How to Make an Interval Version of Data Envelopment Analysis (DEA) More Adequate

Vladik Kreinovich, Allen J. Cutcher, and Minh-Tho Nguyen

Abstract Data Envelopment Analysis (DEA) describes how relatively efficient a production facility can be under the most beneficial combination of prices for its resources and for what it produces. To estimate this efficiency, the original DEA uses the exact amounts of resources used and items produced – e.g., the overall amounts during some period of time. However, the resulting efficiency value does not take into account that in real life, in general, a facility's efficiency fluctuates. To capture these fluctuations, it is desirable to produce not a single efficiency value, but rather an interval of possible values of efficiency. Such interval-valued version of DEA has indeed been proposed and used. In this paper, we argue that while in some cases, the current interval DEA adequately describes these fluctuations, in some other reasonable situations, this method overestimates the fluctuations range. To overcome this limitation, we propose a more economically adequate interval version of DEA.

1 Introduction

What is a practical problem? Many companies have manufacturing facilities all around the world. In some places, labor is cheaper, in some places, electricity is cheaper, etc. Some facilities are run more efficiently, some are just starting and may

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

Allen J. Cutcher
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: ajcutcher@miners.utep.edu

Minh-Tho Nguyen
Faculty of Finance and Banking, Ho Chi Minh City Open University, 35–37 Ho Hao Hon Street
Cau Ong Lanh Ward, Ho Chi Minh City 70000, Vietnam, emailtho.nm@ou.edu.vn

need some help. It is therefore desirable to be able to locate facilities that need help in improving their efficiency. Because of the differences in costs, it is often not easy to directly compare efficiency at different locations. To make such a comparison possible, researchers came up with a technique known as Data Envelopment Analysis (DEA); see, e.g., [1].

Need to take fluctuations into account. The original DEA formulas describe each facility j by a single efficiency value E_j. This value provides an overall description of the facility's efficiency. However, it does not take into account that in real life, in general, efficiency often fluctuates – in some days the production is more efficient, in some days it is less efficient. Instead of a single efficiency value, it is desirable to come up with bounds $\underline{E}_j$ and $\overline{E}_j$ describing the interval $\left[\underline{E}_j, \overline{E}_j\right]$ of possible efficiency values.

What is known and what are the remaining problems. Such an interval version of DEA has indeed been proposed and used [9] (see also [3, 7, 11]). However, as we will show, this version does not always adequately describe the efficiency range: in some realistic situation it overestimates this range. It is therefore desirable to come up with a more economically adequate interval version of DEA.

What we do in this paper. In this paper, we explain the limitations of the current interval DEA, and we describe a new, more economically adequate interval version of DEA.

The structure of this paper. In Section 2, we describe the formulas and algorithms of the original DEA approach. (For readers' convenience, the explanation of why the corresponding algorithms are correct is placed in the special – last – Section 6 of the paper.) In Section 3, we describe the current interval version of DEA. In Section 4, on a simple example, we explain that this version is not always economically adequate. Finally, Section 5 contains the analysis of the problem and the description of the proposed new interval version of DEA.

2 What is DEA and how it is computed: a brief reminder

What is DEA. We consider the situation in which J facilities produce (some or all of) M different products based on N different resources: labor, electricity, components, etc. Let us fix a period of time – e.g., day, week, month, quarter, or year. To describe the efficiency of each facility, we need to introduce some notations.

- Let $y_{j,m}$ denote the amount of the m-th item produced by the j-th facility during the selected time period.
- Let $x_{j,n}$ denote the amount of the n-th resource used by the j-th facility during this time period.
- Let $u_m \geq 0$ denote the price-per-unit of the m-th item.
- Let $v_n \geq 0$ denote the cost-per-unit of the n-th resource.

