Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp616-621)

THE PROBLEM OF INTERVALS IN IMPRECISE DATA ENVELOPMENT ANALYSIS (IDEA)

D. V. DERPANIS and E. FOUNDAS
Department of Informatics
University of Piraeus
80 Karaoli & Dimitriou, 18534 Piraeus
GREECE

Abstract: - The standard data envelopment analysis (DEA) method requires that the values for all inputs and outputs are known exactly. However, this assumption may not be always valid. For example, some outputs and inputs may be only known as in forms of interval data, ordinal data. This model is called imprecise DEA (IDEA). In this paper we try to study the way we could limit the large intervals of DMUs in output level as well as in input level (saving resources) without affecting DMUs' efficiency.

Key-Words: Data envelopment analysis; Interval data; Ordinal data; Imprecise data

1. Introduction

Data envelopment analysis (DEA) [1] is a non–parametric method for evaluating the relative efficiency of decision–making units (DMU) on multiple inputs and outputs .The CCR(Cooper, Charnes, Rhodes) model assumes that data on the outputs and inputs are known exactly. However, this assumption may not be true. For example, some outputs and inputs may not be known as in forms of bounded data, ordinal data, and ratio bounded data. If we incorporate such imprecise data information into the standard linear CCR model, the resulting DEA model is a non linear and non convex program, and is called imprecise DEA (IDEA). Note that IDEA is a deterministic programming approach, although it deals with data variations. IDEA is different from the stochastic or chance constrained DEA approach where imprecise data are estimated with probabilities (see e.g., Cooper et al., 1998) [2].

Cooper et al.(1999) [3] discuss how to deal with bounded data and weak ordinal data and provide a unified IDEA model when weight restriction are also present Kim.(1999) discusses how to deal with bounded data (strong and weak) ordinal data, and ratio bounded data with an application to a set telephone offices.

According to Despotis and Smirlis [4] who have developed an alternative approach for dealing with imprecise data (mixtures of exact, interval and ordinal data in the same setting), they have transformed the non-linear DEA model to a linear programming equivalent by using a straightforward formulation, completely different than that in IDEA. Contrarily to IDEA, theirs transformations on the variables were made on the basis of the original data set, without applying any scale transformations on the data. The original CCR DEA model with exact data, in its multiplier form, is derived then straightforwardly as a special case of theirs model. The potential of theirs transformations enable them to uncover and thoroughly examine some new aspects of efficiency in an imprecise data setting, such as the variation of the efficiency scores of the units. On the basis of their particular transformations, new models were naturally introduced to estimate upper and lower bounds of the efficiency scores of the units, as well to classify and further discriminate the units in terms of the variability of their efficiency scores.

Today organizations want to maximize their outputs by simultaneously minimizing their inputs. So we try in this research to find the minimum of the maximum output values of DMUs' intervals and also the maximum of the minimum input values in which the DMUs lose their efficiency.

In Section 2, we present the existing DEA model for dealing with interval data. Then, on the basis of this model, we define upper and lower bound efficiencies for the units. In Section 3, we proceed still further in formulating another post-DEA model. We limit potentially large intervals of input-output data and thus the uncertainty of input-output estimates, without affecting the efficiency of the DMUs. We deal with the problem of estimating these thresholds. In Section 4 we provide numerical example to illustrate the applications of interval IDEA models. Conclusions are given in Section 5.

2.1 DEA and IDEA models

Assume n units, each using m inputs to produce s outputs. We denote by y_{rj} the level of the rth output (r =1,…, s) from unit j (j=1,…,n) and by x_{ij} the level of the ith input (i=1,…, m) to the jth unit. Let j_0 be the evaluated unit. In such a setting, the following CCR DEA model:

$$\max \quad h_{j_0} = \sum_{r=1}^{s} u_r y_{rj_0}$$

$$\text{st} \quad \sum_{i=1}^{m} v_i x_{ij_0} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, \qquad j=1,\ldots..n$$

$$u_r, v_i \ge \varepsilon \quad , \forall \quad r \quad , i$$

MODEL (1)

