DEA with Missing Data: An Interval Data Assignment Approach

Reza Kazemi Matin\textsuperscript{a,},* Roza Azizi\textsuperscript{b}

\textsuperscript{a} Associate Professor, Department of Mathematics, Islamic Azad University, Karaj Branch, Karaj, Iran
\textsuperscript{b} MSc, Department of Mathematics, Islamic Azad University, Karaj Branch, Karaj, Iran

Received 17 October, 2013; Revised 20 December, 2013; Accepted 02 February, 2014

Abstract

In the classical data envelopment analysis (DEA) models, inputs and outputs are assumed as known variables, and these models cannot deal with unknown amounts of variables directly. In recent years, there are few researches on handling missing data. This paper suggests a new interval based approach to apply missing data, which is the modified version of Kousmanen (2009) approach. First, the proposed approach suggests using an acceptable range for missing inputs and outputs, which is determined by the decision maker (DM). Then, applying the least favourable bounds of missing data along with using the proposed range is suggested in estimating the production frontier. A data set is used to illustrate the approach.

Keywords: Data envelopment analysis, Missing inputs, Missing outputs, Range.

1. Introduction

Data envelopment analysis (DEA) is an approach, which is used for measuring the relative efficiency of a set of decision-making units (DMUs), which convert the same inputs to the same outputs (Charnes et al., 1978). DEA provides efficiency scores and efficient projections for inefficient DMUs. One of the most important qualifications of DEA is that inputs and outputs are known exactly. However, in the real-world DEA applications, there are many cases, in which there are no complete input/output quantities, and data contain some missing values. As a result, the classical DEA models are weak at efficiency evaluation of these kinds of systems. In recent years, there have been a few studies in the DEA literature dealing with missing data. Kao and Liu (2000) proposed an extended DEA model based on fuzzy theory to handle missing data. They suggested replacing the missing values with a fuzzy number and using the observed data to estimate membership functions of fuzzy efficiency scores. Smirlis et al., (2006) suggested an approach to estimating the amount of missing values. Kousmanen (2009) in a systematic fashion proposed a method to handle missing data, in which the missed outputs are replaced with zero and missed inputs with a sufficiently large number. This approach will be discussed in the next section in more details, and our new interval data version of this work will be introduced. Azizi (2013) showed some drawbacks of Smirlis (2006) and replaced it by a new approach. Zha et al. (2013) introduced modified DEA models to calculate proper amounts of missing values.

In this paper, after reviewing some approaches to missing data, a connection among some technologies, which can handle missing data, is presented and finally, by using decision makers (DMs) knowledge about the production system, an acceptable range for missing data is assumed, and missing inputs and outputs will be replaced by the worst case in the interval, respectively. The suggested approach can eliminate some of the disadvantages of other approaches.

The rest of the paper is organized as follows. Section 2 is devoted to introducing the required models and notations. In section 3, three approaches in DEA in dealing with missing data will be discussed. The first technique eliminates DMUs with missing values. The second is based on eliminating outputs and inputs with missing values, and the last one is based on taking zero for missing outputs and taking high value for missing inputs. Afterward, we will present our approach based on having a range for missing data in section 4. Some features of this new approach are explored in this section. Section 5 is devoted to presenting some numerical examples. Conclusions are given in section 6.

2. Models and Assumptions

There are some models and notations, which are required in our discussion about missing data.
2.1 Notations

In general, the production possibility set (PPS) in a specific technology could be stated as
\[ T = \{(x, y) \in \mathbb{R}^{m+n} | x \text{ can produce } y\} \]

To present the required background, assume that there are \( n \) DMUs to be evaluated. Each DMU consumes varying amounts of \( m \) different inputs to produce \( s \) different outputs. Specifically, DMU \( j \) consumes amount \( x_{ij} \) of input \( i \) and produces amount \( y_{jr} \) of output \( r \). In the classical DEA models, non-negative data is assumed and further it is assumed that each DMU has at least one positive input and one positive output value. Under constant returns to scale (CRS) technology and complete data, the underlying production set could be represented based on data set as
\[
\{(x, y) \in \mathbb{R}^{m+n} | x \geq \sum_{j=1}^{n} \lambda_j x_j, y \leq \sum_{j=1}^{n} \lambda_j y_j, \forall j; \lambda_j \geq 0\}
\]