Under these notations, during the selected time period, the j-th facility spent the amount

$$D_j \stackrel{\text{def}}{=} \sum_{n=1}^{N} v_n \cdot x_{j,n}, \tag{1}$$

and produced the items whose overall price is

$$N_j \stackrel{\text{def}}{=} \sum_{m=1}^{M} u_m \cdot y_{j,m} \tag{2}.$$

Thus, for each dollar spent, the facility generates the amount

$$\theta_j \stackrel{\text{def}}{=} \frac{N_j}{D_j} = \frac{\sum\limits_{m=1}^{M} u_m \cdot y_{j,m}}{\sum\limits_{n=1}^{N} v_n \cdot x_{j,n}}. \tag{3}$$

This amount depends on the prices u_m and costs v_n. For some combinations of these prices and costs, one facility may be more efficient, for other combinations of prices and costs, other facilities may be more efficient.

We say that a facility is (relatively) perfectly efficient if for some combination of the prices and costs, its productivity θ_j is the largest, i.e.,

$$\theta_j = \max_k \theta_k. \tag{4}$$

If for all possible combinations of prices and costs, the facility's efficiency lacks behind others, i.e., if always

$$\theta_j < \max_k \theta_k, \tag{5}$$

this means that this facility is not perfectly efficient. As a numerical value of the efficiency of the j-th facility, it makes sense to consider the largest value E_j of the ratio between its productivity and the largest of the facilities' productivities:

$$E_j \stackrel{\text{def}}{=} \max_{u_1 \geq 0, \ldots, u_M \geq 0, v_1 \geq 0, \ldots, v_N \geq 0} \frac{\theta_j}{\max\limits_k \theta_k}. \tag{6}$$

When for some values of the prices and costs the facility is the most productive, then the value E_j is equal to 1. If for all possible values of the prices and costs, the j-th facility is less productive than others, then $E_j < 1$. This inequality is an indication to the managers that something needs to be done to increase the productivity of this facility.

How to compute the efficiencies. The above expression for E_j is non-linear and sounds complicated. Good news is that this problem is equivalent to solving a *linear programming* problem, i.e., a problem of optimizing a linear expression under linear constraints – and for linear programming problems, there are efficient algorithms and use-friendly software implementing these algorithms; see, e.g., [10].

Specifically, the problem of computing the value E_j is equivalent to solving the following linear programming problem:

$$\text{Maximize } \sum_{m=1}^{M} u'_m \cdot y_{j,m} \tag{7}$$

under the constraints

$$\sum_{m=1}^{M} u'_m \cdot y_{k,m} \le \sum_{n=1}^{N} v'_n \cdot x_{k,n} \text{ for all } k, \tag{8}$$

$$\sum_{n=1}^{N} v'_n \cdot x_{j,n} = 1, \text{ and} \tag{9}$$

$$u'_m \ge 0 \text{ and } v'_n \ge 0 \text{ for all } m \text{ and } n. \tag{10}$$

Comment. For the reader's convenience, the equivalence between these two problems is explained in Section 6, the last section of this paper.

3 Why we need an interval version of DEA and which version has been proposed

Limitations of DEA. Many factors affect productivity. As a result, even when the facility remains the same, productivity changes somewhat from day to day, from year to year. If we simply use the production amounts during the whole period of time, we get a general information about the facility's productivity, but nuances are missing. Some facilities may be always lacking behind, while others may be productive on some days while not very productive on others. These two cases require different managerial approaches:

- if a facility always lags behind, then we need to find out why, and to recommend the corresponding changes;
- on the other hand, if the facility is sometimes very efficient and sometimes not, then our problem becomes easier: we just need to look at the times when the productivity was higher, and try to apply what worked then to other periods of time.

It is therefore desirable to change the definition of the DEA so as to provide such an additional information to the managers. A natural idea is to show, to the managers, the bounds on the efficiency, i.e., the interval $\left[\underline{E}_j, \overline{E}_j\right]$ of possible values of the efficiency of the j-th facility.

Comment. In general, interval uncertainty is ubiquitous, it appears in many application areas; see, e.g., [2, 4, 5, 6].

