

# Clustering-Based Method for Data Envelopment Analysis

Hassan Najadat, Kendall E. Nygard, Doug Schesvold  
North Dakota State University  
Fargo, ND 58105

## Abstract.

*Data Envelopment Analysis (DEA) is a powerful performance measurement in economic sector and operations research to assess the relative efficiency for each decision making unit (DMU). In general, there are two assumptions in DEA. Firstly, the DEA assumes that all DMUs are homogenous in their environments and secondly, the DEA is a deterministic approach which refers to not allow to noise or errors in measurements. A large number of papers have addressed the DEA models but not many of them have focused on the heterogenous of DMUs and on the scalability over large datasets (i.e. when datasets contain a large number of DMUs). In this paper, we propose a new method for determining efficiently the performance scores of non-homogenous DMUs based on clustering methods to discover the outliers early. Experimental results presented show big improvements for our approach in assessing a funding transportation system for school districts in North Dakota State.*

**Keywords:** Data envelopment analysis, Data mining, constraint-based cluster, outlier discovery, decision making unit.

## 1. Introduction

The most important factor in economic activities is the productivity of each unit in an organization [SMK00]. Grossman discusses productivity improvement as representing one of the key competitive advantages of an enterprise [Gro93]. Managers use performance measurements to provide them with a strategic plan about organizations [LEB95]. By using the performance measurements, managers can adopt a long-term perspective, make communication more precise, and allocate the organization's resources to the most attractive improvements activities [SZ95].

This paper integrates two important fields of information technology: data mining and data envelopment analysis (DEA) to provide a new tool in measuring the performance of decision making units (DMU). The general motivation for this approach is to achieve synergy-producing results that could not be obtained if each model were operating individually.

In economic sector, the goal is analyzing the performance assessment of actions, productions, or organizational units [Kle04] in order to improve different types of efficiency. This efficiency is calculated as a ratio of set weighted outputs to set weighted inputs. The growth of acceptance of the DEA methodology in measuring the effectiveness of large entities is evidence of its applicability [Emr04, CSZ04].

In data mining field, the goal is to extract useful information from large databases [HK01]. This information can be used in various applications such as financial markets analysis [AIS93, Ben01] and business management [BL99]. The DEA yields a detailed analysis for DMUs to determine the efficient and inefficient units in order to gain useful information for making further improvements. This information can discover unknown relationships among the data which includes identifying the most productive operating scale sizes, the savings in recourses, and the most suitable ways to enhance inefficient units [Tha01]. Thus, both fields (i.e. the data mining and the DEA) serve the goals of management of an organization to get the best guide to improve the productivity of organization. The number of research papers published on various DEA applications [Emr04] and data mining applications [HK01] build a solid base in academic fields and business applications for both areas.

The DEA, which has the ability to measure the productivity of each DMU in the presence of multiple inputs and outputs [Sil86], is a mathematical model based on building a linear program for each DMU under evaluation. This linear program calculates the performance scores by constraining all DMUs to have an efficiency scores less than or equal to one.

The work we are proposing herein takes not only the initiative of developing a framework for measuring the performance of funding school's districts in transportation operations and introducing valuable results from the economical perspective but also providing a new method of integrating the constraint-based clustering into DEA that relieves from the burden of including the whole school districts in calculating the efficiency score for each district. We provide extensive comparison analysis to show the characteristics of our method and how it will be compared under standard DEA in terms of the quality of results. The performance of these school districts is measured several times using different economical models to get the most suitable view of the situation

The plan of development is as follows; first, the problem statement is provided. Second, an overview of different DEA models is shown which includes the CCR and BCC models. We then provide the proposed approach in detail. This is followed by a description of the main characteristics of the data used in this study. Result analysis is then provided. We conclude and present further research opportunities.

## 2. Problem Statement

There are many problems associated with applying the DEA in some applications. The first problem is that the DEA models assume that all DMUs are homogenous and identical in their operations [SS94]. Since various applications have heterogeneous DMUs and there is a high request to evaluate these applications under the DEA due to its acceptance as a performance measurement in different kind of business, we have to modify the DEA to work with these applications. If the heterogeneous DMUs are assessed by DEA without any modifications, the DEA yields a biased performance scores and inaccurate analyses. For example, in North Dakota State, the expenditure of all school districts exceeds more than 28 million annually, which requires to be evaluated in term of its transportation costs and services. An essential requirement in analyzing these districts is to build a fair funding formula for each district to manage and provide a solid plan that improves all inefficient districts and supports all efficient districts. This system can not be assessed under the standard DEA due to the non-homogenous of these districts in terms of their operations, density of population, variety of routes (i.e. rural rides are more expensive than city rides), and variety of district areas (i.e. many districts serve large areas with small number of pupils, which yields high cost in operations, while other districts serve small areas with high number of pupils). These factors will yield unfair funding system evaluation if we apply the standard DEA.

