Estimating DEA Efficiency Using Uniform Distribution

[1]Md. Kamrul Hossain, [1]Anton Abdulbasah Kamil, [2]Adli Mustafa and [3]Md. Azizul Baten

[1]School of Distance Education, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia.
[2]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia
[3]Department of Decision Science, School of Quantitative Sciences, Universiti Utara Malaysia, 01610 UUM Sintok, Kedah, Malaysia
kamrulstat@gmail.com; anton@usm.my; adli@cs.usm.my and baten_math@yahoo.com.

Abstract

The most commonly used non-parametric tool for measuring the relative efficiency of Decision Making Units (DMU) is Data Envelopment Analysis (DEA). In this article, a method for measuring the efficiency level of a DMU when it is in an unfavourable situation as well as estimating the efficiency using uniform distribution is shown. The efficiency score from the traditional BCC-DEA model and the efficiency score in an unfavourable situation form an interval. This interval, known as *interval efficiency*, is used to estimate efficiency using uniform distribution. In an empirical example, a 95% confidence interval (CI) is calculated for the efficiency score using a three-point estimation method. The analysis indicated that the efficiencies that were estimated from uniform distribution are all within the confidence interval. In addition, a statistical test shows that there is no significant different between the estimated efficiency and the efficiency from DEA.

Keywords: *Interval efficiency; Data Envelopment Analysis; Uniform distribution; Three-point estimation.*

AMS Subject Classification: 62xx.

1 Introduction

Data envelopment analysis (DEA) has found amazing development in theory and methodology and widespread applications in the world. Traditional CCR and BCC DEA models do not deal with noise data and are based on exact inputs and outputs. In practical situation, the issues of missing data and imprecise data often arise. Generally speaking, uncertain information can be expressed in interval numbers. Therefore, how to evaluate the performance of a DMU in its unfavourable situation through existing DEA models with uniform distribution is a worth-studying problem. This is the need of both the developments of DEA theory and methodology and its real applications.

In 1957, Farrell [13] proposed some new ideas on relative efficiency. Based on these ideas, Charnes et al. [7] introduced the popular tool, Data Envelopment Analysis (DEA), for measuring the efficiency of

Decision Making Units (DMU). The CCR-DEA model [7] and the BCC-DEA model [5] are the basic models in DEA. These models are deterministic and they require that all inputs and outputs data are known with certainty. In practical situation, however, it is often impossible to obtain data with such consideration. The issues of missing data and imprecise data often arise. Fortunately, the DEA models have been extended to address these issues (see for example [2, 29]). Recently, the stochastic variations of the DEA models have received significant attention. Banker [4], for example, developed an approach of DEA with maximum likelihood methods to make inference in the presence of noise in data. Chance constrained programming [6, 8] was adopted in DEA by Land et al. [17, 18, 19]. Olesen and Petersen [21] modified the chance constrained DEA model to be used with stochastic multiple inputs and outputs. Fuzzy DEA [14], Bootstrap DEA [26, 27, 28], Imprecise DEA [9, 10, 11, 29] and Robust DEA [22, 23] are also developed to work with noise in data. In the presence of noise in data, efficiency is measured either from the optimistic viewpoint or pessimistic viewpoint [1]. Using the optimistic viewpoint and the pessimistic viewpoint, Entani et al. [12] introduced a method to measure interval efficiency. In the paper, influence of the environmental factors on the less efficient DMU is considered to measure the minimum efficiency of a DMU and to carry out interval efficiency.

Simar and Vanhems [24] proposed the first method for statistical inference on efficiency using Free Disposal Hull (FDH) model. Later on, Simar et al. [25] added an assumption of convex production function to the Simar and Vanhems [24] model. They developed the method using consistent bootstrap procedures. However, the DEA Bootstrap is a nonparametric sampling method and is not suitable for small samples. The objective of this paper is to use the interval efficiency to infer the efficiency score of DEA. The proposed method is developed based on parametric sampling techniques (statistical distributions) and three-point-estimation method to overcome the limitation of Simar et al. [25].

The rest of the paper is organized as follows. In Section 2, the proposed method of estimating efficiency using uniform distribution is described. An empirical example and the usage of the three point estimation method to calculate the 95% confidence interval of the efficiency score are given in Section 3. Finally, some concluding remark is given in Section 4.