Current interval version of DEA. To deal with the limitations of DEA, researchers proposed the following interval version of DEA [9] (see also [3, 7, 11]). At different moments of time $t = 1,\ldots,T$, the j-th facility uses different amounts $x_{j,n,t}$ of resources used and different amounts $y_{j,m,t}$ of items produced. So, the researchers suggested to use, instead of the overall values $x_{j,n}$ and $y_{j,m}$, the intervals formed by observed values:

$$\left[\underline{x}_{j,n},\overline{x}_{j,n}\right] = \left[\min_t x_{j,n,t}, \max_t x_{j,n,t}\right]$$

and

$$\left[\underline{y}_{j,m},\overline{y}_{j,m}\right] = \left[\min_t y_{j,m,t}, \max_t y_{j,m,t}\right].$$

In this situation, as a measure of the facility's effectiveness, we take the interval $\left[\underline{E}_j,\overline{E}_j\right]$ of possible values of the ratio E_j corresponding to all possible combinations of values $x_{j,n} \in \left[\underline{x}_{j,n},\overline{x}_{j,n}\right]$ and $y_{j,m} \in \left[\underline{y}_{j,m},\overline{y}_{j,m}\right]$.

Comment. Interestingly, the problem of computing the endpoints $\underline{E}_j$ and $\overline{E}_j$ of this interval also turned out to be equivalent to an appropriate linear programming problem. This interval version is, thus, also feasible to solve.

4 The current interval version of DEA is not always economically adequate: an example

Example. We will illustrate the problem with the current interval version of DEA on the following simple example. Suppose that we have several identical facilities at different locations. We assume that during all the months – with the exception of the three summer months – each facility uses the same number of resources $x_1,\ldots,x_N$ and produces the same number of products $y_1,\ldots,y_M$. In each summer month, due to vacations, production reduces by 20%. So, during these months, each facility uses $0.8\cdot x_n$ resources and produces $0.8\cdot y_m$ products.

From the economic viewpoint, what is the relative efficiency of each facility. Since the facilities are absolutely identical and perform exactly the same way, their productivities θ_j are exactly the same. So, for each j, for each month, we have $\theta_j = \max_k \theta_k$ and thus, $E_j = 1$. We therefore expect all relative efficiencies to be equal to 1, i.e., we expect to have $\left[\underline{E}_j,\overline{E}_j\right] = [1,1]$.

What the current interval version of DEA returns. Let us see what the current version of the DEA returns. For each j, the corresponding intervals are: $\left[\underline{x}_{j,n},\overline{x}_{j,n}\right] = [0.8\cdot x_n, x_n]$ and $\left[\underline{y}_{j,m},\overline{y}_{j,m}\right] = [0.8\cdot y_m, y_m]$.

In this case, one can see that the largest possible values $\overline{E}_j$ of E_j is still 1. However, the smallest possible value $\underline{E}_j$ of the value E_j is now different from 1. Indeed, one can check that it corresponds to the case when:

- the efficiency θ_j of the j-th facility is the smallest possible, i.e., its used resources are the largest possible, while its production is the smallest possible, while

- the efficiencies θ_k of all other facilities $k \neq j$ are the largest possible, i.e., their used resources are the smallest possible, while their production amounts are the largest possible.

So, for the j-th facility, we have $N_j = 0.8 \cdot N_j^{(0)}$ – where for each quantity x, $x^{(0)}$ means its value in an idealized situation where the summer months do not differ from the rest of the year – and $D_j = D_j^{(0)}$. Thus,

$$\theta_j = \frac{N_j}{D_j} = \frac{0.8 \cdot N_j^{(0)}}{D_j^{(0)}} = 0.8 \cdot \frac{N_j^{(0)}}{D_j^{(0)}} = 0.8 \cdot \theta^{(0)}.$$

We skipped the index j for $\theta^{(0)}$, here, since in this situation, all the values θ are the same.

For all other facilities k, we have $N_k = N_k^{(0)}$ and $D_k = 0.8 \cdot D_k^{(0)}$, thus,

$$\theta_k = \frac{N_k}{D_k} = \frac{N_k^{(0)}}{0.8 \cdot D_k^{(0)}} = \frac{1}{0.8} \cdot \frac{N_k^{(0)}}{D_k^{(0)}} = 1.25 \cdot \theta^{(0)}.$$

Thus, $\max_k \theta_k = 1.25 \cdot \theta^{(0)}$ and

$$\underline{E}_j = \frac{\theta_j}{\max_k \theta_k} = \frac{0.8 \cdot \theta^{(0)}}{1.25 \cdot \theta^{(0)}} = 0.64.$$

This is clearly different from the economically meaningful lower endpoint $\underline{E}_j = 1$.