In this paper, a new algorithm is provided to perform the DEA computation in non-homogenous DMUs by introducing the clustering-based technique. In the proposed method, some of the non-homogenous DMUs are not related to the DMU under evaluation to such a degree that we would like to declare some of them to be outliers and fully exclude them from the analysis. This new method can be applied in different applications whose DMUs are not identical in their operations.

The second problem is that DEA assumes that each DMU participates in the performance measurements. But, it may be the case that some of the DMUs have very little effect on the efficiency or inefficiency of particular DMUs, yet they are present in the analysis and require large computational time. Since each DMU requires a separate linear program to evaluate its performance score, this linear program consists of a set of constraints that involve all DMUs in the problem domain, the new method greatly reduces the required computational time and hopefully not affects the solution very much.

The above discussions have evidenced that a new approach is needed to remedy the weakness of DEA. This paper provides an efficient new methodology that is considered to be a preprocessing phase for DEA, which expedites the solution of DEA computation dramatically and considers the efficiency score evaluations in a free outlier environment by eliminating those units that are dissimilar to the unit under evaluation.

## 3. Basic Concepts and Notations

### 3.1. Performance Measurements

There are two main approaches for performance measurements: (i) the parametric approach such as ratio analysis and regression analysis and (ii) the non-parametric approach such as data envelopment analysis (DEA) [Tha01]. The ratio analysis is referred to as a partial productivity measure to distinguish it from total factor measure [CST02]. Partial productivity performs various ratios for multiple outputs to single input or a single output to multiple inputs while the total factor measure takes account of all outputs and all inputs. A major problem with ratio analysis is that, when using multiple ratios, it is difficult to aggregate them into a single numeric judgment [Ier02].

The simple linear regression requires only a single output or an aggregation of outputs. Multiple regression suffers from many drawbacks, which include: (1) there are sets of residuals which can not be interpreted in a clear way in terms of efficiency, (2) the average performance is computed rather than the best performance using regression analysis, and (3) we have to hypothesis the type of model to be estimated. In [Bow98], there are many reasons to prefer the DEA rather than the regression analysis: (1) DEA does not require functional forms to associate inputs to outputs which give DEA more flexibility in recognizing differences in performances between DMUs, (2) DEA utilizes  $n$  optimizations (i.e. if the problem domain has  $n$  DMUs), one for each DMU, instead of a single optimization under the regressions analysis, (3) regression approaches are unable to identify the amount of inefficiency for each DMU while DEA provides both sources and amounts of any inefficiency, and (4) DEA provides the weights for each DMU, while the weights in regression analysis are applicable to a class of DMUs.

### 3.2. Data Envelopment Analysis

DEA is one of the most important performance measurements, which overcomes most of parametric approach shortcomings. The strong nonparametric flavor for DEA provides economic theory ability to estimate efficiency with minimal prior assumptions about the DMUs. DEA was first introduced in the literature in 1978 [CCR81]. It is a powerful performance measurement technique that can be used to assess the relative efficiency of comparable business by considering simultaneous the resources and environmental factors necessary to provide safe and reliable service [CCR+87, Sil86]. Each DMU has multiple inputs and outputs and there is no objective way to aggregate either inputs or outputs. The relative efficiency of a DMU is determined as the ratio of total weighted outputs to total weighted inputs.

DEA permits DMUs to choose the weights for each input and output so as to show the specific DMU in as positive a light as possible [AFS02]. There are two conditions which constraints the weights. First, all weights have to be greater than or equal to zero. Second, the efficiency for each DMU is less or equal to one [CCR81]. DMUs are homogenous such that they perform the same task but at different level of activities.

The data for performance measurements consists of  $n$  DMUs, each of which has  $m$ -dimensional vectors consisting of the set of inputs and outputs in varying values [Avk04]. These DMUs can be separate institutions or branches of a single program such as schools, banks, or hospitals [Emr04]. Each DMU has a set of inputs referred to as resources which transforms into a set of outputs referred to as services. All DMUs and its corresponding resource-service factors should be determined clearly. The performance measurements will yield biased results if we omit any important input or output.

### 3.3. The CCR Model

Among the eighty models listed in [Sil86], CCR was the first model which was established in 1978 by Charnes, Cooper, and Rhodes [CCR81]. The CCR model is an extension of the Farrell model, which was introduced in 1957. The efficiency in the Farrell model permitted considering a single input that yields in two separate outputs or two inputs used to produce a single output. This technique can not work with multiple inputs and multiple outputs simultaneously. This limitation is solved by allowing more than two inputs and outputs simultaneously in the CCR model.