2 Methodology

The performance of a DMU in its unfavourable situation can be evaluated in two stages. In the first stage, the efficiency of a DMU is measured using the existing DEA model and in the second stage, a modified DEA model is used to measure the minimum efficiency.

Stage 1: DEA Model

DEA models are either input oriented or output oriented. We have chosen the output oriented DEA model of Banker et al. [5]. The linear programming (LP) expression of the model is as follows:

$$\text{a)} \underset{\theta,\lambda}{Max} \quad \theta_i^a$$

$$\text{Subject to, } x_i \geq \sum_{i=1}^{n} \lambda_i x_i \qquad (2.1)$$

$$\text{And } \theta_i^a y_i \leq \sum_{i=1}^{n} \lambda_i y_i$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

where $\text{input}\,(x_i) \geq 0, \text{output}(y_i) \geq 0, \text{weight}\,(\lambda_i) \geq 0$ and θ_i^a is a function of y_i. This LP problem is solved *n* times, which is equivalent to the number of DMU.

Stage 2: Modified DEA Model

Generally, production models can be classified as either neoclassical model or frontier model [15]. This classification depends on the interpretation of the deviation terms, $\varepsilon_i \in \Re$. The assumption of neoclassical model is that, all firms are efficient and the deviations $\varepsilon_i \in \Re$ are seen as random, uncorrelated noise terms that satisfy the Gauss-Markov assumptions. But, in the frontier models, all deviations from the frontier are attributed to inefficiency, which implies that, $\varepsilon_i \leq 0, \forall i = 1,2,...,n$ [16]. A function

$$y_i = g(.) + \varepsilon_i \qquad (2.2)$$

is considered a frontier model if $\varepsilon_i \in \Re$ are interpreted as composite error terms that include both the inefficiency and the noise components, where $g(.)$ is the production frontier [3]. However, when the output is due to the environmental factor and the inefficiency ε_i is $\varepsilon_i = g(.) - y_i \geq 0$, it indicate that the frontier output is always greater than or equal to the observed output. Now to maximize the output

ε_i, we use the output oriented DEA method. The method can be expressed as the following the linear programming problem:

$$\text{b)}\ \underset{\lambda}{Max}\ \theta_i^b$$

$$\begin{aligned} \text{Subject to, } & x_i \geq \textstyle\sum_{i=1}^{n} \lambda_i x_i \\ \text{And } & \theta_i^b \varepsilon_i \leq \textstyle\sum_{i=1}^{n} \lambda_i \varepsilon_i \\ & \textstyle\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \qquad (2.3)$$

where θ_i^b is a function of the environmental factor and inefficiency (ε_i). Let us suppose that the maximized quantity of ε_i is $\underset{\max}{\varepsilon_i}$. Now, if we calculate the efficiency of the *i*-th firm using the worst output of its $g(.) - \underset{\max}{\varepsilon_i}$ value, and observed the outputs of other firm, we will have the minimum efficiency level of the firm. This is under the consideration that the inputs are the same and all other factors that have effect on the output are constant. This can be mathematically shown as:

$$\text{c)}\ \underset{\theta,\lambda}{Max}\ \theta_i^c$$

$$\begin{aligned} \text{Subject to, } & x_i \geq \textstyle\sum_{i=1}^{n} \lambda_i x_i \\ \text{And } & \theta_i^c (g(.) - \underset{\max}{\varepsilon_i}) \leq \lambda_i (g(.) - \underset{\max}{\varepsilon_i}) + \textstyle\sum_{j=1, j\neq i}^{n} \lambda_j y_j \\ & \textstyle\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \qquad (2.4)$$

where θ_i^c is the minimum efficiency level of the *i*-th firm.

Estimation of DEA Efficiency using Uniform Distribution:

When an event is equally likely to happen, uniform distribution is frequently used to generate random number. Simar and Wilson [26] proposed a bootstrap procedure to estimate DEA efficiency. They used uniform distribution to generate sample. In this research, the equally likely to happen property and the usage of the bootstrap procedure warrants the use of uniform distribution to estimate DEA efficiency.