5 Analysis of the problem and the proposed new interval version of DEA

Analysis of the problem. The problem appears because:

- in the current version of interval DEA, we consider all possible combinations of values $x_{j,n}$ and $y_{j,m}$, while
- in reality, there may be a correlation between the values $x_{j,n}$ and $y_{j,m}$ corresponding to the same facility, in which case not all pairs $(x_{j,n}, y_{j,m})$ may be possible.

In particular, in the above example, it is not possible to have $N_k = 0.8 \cdot N_k^{(0)}$ and $D_j = D_k^{(0)}$: in the months when we have $N_k = 0.8 \cdot N_k^{(0)}$, we also have $D_k = 0.8 \cdot D_k^{(0)}$.

Comment. This is, by the way, a typical problem that happens when we naively apply interval techniques; see, e.g., [2, 4, 5, 6]. For example, for simple arithmetic operations like $y = x_1 + x_2$, $y = x_1 - x_2$, and $y = x_1 \cdot x_2$, if we know the intervals $[\underline{x}_i, \overline{x}_i]$ of possible values of each inputs x_i, we can find the range of possible values of y:

- for addition, we have
$$\left[\underline{y},\overline{y}\right]=\left[\underline{x}_1+\underline{x}_2,\overline{x}_1+\overline{x}_2\right];$$
- for subtraction, we have
$$\left[\underline{y},\overline{y}\right]=\left[\underline{x}_1-\overline{x}_2,\overline{x}_1-\underline{x}_2\right];$$
- for multiplication, we have
$$\left[\underline{y},\overline{y}\right]=$$
$$[\min(\underline{x}_1\cdot\underline{x}_2,\underline{x}_1\cdot\overline{x}_2,\overline{x}_1\cdot\underline{x}_2,\overline{x}_1\cdot\overline{x}_2),\max(\underline{x}_1\cdot\underline{x}_2,\underline{x}_1\cdot\overline{x}_2,\overline{x}_1\cdot\underline{x}_2,\overline{x}_1\cdot\overline{x}_2)].$$

So, at first glance, to compute the range of the function $f(x)=x-x^2$ when $x\in[0,1]$, it may seem reasonable to first use the multiplication rule to compute the range of $z\stackrel{\text{def}}{=}x^2$, and then use the subtraction rule to find the desired range of $y=f(x)$. This procedure will lead to the following computations:

$$[\underline{z},\overline{z}]=[0,1]\cdot[0,1]=[\min(0\cdot 0,0\cdot 1,1\cdot 0,1\cdot 1),\max(0\cdot 0,0\cdot 1,1\cdot 0,1\cdot 1)]=$$
$$[\min(0,0,0,1),\max(0,0,0,1)]=[0,1]$$

and
$$\left[\underline{y},\overline{y}\right]=[0,1]-[0,1]=[0-1,1-0]=[-1,1].$$

On the other hand, as one can easily check, the actual range of this function is much narrower: it is $[0,0.25]$.

The reason why we have this problem is that the formula for the subtraction implicitly assumes that it is possible to have all possible pairs (x_1,x_2) for which each x_i belongs to the corresponding interval, while in our case the values x and z are correlated: only pairs $(x,z)=(x,x^2)$ are possible.

Our idea. In general, in interval techniques, the solution is to take the dependence between the variables into account. This is exactly what we propose: instead of forming the intervals for $y_{j,m}$ and $x_{j,n}$ by combining values corresponding to different moments of time, let us:

- use the exact values $y_{j,m,t}$ and $x_{j,n,t}$ for each moment of time t to compute the efficiencies $E_{j,t}$ at this moment of time, and
- then combine the resulting efficiencies $E_{j,t}$ into the interval $\left[\underline{E}_j,\overline{E}_j\right]$.