Unlike the original DEA model, we assume further that the levels of inputs and outputs are not known exactly; the true input-output data are known to lie within bounded intervals, i.e. $x_{ij} \in [x_{ij}^L, x_{ij}^U]$ and $y_{rj} \in [y_{rj}^L, y_{rj}^U]$, with the upper and lower bounds of the intervals given as constants and assumed strictly positive. Let j_0 be the evaluated unit. In such a setting, the CCR DEA model is non-linear (non-convex) as, apart from the original variables $u_1, \ldots, u_r, \ldots, u_s$ and $v_1, \ldots, v_i, \ldots, v_m$ (weights for outputs and inputs, respectively), the levels of inputs x_{ij} and outputs y_{rj} are also variables whose exact values are to be estimated. According to Despotis and Smirlis (EJOR 2002) we have

$$x_{ij} = x_{ij}^L + s_{ij}(x_{ij}^U - x_{ij}^L),\ i = 1,\ldots, m;\ j = 1,\ldots, n$$

with $\quad 0 \le s_{ij} \le 1,$

$$y_{rj} = y_{rj}^L + t_{rj}(y_{rj}^U - y_{rj}^L),\ r = 1,\ldots, s\ ;\ j = 1,\ldots, n$$

with $\quad 0 \le t_{rj} \le 1$

With these transformations, the variables x_{ij} and y_{rj} in model (1) are replaced by the new variables s_{ij} and t_{rj}, which locate the levels of inputs and outputs within the bounded intervals $[x_{ij}^L, x_{ij}^U]$ and $[y_{rj}^L, y_{rj}^U]$ respectively. Model (1) still remains non-linear due to the products of variables $v_i s_{ij}$ for inputs and $u_r t_{rj}$ for outputs. We then replace these products with new variables $q_{ij=}$ $v_i s_{ij}$ and $p_{rj} = u_r t_{rj}$. According to these transformations the weighted sum of inputs (composite input) for unit j in model (1) takes the form

$$\sum_{i=1}^{m} v_i x_{ij} = \sum_{i=1}^{m} v_i[x_{ij}^L + s_{ij}(x_{ij}^U - x_{ij}^L)] =$$

$$\sum_{i=1}^{m} v_i x_{ij}^L + v_i s_{ij}(x_{ij}^U - x_{ij}^L) = \sum_{i=1}^{m} v_i x_{ij}^L + q_{ij}(x_{ij}^U - x_{ij}^L)$$

Where the new variables q_{ij} meet the conditions $0 \le q_{ij} \le v_i$ as it is $s_{ij} = q_{ij}/v_i$ with $v_i \ge \varepsilon$ and $0 \le s_{ij} \le 1$ for every i and j. Similarly, the weighted sum of outputs (composite output) for unit j takes the form

$$\sum_{r=1}^{s} u_r y_{rj} = \sum_{r=1}^{s} u_r[y_{rj}^L + t_{ij}(y_{rj}^U - y_{rj}^L)] =$$

$$\sum_{r=1}^{m} u_r y_{rj}^L + u_r t_{rj}(y_{rj}^U - y_{rj}^L) = \sum_{r=1}^{s} u_r y_{rj}^L + p_{rj}(y_{rj}^U - y_{rj}^L)$$

With $0 \le p_{rj} \le u_r$ as it is $t_{ij} = p_{ij}/u_i$ with $u_i \ge \varepsilon$ and $0 \le t_{ij} \le 1$ for every r and j as explained above.

With the above substitutions, model (1) is finally transformed into the following linear program:

$$\max h_{j_0} = \sum_{r=1}^{s} u_r y_{rj_0}^L + p_{rj_0}(y_{rj_0}^U - y_{rj_0}^L)$$

$$s.t.$$

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 1$$

$$\sum_{r=1}^{s} u_r y_{rj}^L + p_{rj}(y_{rj}^U - y_{rj}^L) -$$

$$\sum_{i=1}^{m} v_i x_{ij}^L + q_{ij}(x_{ij}^U - x_{ij}^L) \le 0 \quad j = 1,\ldots, n$$

$$p_{rj} - u_r \le 0 \quad r = 1,\ldots, s\ ;\ j = 1,\ldots, n$$

$$q_{ij} - v_i \le 0 \quad i = 1,\ldots, m\ ;\ j = 1,\ldots, n$$

$$u_r, v_i \ge \varepsilon \quad \forall r, i$$

$$p_{rj} \ge 0,\ q_{ij} \ge 0 \quad \forall r, i, j$$

MODEL (2)

2.2. Upper and lower bounds of efficiency scores

According to Despotis and Smirlis in an interval data setting, many units are likely to be proved efficient, as apart from the flexibility they have in choosing the weights, they are also free to adjust the levels of inputs and outputs in a favorable manner within the intervals. Thus further discrimination of the efficient units becomes more essential in an interval data setting.