Furthermore, TDMUo be the production possibility set, obtained by eliminating DMU O from the observe set (DMU O is a unit with missing input and/or output), and TXY be the production possibility set, obtained by eliminating output \( r' \) and input \( i' \) for all units. These technologies can be introduced as follows:
\[
\begin{align*}
T_{DMUo} &= \{(x, y) | x \geq \sum_{j=1}^{n} \lambda_j x_j, y \leq \sum_{j=1}^{n} \lambda_j y_j, \lambda_j \geq 0, \forall j; \lambda_o = 0\} \\
T_{XY} &= \{(x, y) | x = \hat{x}_o \land y = \hat{y}_o, \forall i \neq i', (\hat{x}_o, \hat{y}_o) \in T_{DEA}\}
\end{align*}
\]

2.2 CCR model

The following table represents envelopment and multiplier forms in both input and output orients of the CCR DEA model introduced by Charnes, Cooper, and Rhodes (1978) when the unit “o” is under evaluation to get its Farrell (1957) radial efficiency score.

<table>
<thead>
<tr>
<th>Input oriented</th>
<th>Envelopment model</th>
<th>Min ( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.t. ( \sum_{i=1}^{n} x_i \lambda_i \leq \theta x_m, \ i = 1,...,m, ) \quad (f.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sum_{r=1}^{s} y_r \lambda_r \geq y_m ), \quad r = 1,...,s, \lambda_r \geq 0, \quad j = 1,...,n.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiplier model</td>
<td>Max ( \theta = \sum_{i=1}^{n} u_i y_m )</td>
<td></td>
</tr>
<tr>
<td>s.t. ( \sum_{r=1}^{s} u_r y_m - \sum_{j=1}^{n} v_j x_j \leq 0, \quad j = 1,...,n, \ u_i \geq 0, \ v_i \geq 0, \forall r, \forall i. )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output oriented</th>
<th>Envelopment model</th>
<th>Max ( \phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.t. ( \sum_{i=1}^{n} x_i \lambda_i \leq x_m, \ i = 1,...,m, ) \quad (f.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sum_{r=1}^{s} y_r \lambda_r \geq \phi y_m ), \quad r = 1,...,s, \lambda_r \geq 0, \quad j = 1,...,n.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiplier model</td>
<td>Min ( \phi = \sum_{i=1}^{n} v_i x_{io} )</td>
<td></td>
</tr>
<tr>
<td>s.t. ( \sum_{j=1}^{n} u_j y_m = 1, ) \quad (f.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sum_{j=1}^{n} v_j x_j - \sum_{i=1}^{n} u_i y_i \leq 0, \quad j = 1,...,n, \ u_i \geq 0, \ v_i \geq 0, \forall r, \forall i. )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Measuring the efficiency score of DMUs with missing data

Several approaches are proposed for dealing with missing data. Here, a classified summary of these approaches is presented.

3.1 Eliminating DMUs with missing values

One of the common strategies in DEA to handle missing data is eliminating DMUs including missing data; see, for example, Neal, Ozcan and Yanqiang (2002) among others. Although this approach is used in some DEA applications, it is not a good solution for a small sample size or cases with a large number of DMUs with missing values. When a considerable number of DMUs is eliminated from the sample, it may influence the efficiency scores of the remaining DMUs badly. In other words, the efficiency score of the remaining DMUs almost increases; this approach causes some inefficient DMUs to become efficient, and bias results are obtained. And finally, in this approach performance of those DMUs including missing data cannot be evaluated.
3.2 Eliminating outputs and inputs with missing values

Another way to handle missing data is eliminating outputs or inputs variables, including missing data in evaluation. This technique is a common method in statistical science, especially in using SPSS software for data analysis. Again, in the presence of a considerable number of eliminated input/output variables, this approach also yields bias efficiency scores for all units. In this case, the efficiency score of DMUs almost decreases and this approach causes some efficient DMUs to become inefficient. The other problem of this approach, which occurs in rare cases, will happen if each input and output of all DMUs includes at least one missing value; then all the inputs and outputs need to be discarded from the sample, and one may not be able to evaluate the relative performance of the units.