Uniform distribution is a continuous distribution ranging from *a* to *b*. The probability density function of this distribution on the interval [a, b] is

$$f(x) = \begin{cases} 0 & for \quad x < a, \quad or \quad x > b \\ \dfrac{1}{b-a} & for \quad a \le x \le b \end{cases} \tag{2.5}$$

and the cumulative distribution function is

$$F(x) = \begin{cases} 0 & for \quad x < a, \quad or \quad x > b \\ \dfrac{x-a}{b-a} & for \quad a \le x \le b \end{cases} \tag{2.6}$$

We use the uniform distribution to estimate efficiency by assuming that the minimum efficiency (θ_i^c) and highest efficiency (*one*) are the lower limit and the upper limit, respectively. The probability density function of the uniform distribution for efficiency is:

$$f(\theta) = \begin{cases} 0 & for \quad \theta > 1, \quad or \quad \theta < \theta_i^c \\ \dfrac{1}{1-\theta_i^c} & for \quad \theta_i^c \le x \le 1 \end{cases} \tag{2.7}$$

Hypothesis:
To determine whether the estimated efficiency represents the efficiency from DEA or not, we use the following hypothesis.

H_{A1}: There is a significant difference between efficiency from DEA and the estimated efficiency using uniform distribution.

A summary of the values that were calculated while performing Stage 1 and Stage 2 of the proposed method is depicted in Table 1. The case study is on the efficiency analysis of rice production in 20 districts. DMU 1, DMU 2, DMU 4, DMU 12, DMU 13 and DMU 20 are efficient DMUs since their efficiency level are 1. DMU 16 has the lowest efficiency score. The *Gap between Observed and Frontier Output* is the amount of output that is lost by a DMU because of inefficiency. Inefficiency is influenced by some unobserved factors. If a DMU can control these factors then it would become an efficient DMU. The *Minimum Output* of a DMU can be calculated as:

$$Minimum \quad Output = (Frontier \quad Output - Maximum \quad Loss \quad of \quad Output) \tag{3.1}$$

The minimum output and the observed output are found to be the same for DMU 7 and DMU 16 since from the beginning these DMUs are in their worst position. When a DMU is faced with the same influence of the environmental as what is being faced by DMU 7 and/or DMU 16, its efficiency is in the worst situation. *Maximum Loss of Output* is the gap between the frontier output and the output in the worst situation. The maximum loss of output of DMU 1 can be explained as: if all the environmental factors could influence DMU 1 as they did on DMU 7 and/or DMU 16, then there would be a shortage of 273 units from the frontier output.

The efficiency of a DMU in its worst situation is calculated by considering the minimum output and the observed input of the DMU together with the observed output and input of the other DMUs. For example, to evaluate the efficiency of DMU 1 in its worst situation, the minimum output (2202 units) and the observed inputs of DMU 1 together with the observed inputs-output of DMU 2 to DMU 20 are used. The efficiency score from the BCC model of DMU16 (0.823) is the lowest. In the worst situation, DMUs 18 and 19 would also obtain the same lowest efficiency score.

Table 1: Efficiency level and projected output using observed inputs and output

DMU	Observed Output	Frontier Output	*Gap between Observed and Frontier Output	Efficiency from BCC model	Maximum Loss of Output	Minimum Output	Efficiency in Worst situation
1	2475	2475	1	1	273	2202.00	0.968
2	2261	2261	1	1	271.40	1989.60	0.928
3	2179	2235.66	56.66	0.975	293.84	1941.82	0.869
4	2075	2075	1	1	288.56	1786.44	0.905
5	1850.2	1925.35	75.15	0.961	267.71	1657.65	0.861
6	1790.7	1993.6	202.9	0.898	276.31	1717.29	0.861
7	1676	1994.42	318.42	0.840	318.42	1676.00	0.840
8	1870	1903.32	33.32	0.982	251.11	1652.21	0.868
9	1874.6	1999.12	124.52	0.938	285.71	1713.41	0.857
10	1616.9	1892.05	275.15	0.855	299.99	1592.06	0.841
11	1734	1939.21	205.21	0.894	249.59	1689.62	0.871
12	1916	1916	1	1	167.91	1748.09	0.983
13	1808	1808	1	1	156.56	1651.44	0.924
14	1850.7	1917.41	66.71	0.965	229.17	1688.24	0.880
15	1831.8	1921.06	89.26	0.954	259.28	1661.78	0.865
16	1500	1821.97	321.97	0.823	321.97	1500.00	0.823
17	1745	1839.02	94.02	0.949	318.91	1520.11	0.827
18	1512.33	1610.94	98.62	0.939	284.68	1326.26	0.823
19	1506.85	1633.42	126.57	0.923	288.65	1344.77	0.823
20	1894.56	1894.56	1	1	298	1596.56	0.948