The CCR model assumes comparing  $n$  DMUs. Each DMU consumes  $m$  inputs denoted by  $x_i, i=1, \dots, m$  and yields  $s$  outputs denoted by  $y_r, r=1, \dots, s$ . The relative efficiency score of the DMU <sub>$p$</sub> ,  $p=1, \dots, n$ , is measured by evaluating the ratio of the set weighted outputs to set weighted inputs of DMU <sub>$p$</sub>  relative to all the ratios of all DMUs [CCR81]. This ratio is generalized to a single virtual output and virtual input. The following FP1 is a fractional program to calculate the performance score of DMU <sub>$p$</sub>  which it can be transformed to a liner program LP1.

$$\begin{aligned}
 \text{(FP1)} \quad & \text{Maximize } h_p = (u_1 y_{1p} + \dots + u_s y_{sp}) / (v_1 x_{1p} + \dots + v_m x_{mp}) \\
 & \text{Subject to} \\
 & (u_1 y_{1j} + \dots + u_s y_{sj}) / (v_1 x_{1j} + \dots + v_m x_{mj}) \leq 1 \quad (j=1, \dots, n) \\
 & v_1, \dots, v_m \geq 0 \\
 & u_1, u_2, \dots, u_s \geq 0
 \end{aligned}$$

The fractional program is transformed into the following linear program [CCR81].

$$\begin{aligned}
 \text{(LP1)} \quad & \text{Maximize } h_p = u_1 y_{1p} + \dots + u_s y_{sp} \\
 & \text{Subject to} \\
 & v_1 x_{1p} + \dots + v_m x_{mp} = 1 \\
 & (u_1 y_{1j} + \dots + u_s y_{sj}) \leq (v_1 x_{1j} + \dots + v_m x_{mj}) \quad (j=1, \dots, n) \\
 & v_1, \dots, v_m \geq 0 \\
 & u_1, u_2, \dots, u_s \geq 0
 \end{aligned}$$

Cooper *et al.* provide a proof that the fractional program FP1 is equivalent to LP1 [CST02]. DEA requires a repeated application of linear programming to check each DMU [Avk04]. The result of LP<sub>1</sub> is combination weights of  $u$  and  $v$  under a constraint requiring that the same weights can not provide any unit with efficiency greater than one. The value of  $h_p$  represents the efficiency score for DMU <sub>$p$</sub>  relative to all other DMUs which ranges from zero to one. If the score is one, the unit is relatively efficient; otherwise, it is relatively inefficient. This model reflects the fact that the DEA efficiency is measured with reference to a production possible set boundary which envelops the input and output levels.

### 3.4. Example

To illustrate different concepts in DEA we adapted the following example from [CST02]. The example shows a simple DEA involving eight school districts. Each has a single input (expenditure) and a single output (number of rides) which are listed in Table 1 **Single Input and Single Output**.

Table 1 Single Input and Single Output

School District	A	B	C	D	E	F	G	H
Expenditure	6	9	9	12	15	18	18	24
# of rides	3	9	6	9	12	6	9	15
Efficiency Score	0.50	1.00	0.67	0.75	0.80	0.33	0.50	0.63

The last row of Table 1 Single Input and Single Output is the efficiency score for each school district. District B is the most efficient school district. Since these school districts use one input and one output, it is possible to know the best practice DMUs.

Figure 1. School districts depicts an efficient frontier line which represents all DMUs whose scores are one while production possibility set represents all DMUs whose scores are less than one. In the same figure, DMU<sub>B</sub> is considered to be such that we can not improve its input or output without worsening some other input or output. All DMUs {A, C, D, E, F, G, H} are inefficient units which represent the production possibility set in this problem.

For each inefficient unit A, C, D, E, F, G, and H, there is a subset of perfectly efficient DMUs called reference set. A reference set can be defined as the set of DMUs whose scores are one and very closed to the unit under evaluation. This enables a comparison to be made between inefficient unit and the units in its reference set.

This comparison is useful in getting information that will help in determining why an inefficient unit is performing poorly and how much projection should be made to reach the efficient frontier [CCR+87, Che04].

The linear program for school district A – district under evaluation here- can be represented using LP<sub>1</sub> as follows:

$$\begin{aligned}
 \text{Max} \quad & h_A = 3u \\
 \text{Subject to} \quad & 6v = 1 \\
 & 3u \leq 6v \quad (\text{A}) \\
 & 9u \leq 9v \quad (\text{B}) \\
 & 6u \leq 9v \quad (\text{C}) \\
 & 9u \leq 12v \quad (\text{D}) \\
 & 12u \leq 15v \quad (\text{E}) \\
 & 6u \leq 18v \quad (\text{F}) \\
 & 9u \leq 18v \quad (\text{G}) \\
 & 15u \leq 24v \quad (\text{H}) \\
 & u \geq 0, v \geq 0
 \end{aligned}$$

This linear program is only for district A, so we have to repeat it to all other districts by just replacing the first equation by suitable weights for the district under evaluation. The optimal solution is  $(v = u = h_A = 0.5)$ . District B is the only efficient unit which represents the reference set for district A. Therefore, the best possible weights for district A are  $u=0.5$  and  $v=0$ . To calculate the efficiency scores for other DMUs, we just replace the optimization equation with a suitable value for  $u$ . There are two types of project orientation for an inefficient unit toward the efficient frontier. The first is the input orientation, which is a maximizing of the movement toward the frontier through a proportional reduction of inputs. The other type is called output orientation, which is a

maximizing of the movement through a proportional augmentation of outputs [DL02]. If the districts are non-homogenous in this example, we would declare a peer group for district A that includes only the most similar districts to district A. This eliminates all outliers and then we build the linear system only for the peer group and district A which reduces the computational time to assess district A. we repeat this procedure for all other districts. This is discussed in the following section.