3.3 Assignment value approach

Among the other novel approaches to handling missing data in DEA literature, Kuosmanen (2009) technique is notable, in which missing values are suggested to be replaced by some pre-specified values. Under constant returns to scale (CRS) technology, this approach proposes to replace missing outputs by zero, and missing inputs by a large positive number and then using the classical CCR DEA model for the new data setting. In the next section, this approach will be discussed in more details and by following its main idea, a new improved technique to handling missing data under DEA framework is proposed.

4. Interval Data Instead of Missing Data

As mentioned earlier, Kuosmanen (2009) suggests the use of dummy entries and replacing missing outputs by zero and missing inputs by sufficiently large positive numbers. This is the first systematic approach to dealing with missing data in the literature. Here, by acknowledging his work, we suggest some points to improve its domain of applicability in the real-world applications. The following discussion shows that Kuosmanen (2009) approach may show some DEA model infeasible. Taking zero for missing outputs and taking large number for missing inputs is not a proper choice for all production systems. As an instance, let assume all the outputs of DMUO are missed, so by using Kuosmanen method, we need to set \( \forall r(r = 1, \ldots, s); y_{i o} = 0 \). Now, in the output-oriented envelopment form of the CCR model, there will be a redundant constraint associated with output variables as \( \sum_i 0 \lambda_i \geq \varphi_0 \). This yields to an unbounded objective function value and it is equal to infeasibility of the multiplier form of the CCR model; both are unacceptable. This simple and of course rare case shows that Kuosmanen (2009) approach in handling missing data needs some modifications in order to increase its applicability.

In general, even with preserving feasibility, such assessed dummy values may lead to misleading efficiency scores due to sensitivity problems in DEA as frontier technique based on extreme points. For a set of efficient units with missed inputs/outputs, accepting these values changes the efficient frontier and leads to bias efficiency estimation for all the units even for those with complete data.

To modify Kuosmanen approach and increase its applicability, we suggest using an acceptable range as an interval value for missing data. In real-world applications, there are some estimation techniques to achieve such intervals; for example, by using decision makers (DMs) information on missed values.

To introduce the modified approach, let’s assume \( x_{i';j'} \) and \( y_{r';j'} \) are missed in data set for some \( i', r', j' \) and \( j \). Based on the decision maker’s experience, the following acceptable intervals for these missing data can be presented: \( a \leq x_{i';j} \leq b \) and \( c \leq y_{r';j} \leq d \), where the constant bounds \( a, b, c, \) and \( d \) are real and positive numbers and depending the application could be selected using statistical or experimental techniques.

Now, in constructing the production set based on estimated interval data, the most pessimistic values for the missing items within their intervals are used in our approach, i.e. \( x_{i';j} = b \) \( y_{r';j} = c \) then the classical CCR DEA models for efficiency evaluation of the new data set will be used.

To show the results of this setting formally, let \( \text{TRD} \) denote the production possibility set determined by “range data” that uses the higher bound of the range for missing inputs and the lower bound of the range for missing outputs. We introduce \( \text{TRD} \) as follows:

\[
T_{RD} = \{ (x, y) | \begin{cases}
x_{ij} = b \ (\forall i \neq i', x_{i'o} = \bar{x}_{i'o}) \\
0 \leq a \leq x_{i'o} \leq b \\
y_{r'o} = c \ (\forall r \neq r', y_{r'o} = \bar{y}_{r'o}) \\
0 \leq c \leq y_{r'o} \leq d \\
y_{oj} = \bar{y}_{oj} \ (\forall j \neq o') \\
x_{oj} = \bar{x}_{oj} \ (\forall j \neq o) \\
(\bar{x}_{oj}, \bar{y}_{oj}) \in T_{DEA}
\end{cases}\}
\]

(3)

\( T_{RD}, T_{DMUo}, T_{DEA} \) and \( T_{XY} \) are related according to the following theorem.