So, the models (2) provide for each unit a bounded interval $[h_j^L, h_j^*]$ which is founded according to follow models in which its possible efficiency scores lie, from the best to worst the case.

$$\max \quad h_j^* = \sum_{r=1}^{s} u_r y_{rj_0}^U$$

$$\text{st} \quad \sum_{i=1}^{m} v_i x_{ij_0}^L = 1$$

$$\sum_{r=1}^{s} u_r y_{rj_0}^U - \sum_{i=1}^{m} v_i x_{rj_0}^L \le 0$$

$$\sum_{r=1}^{s} u_r y_{rj}^L - \sum_{i=1}^{m} v_i x_{rj}^U \le 0 \quad \text{j=1,…..n;}$$

$$j \ne j_0 \qquad u_r, v_i \ge \varepsilon \ , \forall \ r \ , i$$

MODEL (3)

For the evaluated unit, the inputs are adjusted at the lower bounds and the outputs at the upper bounds of the intervals. Unfavorably for the other units, the inputs are contrarily adjusted at their upper bounds and the outputs at their lower bounds.

$$\max \quad h_j^L = \sum_{r=1}^{s} u_r y_{rj_0}^L$$

$$\text{st} \quad \sum_{i=1}^{m} v_i x_{ij_0}^U = 1$$

$$\sum_{r=1}^{s} u_r y_{rj_0}^L - \sum_{i=1}^{m} v_i x_{rj_0}^U \le 0$$

$$\sum_{r=1}^{s} u_r y_{rj}^U - \sum_{i=1}^{m} v_i x_{rj}^L \le 0 \quad \text{j=1,…..n;}$$

$$j \ne j_0$$

$$u_r, v_i \ge \varepsilon \ , \forall \ r \ , i$$

MODEL (4)

For the evaluated unit the inputs are adjusted at their upper bounds and the outputs at their lower bounds. For the other units, the inputs are favorably adjusted at their lower bounds and the outputs at their upper bounds.

On the basis of the above efficiency score intervals, the units can be first classified in three subsets as follows:

$E^{++} = \{J \ \varepsilon \ J / h_j^L \ = 1\}$

$E^{+} = \{J \ \varepsilon \ J / h_j^L \ < 1 \text{ and } h_j^* = 1\}$

$E^{-} = \{ J \varepsilon J / h_j^* < 1\}$

where J stands for the index set (1,…,n) of the units. The set E^{++} consists of the units that are efficient in any case (any combination of input/output levels). The set E^{+} consists of units that are efficient in a maximal sense, but there are input/output adjustments under which they cannot maintain their efficiency.

Finally, the set E^{-} consists of the definitely inefficient units. Moreover, the range of possible efficiency scores can be used to rank further the units in the set E^{+}

3. An extension of the interval DEA model for dealing with imprecise data

We will examine only the DMUs belonging to E^{+} because the units of E^{++} set conserve their efficiency for any value input/output levels (so they can take the minimum input value – the minimum possible cost- and the lowest output value) while the units of E^{-} set never succeed to become efficient. We examine unit j_0 : We assume that an input (say input g) exists which does not need to take the minimum value so that the DMU becomes efficient. That is to say

$$x_{gj_0} = x_{gj_0}^L + s_{gj_0}(x_{gj_0}^U - x_{gj_0}^l) .$$

That means that we want to find a value for x_{gj_0} that would be smaller than $x_{gj_0}^U$. This value can be achieved by estimating q_{gj_0} and u_g that maximize $s_{gj_0} = q_{gj_0} /$ u_g. The model below accomplishes

max z

s.t (u,v,Q,P) ε S,

q_{gj_0} -z v_g ≥ 0

where u =(u_r r=1,…s) , v=(v_i i=1,….m), Q=(q_{ij} i=1,…,m; j=1,..n) and P=($P_{rj,}$ r=1,….s; j=1,..,n) are the decision variables in vector form, the variable z represents the maximum value of the ratio $s_{gj_0} = q_{gj_0} /$ u_g and S is the solution space formed by the following set of constraints:

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 1$$

$$\sum_{r=1}^{s} u_r y_{rj_0}^L + p_{rj_0}(y_{rj_0}^U - y_{rj_0}^L) -$$

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 0$$

$$\sum_{r=1}^{s} u_r y_{rj}^L + p_{rj}(y_{rj}^U - y_{rj}^L) - \sum_{i=1}^{m} v_i x_{ij}^L + q_{ij}(x_{ij}^U - x_{ij}^L) \le 0$$

j=1,…n(j$\neq j_0$)