The main limitation of the CCR model is the constant return to scale which implies that a change in the amounts of the inputs leads to a similar change in the amounts of the outputs. DEA has been further extended in the BCC model in 1984 by Banker, Charnes and Cooper [CSZ04]. The BCC model allows a variable return to scale assumption, which is able to distinguish between technical and scale inefficiencies. Technical inefficiency is calculated by measuring how well the unit uses its inputs to create outputs. Scale inefficiency identifies whether increasing, decreasing, or constant returns to scale exist for further exploitation.

#### 4. Research Design

The DEA has been applied to many areas such as production of education [CCR81], performance evaluation in education [CCR81], efficiency of providing the educational services for New York State school districts [Rug98], Australian universities efficiency [Avk02], health care applications [DL02], public health

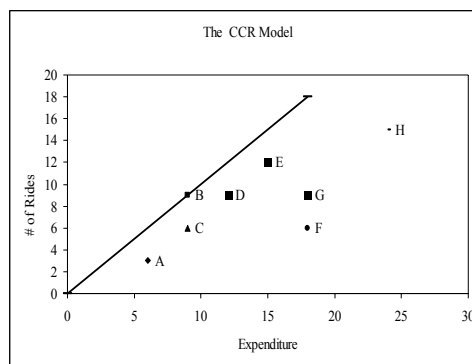


Figure 1. School districts performance score

service performance in the U.K. [PT92, TBD95], an assessment of the efficiency of New Jersey hospitals [BOR88], the relationship between hospital efficiency and type of hospital ownership [Bor88], nursing home care in The Netherlands [KOO94], financial institutions [Emr04], Spanish banking evaluation [DJ04, GL97], banking firms efficiencies [PPQ97], crime management [Sil86], and police offices evaluation [Sil86].

In this paper, all school districts in North Dakota are studied by employing two different methods: (i) standard DEA and (ii) clustering DEA. The goal of studying these districts is to provide a fair funding assessment for each district in term of transportation costs and services provided. The data set consists of 194 school districts, each of which consists of a single input and four outputs. The input is the total transportation expenditures which represents the summation of salary, benefits, purchased services, supplies, equipments, and others costs for each district. The outputs are the following: (1) annual total rides, (2) annual total rural rides, (3) surplus ride minutes (route maximum), and (4) surplus ride minutes (route average) [18]. These input-outputs represent the five-dimensional vector for each district. The reader is referred to [18] for more details on this issue.

Due to the heterogeneity of school districts in North Dakota, There is a set of attributes called site characteristics that are much related to all districts. It will help us in identifying the best similar districts for each district. These attributes are: (1) student density which represents the total number of annual rides divided by the square mileage of the district- rides per road miles, (2) usable road density which is calculated as the total mileage of usable roads within district boundaries divided by the square mileage of the district, (3) percent roads gravel, paved, and other, and (4) rides per road mile and square mileage. Based on the site characteristics, we define the similarity among the districts to form a representative cluster around the district under evaluation.

Having identified the input and outputs, the analysis options are addressed next. The management focuses on raising productivity without increasing the expenditures. This direction is called the output orientation which we applied in this study. We applied both analysis of constant returns to scale and variable returns to scale. Because of the space, we could not include all school districts data and its analysis. We provide a sample for some districts. All results can be viewed online at <http://www.cs.ndsu.nodak.edu/~najadat/District/all.html>.

#### 4.1. The Proposed Approach

Many previous studies such as [ST04, THP97, HL01, DJ05] have been proposed clustering method in various DEA applications but after evaluating the efficiency score for DMUs. Based on the results of DEA, they built clusters for each DMU and its reference set to show the degree of sensitivity in the presence of a particular DMU in the cluster. This direction is not suitable in analyzing non-homogenous DMUs due to the following: (1) "DEA will overestimate the efficiency scores of those operating under favorable conditions, and (2) DEA will underestimate the efficiency scores of those operating under unfavorable conditions." [Northcarolina].

Another popular approach that falls in non-homogenous DMUs as well is outlier discovery in DEA [outliers articles]. The outlier is redefined as those DMUs who have extreme efficiency scores putting in their consideration that these DMUs are relatively efficient because they lie in extreme positions and eventually bias the analysis. It does not work with non-homogenous DMUs due to the following: (1) the outliers are discovered after applying the DEA model which affects all analysis (i.e. all DMUs participate in the analysis), and (2) the original data may have outliers. In [Northcarolina], they integrated DEA into regression analysis. They calculate the efficiency scores first (unadjusted scores) then regress these scores with other variables other than the input-output variables, and then use the outcomes to recalculate the efficiency scores. The DEA is calculated twice and they did not check the presence of outliers in their data.