Theorem 1: Production possibility sets \( T_{RD}, T_{DEA}, T_{DMUo}, \) and \( T_{XY} \) are nested as:

\[
T_{DMUo} \subseteq T_{RD} \subseteq T_{DEA}
\]

(4)

Proof: Under constant returns to scale assumption, \( T_{RD}, T_{DEA}, T_{DMUo}, \) and \( T_{XY} \) could be defined through the following formulas:
Let the \( r \)th output of DMU\(_{0} \) is missing.

1) First we show that \( T_{RD} \subseteq T_{DEA} \): Due to the formulation of technologies in section 2.1, the only difference between sets \( T_{RD} \) and \( T_{DEA} \) is in \( r \)th output of DMU\(_{0} \). For \( T_{DEA} \) this constraint reads as \( y_{r,\omega} \leq \sum_{j=1}^{m} \lambda_{j} y_{r,j} \). Since output \( r' \) is missing for DMU\(_{0} \) in \( T_{RD} \), the constraint reads as
\[
y_{r,\omega} \leq \sum_{j=1}^{m} \lambda_{j} y_{r,j} + \lambda_{0} \times c + \sum_{j=m+1}^{n} \lambda_{j} y_{r,j}
\]
(5)
Thus, the amount of acceptable \( y_{r,\omega} \) in \( T_{RD} \) must always be less than or equal to the corresponding value of \( y_{r,\omega} \) in \( T_{DEA} \). Since the two sets are otherwise identical, it is proved that \( T_{RD} \subseteq T_{DEA} \).

2) In the second step, it will be shown that \( T_{XY} \subseteq T_{RD} \). The constraint for output \( r' \) in \( T_{XY} \) reads as
\[
y_{r,\omega} \leq \sum_{j=1}^{m} \lambda_{j} y_{r,j} + \lambda_{0} \times c + \sum_{j=m+1}^{n} \lambda_{j} y_{r,j}
\]
and the constraint for output \( r' \) in \( T_{RD} \) reads as
\[
y_{r,\omega} \leq \sum_{j=1}^{m} \lambda_{j} y_{r,j} + \lambda_{0} \times c + \sum_{j=m+1}^{n} \lambda_{j} y_{r,j}
\]
It is clear that
\[
0 \leq \sum_{j=m+1}^{n} \lambda_{j} y_{r,j} \leq \sum_{j=1}^{m} \lambda_{j} y_{r,j}
\]
So, the acceptable amounts of output \( r' \) in set \( T_{RD} \) are greater than or equal to the zero value implied by \( T_{XY} \). Since the two sets are otherwise identical, the two sets position are related as \( T_{XY} \subseteq T_{RD} \).

3) And finally let show \( T_{DMU} \subseteq T_{RD} \) : Comparing \( T_{DMU} \) and \( T_{RD} \), easily can be observed that
\[
y_{r,\omega} \leq \sum_{j=1}^{m} \lambda_{j} y_{r,j} + \sum_{j=m+1}^{n} \lambda_{j} y_{r,j}
\]
(6)
Now, \( y_{\omega} = \sum_{j=1}^{n} \lambda_{j} y_{\omega,j} + \sum_{j=m+1}^{n} \lambda_{j} y_{\omega,j} \) for output \( r (r \neq r') \), in \( T_{DMU} \) but \( y_{\omega} = \sum_{j=1}^{n} \lambda_{j} y_{\omega,j} (r \neq r') \) in \( T_{RD} \) and it is clear that:
\[
\sum_{j=1}^{m} \lambda_{j} y_{\omega,j} + \sum_{j=m+1}^{n} \lambda_{j} y_{\omega,j} \leq \sum_{j=1}^{m} \lambda_{j} y_{\omega,j} (r \neq r')
\]
Thus, the amount of acceptable \( y_{\omega,\omega} \) in \( T_{DMU} \) must always be less than or equal to the corresponding value of \( y_{r,\omega} \) in \( T_{RD} \) for any output. So, it is proved that \( T_{DMU} \subseteq T_{RD} \).