*We use one instead of zero in Gap between Observed and Frontier Output which do not influence the result.

The Three-Point Estimate was developed as part of the Program Evaluation and Review Technique (PERT) which is used in project management. The method estimates the expected case (E_i) based on three estimates [20]:

1) The Most Likely case (O_i)
2) The Optimistic case (H_i)
3) The Pessimistic case (W_i)

Then, the expected case (E_i) from the $O_i, H_i \; and \quad W_i$ is calculated as follows:

$$E_i = \frac{H_i + 4O_i + W_i}{6} \tag{3.2}$$

The 95% confidence interval for the estimated efficiency can be expressed as:

$$CI^* = E_i \pm 2 * SD_i \tag{3.3}$$

where SD_i is the standard deviation.

In this study, the following assumptions are made to estimate the expected case. (1) The efficiency score from the BCC model is considered as the most likely case (O_i), (2) the value of the optimistic case (H_i) is one, which is the highest efficiency score that can be obtained and (3) the efficiency score of the worst situation is considered as the pessimistic case (W_i).

Definition 1 *Efficiency is the ratio of weighted outputs and weighted inputs. Inputs and outputs values are non-negative i.e.* $input(x_i) \geq 0, \quad output(y_i) \geq 0 \quad and \quad weight(\lambda_i) \geq 0$. *Thus, the efficiency score of all DMUs are greater than or equal to zero and the expression of the efficiency score is* $\theta \geq 0$.

Definition 2 *Efficiency is measured with the constrain that it cannot exceed the value 1, that is* $\theta \leq 1$. *If* $\theta = 1$, *then the DMU is efficient and if* $\theta < 1$, *then the DMU is inefficient.*

Based on Definition 1 and Definition 2, the confidence interval of efficiency score is defined as:

$$CI = \begin{cases} 1 & if \quad CI^* > 1 \\ CI^* & if \quad 0 \leq CI^* \geq 1 \\ 0 & if \quad CI^* < 0 \end{cases} \tag{3.4}$$

The expected efficiency and confidence interval that were calculated using Equations 3.2, 3.3 and 3.4 are presented in Table 2. The estimated efficiency is the mean of 1000 random values that were generated from uniform distribution in the interval of $[W_i, H_i]$. 12 DMUs have upper limit of less than 1. All the estimated efficiencies from the uniform distribution lie in the 95% confidence interval.

Table 2: Inferential statistics of efficiency scores

DMU	Efficiency From BCC Model (O_i)	Efficiency in Worst Situation (W_i)	Estimated efficiency from uniform (U_i)	Biasness between Observed and Uniform Estimation	Expected efficiency (E_i)	Standard deviation (SD_i)	95% CI	
							Lower limit	Upper Limit
1	1	0.968	0.984	0.016	0.995	0.005	0.985	1
2	1	0.928	0.963	0.037	0.988	0.012	0.964	1
3	0.975	0.869	0.934	0.041	0.962	0.022	0.918	1
4	1	0.905	0.952	0.048	0.984	0.016	0.952	1
5	0.961	0.861	0.93	0.031	0.951	0.023	0.905	0.997
6	0.898	0.861	0.931	-0.033	0.909	0.023	0.863	0.955
7	0.84	0.84	0.919	-0.079	0.867	0.027	0.813	0.921
8	0.982	0.868	0.936	0.046	0.966	0.022	0.922	1
9	0.938	0.857	0.928	0.01	0.935	0.023	0.889	0.981
10	0.855	0.841	0.92	-0.065	0.877	0.027	0.823	0.931
11	0.894	0.871	0.933	-0.039	0.908	0.022	0.864	0.952
12	1	0.983	0.991	0.009	0.997	0.003	0.991	1
13	1	0.924	0.962	0.038	0.987	0.013	0.961	1
14	0.965	0.88	0.94	0.025	0.957	0.02	0.917	0.997
15	0.954	0.865	0.931	0.023	0.947	0.023	0.901	0.993
16	0.823	0.823	0.912	-0.089	0.853	0.03	0.793	0.913
17	0.949	0.827	0.914	0.035	0.937	0.029	0.879	0.995
18	0.939	0.823	0.911	0.028	0.93	0.03	0.87	0.99
19	0.923	0.823	0.912	0.011	0.919	0.03	0.859	0.979
20	1	0.948	0.974	0.026	0.991	0.009	0.973	1