The approach proposed herein belongs to the non-homogenous DMUs. It uses the outliers' discovery process by introducing a peer group for each DMU. This peer group contains only the best similar DMUs to the DMU under evaluation. All non-related DMUs are not participated in the analysis of the DMU under evaluation. The peer group for the DMU under evaluation is then assessed by any DEA solver system. As we shall demonstrate latter, our approach is specifically targeted for very large datasets. To that end, we provide a detailed experimental analysis over school districts dataset.

#### 4.2. The Algorithm

In this section we discuss our algorithm which greatly resembles the outlier discovery in datamining area. This algorithm reduces the computational time of analyzing non-homogenous DMUs using DEA by defining a peer group for each DMU under evaluation. The algorithm is formally outlined in Figure 2: **The proposed algorithm**

#### 4.3. Process of Outlier Detection

The process of determining outliers in each district group starts by building the dissimilarity vector that represents the distance between the district under the evaluation and all other district. Suppose we have  $n$  school districts; we build  $n$  vectors so that each vector has  $[d(i,1), d(i,2), d(i,3), \dots, d(i,n)]$  where  $d(i,j)$  is a

non-negative number that is close to 0 when the districts  $i$  and  $j$  are near each other. There are many distance metrics used to measure the similarity between two entities in a database literature [4,14]. The most famous is Euclidean distance:

$d(j,i) = \text{sqrt} ( |f_{j1} - f_{i1}|^2 + |f_{j2} - f_{i2}|^2 + \dots + |f_{jk} - f_{ik}|^2 )$  where  $i = (f_{i1}, \dots, f_{ik})$  and  $j = (f_{j1}, \dots, f_{jk})$  are two  $k$ -dimensional data districts. In our case study, we use the site characteristics values as a set of features for each district.

Statisticians determine the outliers based on the data distribution. There are many different methods for normally distributed data based on robust regression methods: z-score method, modified z-score method, and boxplot method [30]. All of the experimental observations are standardized and the standardized values outside a predetermined bound are labeled as outliers [17]. The values of the distance vector  $d(i,j)$  are standardized by using the modified z-score method.

In step 1.4, we proposed mean absolute deviation which is more robust than using standard deviation since the standard deviation is affected by the outliers. This modified z-score is evaluated based on an outlier resistant estimator such that the values of z-scores do not become too small [26]. The deviations from the mean  $|v_j - m|$  are not squared; therefore, the effect of the outlier is reduced. According to Chebyshev theorem [28] which stated that "for any data set with mean  $x$  and the standard deviation  $s$  at least 75% of the values will fall within the interval  $x \pm 3*s$  and at least 89% of the values will fall within the interval  $x \pm 3*s$ ".

<p><b>Inputs:</b></p> <ol style="list-style-type: none"> <li>1. A database <math>D</math> contains all <math>n</math> DMUs (Each DMU is represented by a tuple with a set of attributes (i.e. features <math>(f_1, \dots, f_k)</math>, the site characteristics in our case study).</li> <li>2. Define the following parameters: <ol style="list-style-type: none"> <li>a. <b>min</b> is the minimum number of DMUs for each peer group.</li> <li>b. <b>max</b> is the maximum number of DMUs for each peer group.</li> </ol> </li> </ol> <p><b>Algorithm:</b></p> <ol style="list-style-type: none"> <li>1. For every DMU<math>_j</math> (<math>j=1, \dots, n</math>) <b>BEGIN</b> <ol style="list-style-type: none"> <li>1.1. Generate two vectors <math>V</math> and <math>ID</math>: <ol style="list-style-type: none"> <li>1.1.1. A distance vector <math>V=[v_1, v_2, \dots, v_n]</math> which refers to the similarity between DMU<math>_j</math> and all DMU, each <math>v</math> is calculated using Euclidian metric as follows: for <math>i=1</math> to <math>n</math> <math display="block">v(i) = \text{sqrt} (  f_{j1} - f_{i1} ^2 +  f_{j2} - f_{i2} ^2 + \dots +  f_{jk} - f_{ik} ^2 )</math></li> <li>1.1.2. Generate an identification number for each DMU and store them in a vector called <math>ID_j=[DMU_1, \dots, DMU_n]</math></li> </ol> </li> <li>1.2. Sort all values of <math>V</math> in ascending order. (All <math>ID</math>s of DMUs whose <math>v</math> is swapped have to be swapped also to keep the matching)</li> <li>1.3. Generate a vector <math>P_j</math> whose values are the first <math>max</math> values of distance vector <math>V</math>.</li> <li>1.4. Generate the z-score values as follows: <ol style="list-style-type: none"> <li>1.4.1. Calculate the mean <math>m</math> of <math>P_j</math>: <math display="block">m_j = (v_1 + v_2 + \dots + v_{max}) / max</math></li> <li>1.4.2. Calculate the modified standardized measurement <math>S_j</math> as follows: <math display="block">S_j = ( v_1 - m  + \dots +  v_{max} - m ) / max</math></li> <li>1.4.3. Calculate z-score for each value in <math>P_j</math> for <math>i=1</math> to <math>n</math> <math display="block">z_j = (v_j - m) / S_j</math></li> <li>1.4.4. for <math>i=1</math> to <math>max</math> if <math>(z_j \geq m + 3 S_j)</math> Do the following: if (number of DMUs in <math>P_j \geq min</math>) then delete <math>v(i)</math> from <math>P_j</math> and delete <math>ID_i</math> from <math>ID_j</math></li> </ol> </li> </ol> </li> </ol> <p><b>END</b></p> <p><b>Output:</b> A set of DMUs that are very similar to DMU<math>_j</math> stored in <math>ID_j</math></p>
--