$p_{rj} - u_r \le 0$ r=1,..,s ; j=1,…n

$q_{ij} - v_i \le 0$ i=1,..,m ; j=1,…n

$p_{rj} \ge 0$ $\forall r, j$

$q_{ij} \ge 0$ $\forall i, j$

$u_r, v_i \ge \varepsilon$,$\forall r, i$

MODEL (5)

Model (5) is a non linear due to the last constraint. However it is possible to solve by resorting to standard LP software, with a two stage procedure as follows

Stage 1 we solve the linear program

max q_{gj_0}

s.t (u,v,Q,P) ε S,

If $q_{gj_0}^0$, v_g^0 are the values of the variables q_{gj_0}, v_g in the optimal solution of (5), then the ratio $q_{gj_0}^0 / v_g^0$ is a value of z for which the unit j_0 becomes efficient as it satisfies, among others, the first two constraints. On the other hand z > 0 as q_{gj_0}>0. So the optimal (maximum) value of z will lie in the bounded interval [0, ($q_{gj_0}^0 / v_g^0$)].

Stage 2. On the basis of model (5), we perform bisection search in the interval [0, ($q_{gj_0}^0 / v_g^0$)] as follows. Let $\underline{z}$ be a value of z for which the constraints of model (5) are consistent (initially $\underline{z} = q_{gj_0}^0 / v_g^0$) and $\bar{Z}$ a value of z for which the constraints of (5) are not consistent (initially $\bar{Z}$=0). Then the consistency of the constraints is investigated for **z′** = ($\underline{Z}$ + $\bar{Z}$)/2. If they are consistent **z′** will replace $\underline{z}$ if they are not it will replace $\bar{z}$. The bisection is continued until $\underline{z}$ and $\bar{Z}$ come sufficiently close to each other, at a desirable degree of accuracy. We end this iterative stage with z*= $\underline{z} \cong \bar{Z}$ and (u*,v*,Q*,P*) that is an optimal solution of model (3) (i.e., $z^* = q_{kj_0}^* / v_g^*$). The threshold of efficiency $\bar{x}_{gj_0}$ derives then from $\bar{x}_{gj_0} = x_{gj_0}^L + (q_{gj_0}^* / v_g^*)(x_{gj_0}^U - x_{gj_0}^L)$.

However, if we want to be absolutely certain we must examine the worst case of DMU j_0 which it means that all the rest DMUs have taken their best position (minimum inputs and maximum outputs)

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 1$$

$$\sum_{r=1}^{s} u_r y_{rj_0}^L + p_{rj_0}(y_{rj_0}^U - y_{rj_0}^L) -$$

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 0$$

$$\sum_{r=1}^{s} u_r y_{rj}^U - \sum_{i=1}^{m} v_i x_{ij}^L \le 0 \quad \text{j=1,…n} \quad (j \neq j_0)$$

$p_{rj_0} - u_r \le 0$ r=1,..,s ;

$q_{ij_0} - v_i \le 0$ i=1,..,m ;

$p_{rj} \ge 0$ $\forall r, j$

$q_{ij} \ge 0$ $\forall i, j$

$u_r, v_i \ge \varepsilon$,$\forall r, i$

MODEL (6)

The things are somehow different for the output. Thus model (5) becomes:

min p_{gj_0}

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 1$$

$$\sum_{r=1}^{s} u_r y_{rj_0}^L + p_{rj_0}(y_{rj_0}^U - y_{rj_0}^L) -$$

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 0$$

$$\sum_{r=1}^{s} u_r y_{rj}^L + p_{rj}(y_{rj}^U - y_{rj}^L) - \sum_{i=1}^{m} v_i x_{ij}^L + q_{ij}(x_{ij}^U - x_{ij}^L) \le 0$$

j=1,…n (j$\neq j_0$)

$p_{rj_0} - u_r \le 0$ r=1,..,s ;

$q_{ij_0} - v_i \le 0$ i=1,..,m ;

$u_r, v_i \ge \varepsilon$,$\forall r, i$

$p_{rj} \ge 0$ $\forall r \neq g, j \neq j_0$

$q_{ij} \ge 0$ $\forall i, j$

p_{gj_0} free

MODEL (7)

Since p_{gj_0} can take also negative values provided that DMU j₀ can be efficient even if it has an output smaller than the low limit of his interval.