Based on the result in Table 3, hypothesis H_{A1} is rejected, meaning that there is no different between the estimated efficiency and the efficiency from DEA.

Table 3: Comparison between efficiency from DEA and the estimated Efficiency

Mean Difference	Standard. Deviation	95% CI		Calculated t-score	Degree of freedom	p-value
		Lower	Upper			
0.00595	0.04265	-0.01401	0.02591	0.624	19	0.540

In Figure 1, all the observed and the estimated efficiencies lie in the 95% CI of the uniform distribution. The upper limit values of 12 DMUs are less than one. If the confidence interval is considered as interval efficiency, then DMU 16 is the dominated DMU.

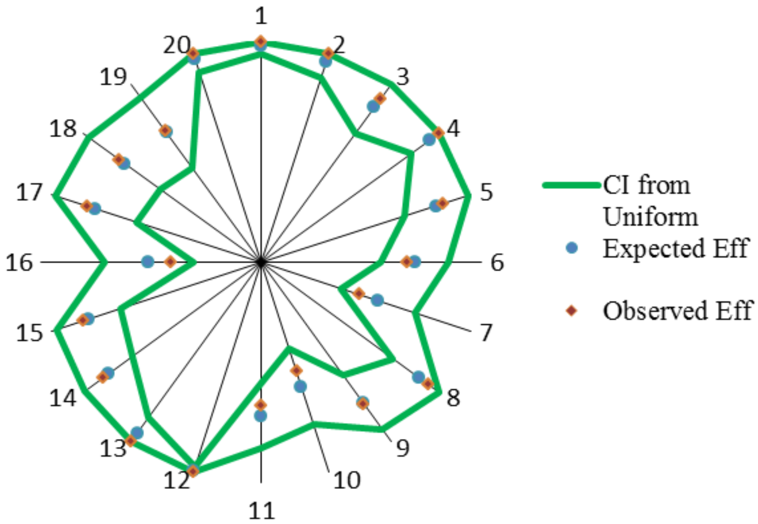

Figure 1. Confidence interval of DEA efficiency

4 Conclusion

In this article, an approach to estimate DEA efficiency using uniform distribution is presented. In an empirical example, the estimated efficiency score and the efficiency score from the BCC model is compared and no significant different between them were found. Confidence interval is estimated with three-point estimate method. All of the estimated efficiencies from uniform distribution are in the interval. This approach will help managers to determine the optimal size of inventory for their DMUs and researchers to estimate the statistical properties of efficiency score from DEA.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

[1] H. Azizi and R. Jahed, Improved data envelopment analysis models for evaluating interval efficiencies of decision-making units, *Computers & Industrial Engineering* **61** (2011), 897–901.

[2] H. Azizi, A note on data envelopment analysis with missing values: an interval DEA approach, *The International Journal of Advanced Manufacturing Technology* **66** (2013), 1817–1823.

[3] D. J. Aigner and S. F. Chu, On estimating the industry production function, *The American Economic Review* **58** (1968), 826–839.

[4] R. D. Banker, Maximum likelihood, consistency and data envelopment analysis: a statistical foundation, *Management Science* **39** (1993), 1265–1273.

[5] R. D. Banker, A. Charnes and W. W. Cooper, Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science* **30** (1984), 1078–1092.

[6] A. Charnes and W. W. Cooper, Chance-constrained programming, *Management Science* **6** (1959), 73–79.