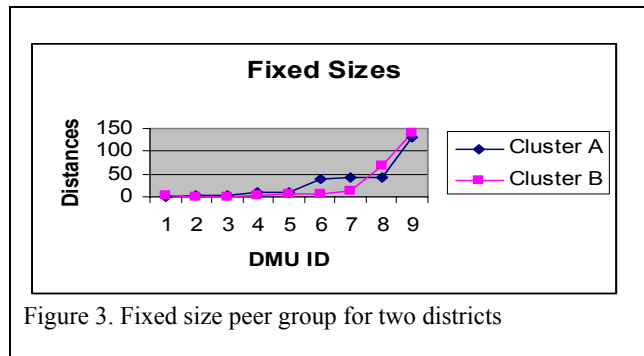
Figure 2: The proposed algorithm

#### 4.4. Clustering-Based DEA

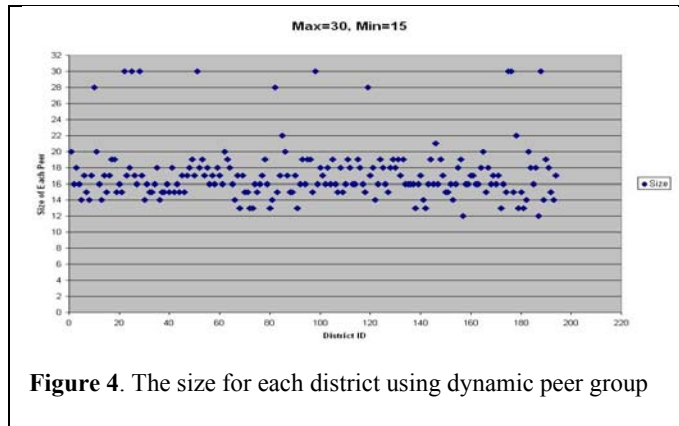
The main limitation and weakness of a standard formulation of DEA in a large scale problem is to build a separate linear program for each DMU. Solving this linear system is considered to be a computationally

intensive. Letting  $n$  be the number of DMUs in the database  $D$ , there are  $n$  separate linear systems each consisting of  $n$  constraints. The algorithm steps define the group of DMUs that are similar to the  $DMU_j$  under evaluation. This peer group represents a cluster around  $DMU_j$ . We have built this cluster based on many constraints such as *max*, *min*, and *threshold* of *z*-score to discover the outliers early. Consequently, the size of this cluster is variable from one DMU to another which can be considered as a factor in accelerating the DEA calculations. Also, this yields to get an accurate weights and efficiency scores for each DMU. Taking into consideration the predefined min, max, and threshold parameters in building the cluster around the  $DMU_j$ , the cluster will be dynamic size with homogenous members which solves the DEA problems.

One could claim that using a fixed number of members in each cluster for each DMU. This direction yields non-homogeneous members in each cluster (i.e. this results inclusion non-related DMUs in the cluster and produces unfair comparisons among the clusters). Figure 3. Fixed size peer group for two districts shows two peer groups A and B under fixed size assumption which yields non-homogenous distances between the members in each cluster. These irregular distances between each element cause inclusion outliers, which are dissimilar to the district under evaluation.



Oppositely, dynamic peer group assumed the variability in the number of elements in each cluster and the distances reached. Figure 4 shows the size of each peer group based on the clustering outlier detection. It shows 194 district, each district has a cluster contains 15 to 30 members. Most clusters have a uniform size that leads to a systematic inclusion of each member in each group and the similarities among the members of each district group.



It is simple to show the distances reached in each cluster under the dynamic sizes assumption. Figure 5 Area covered in each cluster shows 194 clusters –each district has a cluster- with the farthest distances reached by most dissimilar member in each cluster. It confirms that sometimes a peer group might contain 20 members while it reaches the same distance of another district with less than 15 members.

In fact, there were many modifications applied in DEA basics in ranking the units. One ranks a unit by excluding it from the evaluation of the whole system [4] and some others identify the influential units if they are frequently used in the calculation of efficiency scores [33]. But all of them did not consider the existence of outlier among the DMUs. Clustering the outlier far away from DEA calculation will lead to a successful

evaluation of efficiency score for each unit that eventually affects the whole system results based on accuracy and time consumption.