The similar proof can be presented for missing inputs or for the case that missing inputs and outputs are existed.

5. Numerical Examples

In this part of the paper, through three numerical examples, a comparison on the results of the above-mentioned approaches in dealing with missing data will be made, and it will be shown that our proposed approach is more applicable than the other ones. The same data set as example 1 was analyzed in Kuosmanen (2009) paper. Data set of examples 2 and 3 are similar to the data set of some examples in research of Kuosmanen (2009).

For simplicity of presentation, let \( \theta \) be used to show the efficiency score of units when all the data are available; \( \theta_{\omega} \) for the efficiency score of units when the DMUs with missing values are omitted from the observation; \( \theta_{\omega} \) shows the efficiency score of units when the missing outputs or missing inputs are omitted from the data set; \( \theta_{\omega} \) also shows the efficiency score of DMUs obtained by Kuosmanen (2009) approach and finally \( \theta \) denotes the efficiency score of DMUs, which are obtained by our interval approach. The data, which are in parentheses, are supposed be the exact values of the missing inputs or outputs.

Example 1. Table 1 shows data for 5 hypothetical DMUs with one input \( X \) and two outputs \( Y \). The three last columns show an acceptable range of the missing items.

<table>
<thead>
<tr>
<th>DMU</th>
<th>( X )</th>
<th>( Y_{1} )</th>
<th>( Y_{2} )</th>
<th>Missing ( X )</th>
<th>Missing ( Y_{1} )</th>
<th>Missing ( Y_{2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>15</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>(20)</td>
<td>60</td>
<td>-</td>
<td>[15,30]</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>35</td>
<td>40</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>(45)</td>
<td>30</td>
<td>-</td>
<td>[30,50]</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>50</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2 reports efficiency score of the units computed by the five approaches by using output orientation of the CCR model.

<table>
<thead>
<tr>
<th>DMU</th>
<th>( \theta_{I} )</th>
<th>( \theta_{II} )</th>
<th>( \theta_{III} )</th>
<th>( \theta_{IV} )</th>
<th>( \theta_{V} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.75</td>
<td>1</td>
<td>0.75</td>
<td>0.89</td>
<td>0.8</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0.98</td>
<td>1</td>
<td>0.67</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>-</td>
<td>0.5</td>
<td>0.5</td>
<td>0.82</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>0.17</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Based on the results, when all data are known, the first approach shows DMUs B, D, and E are efficient. The
results in the second column show the strategy of omitting units, including missing data from the observation is beneficial for remaining units, for example, DMU A, which is not efficient in the original data set, becomes efficient. In contrast with the first approach, the results in the third column show when the output variable concluding missing value (Y2) is omitted, some of the efficient units, like D and E become inefficient in the new data set. It is caused by neglecting their good output performance in comparison with the other units. The fourth column shows efficiency scores computed by the Kuosmanen (2009) approach. In comparison with the second column, this approach suggests efficiency scores for all the units, and the results are more accurate in comparison with the previous approaches. The results of the new interval approach are reported in the last column of Table 3. The results show that the new approach gives a more accurate efficiency score in comparison with the original data results, as well as the Kuosmanen (2009) approach. However, there is a notable difference; DMU D, which is an efficient unit regarding the complete data set, gets a better rank in the interval approach compared to its score in Kuosmanen (2009) approach.

The differences could be even more in real data set when more accurate approximations are used in the form of intervals, in which the unknown missing values are likely to belong, and the computed results of the modified approach will be more reliable.

Example 2. Table 3 shows data for five hypothetical DMUs with two inputs X, and two outputs Y. The four last columns are used to show a selected range of missing inputs and outputs.