[7] A. Charnes, W. W. Cooper, E. Rhodes, Measuring the efficiency of decision making units, *European Journal of Operational Research* **2** (1978), 429–444.

[8] A. Charnes, W. W. Cooper and G. H. Symonds, Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil, *Management Science* **4** (1958), 235–263.

[9] W. W. Cooper, K. S. Park and G. Yu, IDEA and ARIDEA: Models for dealing with imprecise data in DEA, *Management Science* 45 (1999), 597–607.

[10] W. W. Cooper, K. S. Park and G. Yu, IDEA (imprecise data envelopment analysis) with CMDs (column maximum decision making units), *Journal of the Operational Research Society* **52** (2001a), 176–181.

[11] W. W. Cooper, K. S. Park and G. Yu, An illustrative application of IDEA (Imprecise Data Envelopment Analysis) to a Korean mobile telecommunication company, *Operations Research* **49** (2001b), 807–820.

[12] T. Entani, Y. Maeda and H. Tanaka, Dual models of interval DEA and its extension to interval data, *European Journal of Operational Research* **136** (2002), 32–45.

[13] M. J. Farrell, The measurement of productive efficiency, *Journal of the Royal Statistical Society, Series A (General)* **120** (1957), 253–289.

[14] P. Guo and H. Tanaka, Fuzzy DEA: a perceptual evaluation method, *Fuzzy Sets and Systems* **119** (2001), 149–160.

[15] T. Kuosmanen and M. Fosgerau, Neoclassical versus frontier production models? Testing for the skewness of regression residuals, *The Scandinavian Journal of Economics* **111** (2009), 351–367.

[16] T. Kuosmanen and A. L. Johnson, Data envelopment analysis as nonparametric least-squares regression, *Operations Research* **58** (2010), 149–160.

[17] K. C. Land, C. K. Lovell and S. Thore, Productivity and efficiency under capitalism and state socialism: the chance-constrained programming approach, In Proceedings of *The 47th Congress of the International Institute of Public Finance*, St. Petersburg, (edited by P. Pestieau), 1992, pp. 109–121.

[18] K. C. Land, C. K. Lovell and S. Thore, Chance-constrained data envelopment analysis, *Managerial and Decision Economics* **14** (1993), 541–554.

[19] K. C. Land, C. K. Lovell and S. Thore, Productive efficiency under capitalism and state socialism: an empirical inquiry using chance-constrained data envelopment analysis, *Technological Forecasting and Social Change* **46** (1994), 139–152.

[20] S. McConnell, *Software Estimation: Demystifying the Black Art*, Microsoft press, Redmond, WA, 2009.

[21] O. B. Olesen and N. Petersen, Chance constrained efficiency evaluation, *Management Science* **41** (1995), 442–457.

[22] S. J. Sadjadi and H. Omrani, Data envelopment analysis with uncertain data: an application for Iranian electricity distribution companies, *Energy Policy* **36** (2008), 4247–4254.

[23] S. J. Sadjadi and H. Omrani, A bootstrapped robust data envelopment analysis model for efficiency estimating of telecommunication companies in Iran, *Telecommunications Policy* **34** (2010), 221–232.

[24] L. Simar and A. Vanhems, Probabilistic characterization of directional distances and their robust versions, *Journal of Econometrics* **166** (2012), 342–354.

[25] L. Simar, A. Vanhems and P. W. Wilson, Statistical inference for DEA estimators of directional distances, *European Journal of Operational Research* **220** (2012), 853–864.

[26] L. Simar and P. Wilson, Sensitivity analysis of efficiency scores: how to bootstrap in non parametric frontier models, *Management Science* **44** (1998), 49–61.

[27] L. Simar and P. Wilson, A general methodology for bootstrapping non parametric frontier models, *Journal of Applied Statistics* **27** (2000a), 779–802.

[28] L. Simar and P. Wilson, Statistical inference in nonparametric frontier models: the state of the art, *Journal of Productivity Analysis* **13** (2000b), 49–78.

[29] J. Zhu, Imprecise data envelopment analysis (IDEA): a review and improvement with an application, *European Journal of Operational Research* **144** (2003), 513–529.