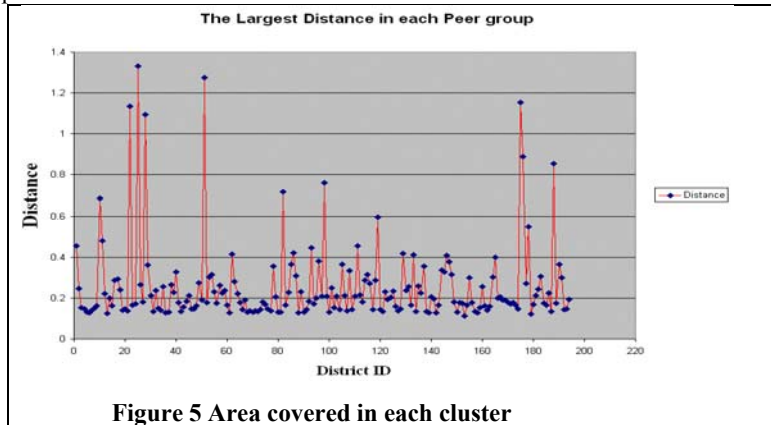


Figure 5 Area covered in each cluster

One could claim that fixing peer group size causes the simplicity of DEA calculations, but it is a simple to state that excluding the outliers in the DEA calculation is the best way to predict accurate efficiencies. If the outlier is not studied, DEA calculations can mislead decision makers by introducing a misleading reduction of resources with invalid weights. Using imprecise weights will not only influences the efficiency score of the unit under evaluation but will also change the efficiency scores of other DMUs since the unit under evaluation may be included in another peer group to evaluate other units.

## 5. Results and Analysis

As far as models are concerned, in the present study, the output-oriented model has been preferred to the input-oriented model for management purposes. In constructing the most similar group of school districts around the district under evaluation, we consider the site characteristics as a set of features in forming a suitable cluster as discussed in section 4.4.

Figure 6 Architecture of the proposed Alg.orithm.

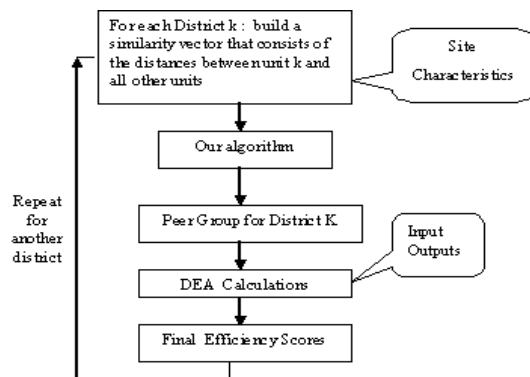


Figure 6 Architecture of the proposed Alg.

In the following sections, we provide a detailed analysis of the various models. We executed the VRS model which is a piecewise linear frontier line providing more efficient units. Also, CRS was executed on the same data set to distinguish the DMU performance in both standard DEA and cluster DEA. We discuss the results of the CRS and the VRS in order to illustrate the improvements in score efficiency measurements.

### 5.1. Constant Return to Scale –CCR model

As mentioned in the CCR review, this study applies both standard and cluster DEA under the CRS assumption. Cluster-based DEA was able to discover thirty efficient school districts compared to the standard DEA which discovered only three efficient districts among 194 districts. As shown in Figure 7. **Efficiency Scores Distribution under CRS** we divide the school performance scores into 11 intervals. The reader should observe that the scores of school districts are mainly concentrated in lower intervals from 0 to 0.4 using standard DEA which consequently implies poor performance in these intervals. Moreover upper intervals have

sparse districts, which do not exceed 10% of the data. On the other hand, using our algorithm, the high scores are concentrated in the intervals from 0.5 to 1 and scattered in the lower intervals. This indicates the power of this algorithm in producing more efficient units compared to the standard DEA.

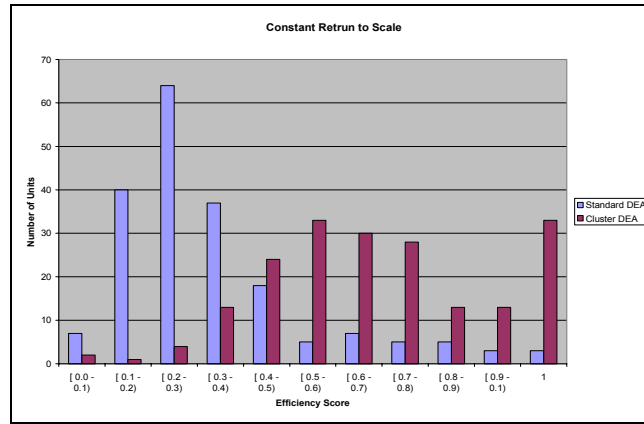


Figure 7. Efficiency Scores Distribution under CRS in CCR

### 5.2. Variable Return to Scale – BCC model

The second computational model is the model that handles VRS in both standard and cluster-based DEA. Figure 8 **VRS efficiency scores under standard and cluster DEA** depicts the differences in the quality of the improvements and the number of efficient units discovered. For example, 30% of districts in the highest interval are classified to be efficient districts using the cluster DEA while only 9% using the standard DEA. It is also simple to recognize that in the lower intervals, the number of districts under cluster DEA is less than the number of districts under standard DEA. This potential improvement is discussed in the following section.

