi:10.1023/a:1011103303616

[10] Chen, W. C., & Cho, W. JA., procedure for large-scale DEA computations, *Computers & Operations Research*, 2009; 36(6), 1813-1824. Doi:10.1016/j.cor.2008.05.006

[11] Dulá, J. H., & López, F. J., DEA with streaming data, *Omega*, 2013; 41(1), 41-47. Doi: 10.1016/j.omega.2011.07.010

[12] Chen, W. C., & Lai, S. Y., Determining radial efficiency with a large data set by solving small-size linear programs, *Annals of Operations Research*, 2017; 250(1), 147-166. Doi: 10.1007/s10479-015-1968-4.

[13] Zhu, Q., Wu, J., & Song, M., Efficiency evaluation based on data envelopment analysis in the big data context, *Computers & Operations Research*, 2018; 98, 291-300. Doi: 10.1016/j.cor.2017.06.017

[14] Khezrimotlagh, D., Zhu, J., Cook, W. D., & Toloo, M., Data envelopment analysis and big data, *European Journal of Operational Research*, 2019; 274(3), 1047-1054. Doi: 10.1016/j.ejor.2018.10.044.

[15] Dellnitz, A Big data efficiency analysis: Improved algorithms for data envelopment analysis involving large datasets, *Computers & Operations Research*, 2022; Volume 137,2022,105553. Doi: 10.1016/j.cor.2021.105553.

[16] Vijayarani, S., & Nithya, S. An efficient clustering algorithm for outlier detection, *International Journal of Computer Applications*, 2011; 32(7), 22-27.

[17] Vijayarani, S., & Nithya, S., Sensitive Outlier Protection in Privacy Preserving Data Mining, *International Journal of Computer Applications*, 2011; 33(3).