<table>
<thead>
<tr>
<th>DMU</th>
<th>X1</th>
<th>X2</th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>15</td>
<td>45</td>
<td>30</td>
</tr>
<tr>
<td>B</td>
<td>35</td>
<td>20</td>
<td>60</td>
<td>20</td>
</tr>
<tr>
<td>C</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>45</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>E</td>
<td>25</td>
<td>50</td>
<td>10</td>
<td>45</td>
</tr>
</tbody>
</table>

DMU Missing X1 Missing X2 Missing Y1 Missing Y2
A - - - -
B [25,40] - - -
C - - [30,45] -
D [30,50] - - -
E - - [5,15] -

The computed efficiency scores for this data setting in the five models are summarized in the table 4.

<table>
<thead>
<tr>
<th>DMU</th>
<th>$\theta_I$</th>
<th>$\theta_{II}$</th>
<th>$\theta_{III}$</th>
<th>$\theta_{IV}$</th>
<th>$\theta_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0.59</td>
<td>-</td>
<td>0.21</td>
<td>0.31</td>
<td>0.44</td>
</tr>
<tr>
<td>D</td>
<td>0.79</td>
<td>-</td>
<td>0.56</td>
<td>0.56</td>
<td>0.65</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>-</td>
<td>0.45</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

As it can be seen, scores in the last column computed using the new approach, gives the best approximation of the original efficiency scores among the other techniques.

Example 3. As the last illustration example, consider the table 5, which shows data for five hypothetical DMUs with one input X, and one output Y. The two last columns of the table shows suggested ranges for missing data.

<table>
<thead>
<tr>
<th>DMU</th>
<th>X1</th>
<th>Y1</th>
<th>Missing X1</th>
<th>Missing Y1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>45</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>35</td>
<td>60</td>
<td>[25,40]</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>30</td>
<td>40</td>
<td>[30,45]</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>30</td>
<td>[30,50]</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>25</td>
<td>10</td>
<td>[5,15]</td>
<td>-</td>
</tr>
</tbody>
</table>

The computed efficiency scores using five approaches are included in Table 6.

<table>
<thead>
<tr>
<th>DMU</th>
<th>$\theta_I$</th>
<th>$\theta_{II}$</th>
<th>$\theta_{III}$</th>
<th>$\theta_{IV}$</th>
<th>$\theta_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0.76</td>
<td>-</td>
<td>0.27</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.59</td>
<td>-</td>
<td>Infeasible</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.33</td>
<td>-</td>
<td>0.13</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0.18</td>
<td>-</td>
<td>Infeasible</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>

As the results show, two approaches are failed in computing efficiency scores for the units, including missing data. In addition, the method proposed by Kuosmanen (2009) cannot suggest an efficiency score for units C and D, and the new approach is the only technique, which finds feasible efficiency scores near to the original values.

6. Conclusion

One of the basic assumptions of the original DEA is that inputs and outputs are known on a ratio scale for all units. However, in many DEA applications, it is practically impossible to collect complete data sets, and missing values in some input/output variables are inevitable. In DEA literature, there have been few approaches, which are introduced for dealing with missing data. Through some simple numerical examples, it has been shown that the previous works like eliminating DMUs with missing values or eliminating missing inputs or
outputs may destroy the true classification of the units into efficient and inefficient ones. Also, these approaches may fail in measuring the efficiency scores of some units. One attempt in using missing data was done by Kuosmanen (2009), which may make infeasible solution for some DMUs. The current paper follows and extends the way proposed by Kuosmanen (2009) by assigning appropriate values as an acceptable interval for the missing data applying decision maker’s information. Then the worst bounds of the suggestion interval for variables with missing data are used to estimate the performance of DMUs. With some numerical examples, it was shown that the new approach is more applicable than the others in the existing literature.

To address the many practical problems encountered in settings such as these, we suggest that more extensive research be done on evaluating economical efficiencies such as cost efficiency or profit efficiency for units with missing prices.

References