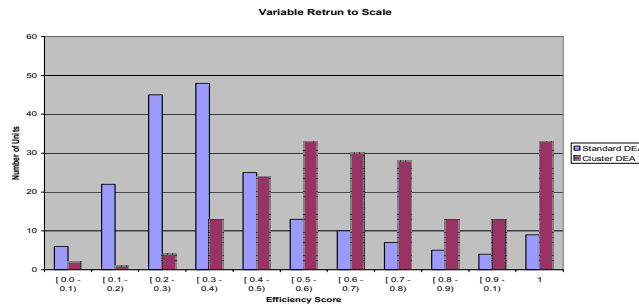


Figure 8 VRS efficiency scores under standard and cluster DEA

### 5.3. Potential Improvements Analysis

The summary improvement analysis provides the best quality of service after applying clustering DEA. This observation is capable of increasing the quality of granted services in each district. Table 2 shows that the efficient school districts are increased in both constant and VRS under clustering DEA. The results show that performance scores of 100% efficiency is attained by 92% and 87% of the districts using VRS and CRS respectively. In general, average performance scores are improved by 67%. This percentage includes all score intervals from zero to 100. This improvement covers all factors in the funding transportation system. This helps the leaders of the instruction sector in North Dakota to rearrange their priorities for improving inefficient schools.

Table 2. DEA Summeries for all Districts

	# of Efficient districts	Averages of all efficient & inefficient districts performance score
CRS-Standard DEA	3	0.335
VRS-Standard DEA	9	0.407
CRS-Cluster DEA	33	0.677
VRS-Cluster DEA	58	0.779

As shown in Table 3 Efficiency Scores Improvements, using the cluster-based DEA yields an improvement in each school district. 97% of the school districts were improved in CRS and 95% in VRS. The performance scores itself also were improved by 42% in CRS and 72% in VRS. On average, 96% of the school district scores have shown an increased performance which is mainly due to the reallocation of districts in such a way that they are only compared with peer districts using the dynamic group under an outlier free environment.

Table 3 Efficiency Scores Improvements

	CRS	VRS
Average improvement for all district scores	42%	72 %
Percentage of the number school district improved	97 %	95 %

Because of the lack of space herein, we provide a detailed analysis for each district in order to discover the weaknesses and strengths of the transportation funding system and make it simple to let the leaders of school districts to see it at <http://www.cs.ndsu.nodak.edu/~najadat/District/all.html>. We summarize the potential improvements in all services produced from the funding system in Table 4. This analysis compares the projection values that have shown significant improvements.

The first service for each district is the annual total rides, which has improved to 81.35% and 63.02% in CRS and VRS, respectively. The second service is the annual total rural rides, which has improved to 95.46% and 65.50% in CRS and VRS respectively. The largest improvement was in the surplus ride minutes (route maximum) and the surplus ride minutes (route average) with 147.72% and 126.38% in CRS and VRS, respectively.

Table 4. Potential improvements averages without additional costs

	Annual total rides		Annual total rural rides		Surplus ride minutes (route maximum)		Surplus ride minutes (route average)	
	CRS	VRS	CRS	VRS	CRS	VRS	CRS	VRS
Actual	0.10	0.103	0.13	0.129	0.06	0.063	0.07	0.073
Projection	0.17	0.127	0.23	0.166	0.11	0.090	0.13	0.097
Difference	0.06	0.024	0.10	0.037	0.05	0.027	0.05	0.024
Improvement %	81.35%	63.02%	95.46%	65.50%	147.72%	105.57%	26.38%	93.26%

We will end this section by emphasizing that we implement the proposed algorithm using visual C++, and the analysis of school districts was done using DEA-Solver-LV [CST02] and DEAP software which can be downloaded at <http://www.une.edu.au/econometrics/cepawp.htm>.

## 6. Conclusion

In this work we have shown to efficiently produce a peer group for each DMU under evaluation in order to solve the non-homogenous DMUs and discover the outliers early. This study have shown that using the proposed algorithm before applying the DEA in large scale problems then the performance of DEA will be improved dramatically in computational time and quality of the efficiency scores. We provide a framework of integrating constraint-based clustering into DEA and efficiently determine the best similar DMUs to the DMU under evaluation by fully excluding the outliers from analysis. We applied clustering DEA in the transportation funding system for the school districts in North Dakota. The results show the fairness of evaluating each district. We plan to integrate our work with different methods of clustering techniques such as the hierarchical methods and outlier techniques in data mining area to provide a new tool of building the peer groups for DMUs.