Thus, we arrive at the following method.

The proposed new version of interval DEA. First, for each moment of time t and for each facility j, we compute the value

$$E_{j,t}\stackrel{\text{def}}{=}\max_{u_1\ge 0,\ldots,u_M\ge 0,v_1\ge 0,\ldots,v_N\ge 0}\frac{\theta_{j,t}}{\max\limits_k \theta_{k,t}},\tag{11}$$

where we denoted

$$\theta_{j,t} \stackrel{\text{def}}{=} \frac{\sum\limits_{m=1}^{M} u_m \cdot y_{j,m,t}}{\sum\limits_{n=1}^{N} v_n \cdot x_{j,n,t}}. \tag{12}$$

Each of these values $E_{j,t}$ can be computed by solving the following linear programming problem:

$$\text{Maximize } \sum_{m=1}^{M} u'_m \cdot y_{j,m,t} \tag{13}$$

under the constraints

$$\sum_{m=1}^{M} u'_m \cdot y_{k,m,t} \le \sum_{n=1}^{N} v'_n \cdot x_{k,n,t} \text{ for all } k, \tag{14}$$

$$\sum_{n=1}^{N} v'_n \cdot x_{j,n,t} = 1, \text{ and} \tag{15}$$

$$u'_m \ge 0 \text{ and } v'_n \ge 0 \text{ for all } m \text{ and } n. \tag{16}$$

Once all the values $E_{j,t}$ are computed, we can compute the desired intervals $\left[\underline{E}_j, \overline{E}_j\right]$ as follows:

$$\left[\underline{E}_j, \overline{E}_j\right] = \left[\min_t E_{j,t}, \max_t E_{j,t}\right].$$

Comment. For each facility j, in addition to knowing the range $\left[\underline{E}_j, \overline{E}_j\right]$ of possible values of its efficiency, we may want to know how frequent are different efficiency values E_j from this range. In other words, we may want to know the probability distribution on this interval.

Good news is that, based on our computations, we have a sample of values $E_{j,1}, \ldots, E_{j,T}$ from this distribution. Thus, we can use the usual statistical methods to estimate this distribution based on the sample; see, e.g., [8].

6 Justification of reducing the original DEA problem to its linear programming reformulation

Let us show that the original DEA problem (6) is indeed equivalent to the linear programming problem (7)-(10). For this purpose, we will show that the value E_j as determined by the formula (6) is equal to the maximum corresponding to the problem (7)-(10); we will denote this maximum by E'_j.

1°. Let us first prove that for each combination of values u_m and v_n, there exist values u'_m and v'_n that satisfy the conditions (8)-(10) for which the sum

$$\sum u'_m \cdot y_{j,m}$$

– that is maximized in the formula (7) – is equal to the ratio $\theta_j/\max\theta_k$ corresponding to u_m and v_n. This would imply that $E'_j \geq E_j$.

To prove this, first, we observe that if we divide all the values u_m by the same positive constant U – i.e., replace the original values u_m with the new values $u'_m = u_m/U$ – then all the numerators N_k will be divide by U, i.e., their new values will be $N'_k = N_k/U$. Indeed, we have

$$N'_k = \sum_{m=1}^{M} u'_m \cdot y_{k,m} = \sum_{m=1}^{M} \frac{u_m}{U} \cdot y_{k,m} = \frac{1}{U} \cdot \sum_{m=1}^{M} u_m \cdot y_{k,m} = \frac{N_k}{U}.$$

Similarly, if we divide all the values v_n by the same positive constant V – i.e., replace the original values v_n with the new values $v'_n = v_n/V$ – then all the denominators D_k will be divide by V, i.e., their new values will be $D'_k = D_k/V$. Indeed, we have

$$D'_k = \sum_{n=1}^{N} v'_n \cdot x_{k,n} = \sum_{n=1}^{N} \frac{v_n}{V} \cdot x_{k,n} = \frac{1}{V} \cdot \sum_{n=1}^{N} v_n \cdot x_{k,n} = \frac{D_k}{V}.$$