As long as the model (5) in the case of output concerns it is precisely itself

Min p_{gj_0}

$$\sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 1$$

$$\sum_{r=1}^{s} u_r y_{rj_0}^L + p_{rj_0}(y_{rj_0}^U - y_{rj_0}^L) - \sum_{i=1}^{m} v_i x_{ij_0}^L + q_{ij_0}(x_{ij_0}^U - x_{ij_0}^L) = 0$$

$$\sum_{r=1}^{s} u_r y_{rj}^U - \sum_{i=1}^{m} v_i x_{ij}^L \le 0$$ j=1,…n (j$\neq j_0$)

$p_{rj_0} - u_r \le 0$ r=1,..,s ;

$q_{ij_0} - v_i \le 0$ i=1,..,m ;

$p_{rj} \ge 0$ $\forall r \neq g, j \neq j_0$

$q_{ij} \ge 0$ $\forall i, j$

p_{gj_0} free

MODEL (8)

4. A numerical example

Assume that eight units are evaluated based on their efficiency according to the inputs / outputs of the following table, all with imprecise data and with no information given for the price allocation in the intervals.

	INPUT		OUTPUT	
DMU J	***X₁J***	***X₂J***	***Y₁J***	***Y₂J***
1	16 21	0.30 0.50	120 125	19 21
2	18 25	0.44 0.53	122 130	20 21
3	20 27	0.41 0.61	124 131	16 24
4	12 15	0.21 0.48	138 144	21 22
5	10 17	0.1 0.7	143 159	28 35
6	4 30	0.16 0.35	157 198	21 29
7	19 22	0.12 0.19	158 181	21 25
8	14 15	0.06 0.09	157 161	28 40

DMU J	**E**
1	**E⁻**
2	**E⁻**
3	**E⁻**
4	**E⁺**
5	**E⁺**
6	**E⁺**
7	**E⁻**
8	**E⁺⁺**

We apply model (5) in DMU6 and we have: q_{16}^*=0,030328 and v_6^*=0,052864.

So $x_{gj_0} = x_{gj_0}^L + (q_{gj_0}^* / v_g^*)(x_{gj_0}^U - x_{gj_0}^l) =$ 4+0.030328 / 0.052864(30-4) =18.9 which declares that the input of DMU6 can take values between 18.9 and 30 because for any given value from 4 to 18.9 DMU6 is efficient.

We apply model (6) in DMU6: q_{16}^*=0.026105 and v_1^* =0.080314.

So $x_{gj_0} = x_{gj_0}^L + (q_{gj_0}^* / v_g^*)(x_{gj_0}^U - x_{gj_0}^l) =$ 4+0.026105 / 0.080314(30-4)=12.4 which declares that the input of DMU6 can take values between 12.4 and 30 because for any given value from 4 to 12.4 DMU6 is efficient.

We apply model (7) in DMU4 and we have: p_{14}^*= -0.016449 and u_1^* = 0,007959.

So $y_{gj_0} = y_{gj_0}^L + (p_{gj_0}^* / u_g^*)(y_{gj_0}^U - y_{gj_0}^l) =$

138-0.016449 / 0.007959(144-138) =125.6 which means that DMU4 can be efficient even if its output is less than its low limit

We apply model (8) in DMU4 and we find that solution does not exist inside the interval.

5. Conclusion

We developed in this paper an alternative approach for dealing with imprecise data in DEA.
Firstly we wanted to restrict the efforts which each DMU does but at the same time remains efficient. Secondly we wanted to minimize the intervals without losing significant information. To sum up we want to find the thresholds of inputs and outputs where DMU find and lose their efficient

References

[1] A. Charnes, W.W. Cooper, A.Y. Lewin and L. M. Seiford, *Data Envelopment Analysis: Theory, Methodology and Applications*, Kluwer Academic Publishers, Norwell, MA, 1994.
[2] W. W. Copper ,Huang V.Lelas ,S.X Li and O.B Olesen, “Chance Constrained Programming Formulations For Stochastic Characterizations of Efficiency and Dominance in DEA’’ *Journal of productivity Analysis,* Vol.9, No. 1, 1998, pp. 53-79.
[3] W. W. Copper, K. S. Park, and G. Yu, “IDEA and AR-IDEA: Models for dealing with imprecise data in DEA”, *Management Science* Vol. 45, 1999, pp. 597-607.
[4] D. K. Despotis and Y. G. Smirlis, “Data envelopment analysis with imprecise data”, *European Journal of Operational Research* Vol. 140, 2002, pp. 24-36.