Let us take $U = D_j/\max\theta_k$ and $V = D_j$. Then, we have $D'_j = D_j/V = 1$, i.e., we have the formula (9). For each k, we have

$$\frac{N'_k}{D'_k} = \frac{N_k/(D_j \cdot \max\theta_\ell)}{V_k/D_j} = \frac{N_k}{D_k} \cdot \frac{1}{\max\theta_\ell} = \frac{\theta_k}{\max\limits_{\ell}\theta_\ell}.$$

By definition, each value is smaller or equal than the maximum, thus, we have $N'_k/D_k \leq 1$, i.e., $N'_k \leq D'_k$, which is exactly the condition (8). In particular, for j, we have

$$\frac{N'_j}{D'_j} = \frac{\theta_j}{\max\limits_{k}\theta_k}.$$

Since $D'_j = 1$, we conclude that

$$N'_j = \frac{\theta_j}{\max\theta_k},$$

i.e., indeed, that the value of the expression (7) corresponding to the new values u'_m and v'_n is equal to the ratio $\theta_j/\max\theta_k$ corresponding to the original values u_m and v_n.

2°. To complete our justification, let us show that, vice versa, for every combination of values u'_m and v'_n that satisfy the conditions (8)-(10), there exist values u_m and v_n for which the ratio $\theta_j/\max\theta_k$ is larger than or equal to the sum $\sum u'_m \cdot y_{j,m}$ – the sum that is maximized in the formula (7). This would imply that $E'_j \leq E_j$ – which, together with the inequality $E'_j \geq E_j$ proved in the first part of this justification, will imply that $E'_j = E_j$.

Indeed, let us take $u_m = u'_m$ and $v_n = v'_n$. Due to (8), we have $N_k \le D_k$ and thus, $\theta_k N_k / D_k \le 1$ for all k. This implies that $\max_k \theta_k \le 1$. Due to (9), we have $D_j = 1$, so $\theta_j = N_j / D_j = N_j$ is equal to the expression (7). Since $\max \theta_k \le 1$, we conclude that

$$\frac{\theta_j}{\sum_k \theta_k} \ge \theta_j = D_j.$$

Thus indeed the resulting ratio $\theta_j / \max \theta_k$ is larger than or equal to the value (7).

The justification is thus proven.

Acknowledgments

This work was supported by the AT&T Fellowship in Information Technology, by the Institute for Risk and Reliability, Leibniz Universitaet Hannover, Germany, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329, by the European Union under the project ROBOPROX (No. CZ.02.01.01/00/22 008/0004590), by the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand, by the Ho Chi Minh City University of Banking, Vietnam, and by Thang Long University, Hanoi, Vietnam.

References

1. S. Boubaker and T. Ngo, *Handbook on Data Envelopment Analysis in Business, Finance, and Sustainability: Recent Trends and Developments*, World Scientific, London, UK, 2024.
2. L. Jaulin, M. Kiefer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics*, Springer, London, 2012.
3. M. Kazemi and A. Alimi, "A fully fuzzy approach to data envelopment analysis", *Journal of Mathematics and Computer Science*, 2014. Vol. 11, pp. 238–245.
4. B. J. Kubica, *Interval Methods for Solving Nonlinear Constraint Satisfaction, Optimization, and Similar Problems: from Inequalities Systems to Game Solutions*, Springer, Cham, Switzerland, 2019.
5. G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.
6. R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.
7. P. C. Pendharkar, "Data envelopement analysis models for probabilistic classification", *Computers & Industrial Engineering*, 2018, Vol. 119, pp. 181–192.
8. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
9. Y. G. Smirlis, E. K. Maragos, and D. K. Despotis, "Data envelopment analysis with missing values: an interval DEA approach", *Applied Mathematics and Computation*, 2006, Vol. 177, pp. 1–10.
10. R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, Springer, New York, 2014.
11. J. Zhu, "Imprecise data envelopment analysis (IDEA): a review and improvement with an application", *European Journal of Operational Research*, 2003, Vol. 144, pp. 513–529.