International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:12, No:6, 2018

Classifying and Predicting Efficiencies Using Interval DEA Grid Setting

Yiannis G. Smirlis

***Abstract*—**The classification and the prediction of efficiencies in Data Envelopment Analysis (DEA) is an important issue, especially in large scale problems or when new units frequently enter the under-assessment set. In this paper, we contribute to the subject by proposing a grid structure based on interval segmentations of the range of values for the inputs and outputs. Such intervals combined, define hyper-rectangles that partition the space of the problem. This structure, exploited by Interval DEA models and a dominance relation, acts as a DEA pre-processor, enabling the classification and prediction of efficiency scores, without applying any DEA models.

***Keywords*—**Data envelopment analysis, interval DEA, efficiency classification, efficiency prediction.

I. Introduction

DATA Envelopment Analysis (DEA) [1] is a non-parametric linear programming method for measuring the relative efficiency of organizational units on the basis of multiple inputs and outputs. DEA achieves to classify the units in efficient and non-efficient and to estimate their efficiency score.

For the detection of efficient and inefficient units, except the typical DEA LP formulation, several other techniques have been developed, falling to the category of DEA preprocessors [2]-[4]. These are simple and quick computational procedures that provide useful, initial information about the efficiency of the units, prior to solving any DEA problem. The efficiency information provided by DEA pre-processors, may result to substantial savings in running LP programs especially in the case of large scale DEA problems. Ali [5], based to the idea that the dominated units are evidently inefficient, developed the so called reduced basis entry (RBE) technique to speed up the computational effort for additive and multiplicative DEA models.

For the problem of detecting inefficient units in DEA, in this paper we present a technique that partitions the hyperspace of a DEA problem using a grid. The areas that comprise the grid have the form of hyper-rectangles, defined as the Cartesian products of intervals that segmentize the inputs and outputs of the problem. The grid and the particular shape of the hyper-rectangles enable to extend the dominating relation from single units to areas of hyperspace so to identify whole groups of inefficient units. Additionally, it defines a data setting which is exploited by Interval DEA (IDEA) and classifies the hyper-rectangles into efficient, efficient in a maximal sense, and inefficient. All the units belonging to inefficient hyper-rectangles are certainly inefficient units. The two techniques, act as a DEA preprocessor in the sense that can detect inefficient units in a DEA problem. This is particularly useful in the case of large-scale DEA problems for which the proposed technique can reduce their size, without solving it. The proposed partitioning grid of hyper-rectangles has the additional advantage to predict the efficiency class of any new unit that enters the problem using only comparisons.

The rest of this paper is unfolded as follows: Section II describes the definition of the partitioning grid of hyper-rectangles, Section III presents the IDEA modelling that assesses the hypr-rectangle efficiency, Section IV extends the domination issue to the groups of units, and Section V presents the exploitation of the methodology using an arithmetic example.

Yiannis G. Smirlis is with University of Piraeus, School of Economics, Business and International Studies, Address. 80 Karaoli & Dimitriou str, 18534 Piraeus, Greece (phone: +3210-4142173, fax. +32104142180, e-mail: smirlis@unipi.gr).

II. Definition of a Partitioning Grid of Hyper-Rectangles

Assume a typical DEA problem consisted of a set of n units $D=\{d_1,d_2,...,d_n\}$ evaluated by m inputs and s outputs. For a given unit j, let y_{rj} be the level of the rth output (r =1,…, s) and x_{ij} the level of the ith input (i=1,…, m). Each unit j can be regarded as a data point represented by the vector $(X_j,Y_j)=(x_{1j},x_{2j},..,x_{mj},y_{1j},y_{2j},...,y_{sj})$ that lies in an $m+s$ multidimensional space (the m inputs and the s outputs are its dimensions). DEA theory underlies that such points that correspond to the under-evaluation units define a polyhedral set whose boundary is formed by the efficient units. The shape of the polyhedron is depended on the returns to scale assumption and can range from an unbounded polyhedron to a convex hull. Inefficient units are interior points in this structure. The problem of discriminating efficient and inefficient units is equivalent to determine the boundary set of this polyhedral set of points.

To approach the efficiency classification problem, we first partition the range of values in inputs, outputs by using a number of breakpoints. The breakpoints form intervals that separate the area of the polyhedral set. The Cartesian product of these intervals, seen from geometrical viewpoint, defines a grid of hyper-rectangles (hyper-rectangle is the generalization in the multidimensional space of a typical two-dimensional rectangle) that partition the area of the polyhedral set.

The so formed grid can be used for the efficiency classification analysis. The notation and formal definition of the grid is as follows.

For a given input i, let $[l_i, h_i]$ be the range of values over the

entire set of DMUs (i.e. $l_i = min_j\{x_{ij}\}$ and $h_i = max_j\{x_{ij}\}$). We segmentize the interval [l_i, h_i] by considering a number α_i of breakpoints $x_i^1, x_i^2, ..., x_i^{a_i}$ with $x_i^1 = l_i$ and $x_i^{a_i} = h_i$ so the range [l_i, h_i] to be covered by the intervals $[x_i^1,..,x_i^2],[x_i^2,..,x_i^3],..,[x_i^{a_i-1}...,x_i^{a_i}]$. The breakpoints $x_i^1, x_i^2, ..., x_i^{a_i}$ are defined so every intermediate interval to contain at least one value x_{ij}. As the intervals overlap at their upper bound point, in order to make them distinct (non-overlapping), we introduce a discrimination factor δ that is added to the upper bound of every interval. This arrangement makes the intervals take the form $[x_i^1,..,x_i^2],[x_i^2+\delta,..,x_i^3],..,[x_i^{a_i-1}+\delta...,x_i^{a_i}]$. Factor δ takes small, non-significant values, e.g. 10^{-3}, 10^{-6} depending on the unit of measurement of the particular input, output. The above interval segmentation is defined in all inputs and outputs of the problem. Similar is the arrangement for the outputs: the range [l_r, h_r] between the minimum and maximum value for any output r can be segmented by using b_r in number breakpoints $y_i^1, y_i^2, ..., y_i^{b_r}$ so the associated intervals will take the form $[y_r^1,..,y_r^2],[y_r^2+\delta,..,y_r^3],..,[y_r^{b_r-1}+\delta,..,y_r^{b_r}]$.

The Cartesian product of segmenting intervals for all inputs and outputs (i.e. all possible combinations of intervals in inputs and outputs) defines non-overlapping, non-empty hyper-rectangles that cover the space of the polyhedral set. Note that the number p of all the covering hyper-rectangles is $p = (a_1-1)(a_2-1)...(a_m-1)(b_1-1)(b_2-1)...(b_s-1)$.

Each different hyper-rectangle combination of intervals in Table I presents the whole set (grid) of such hyper-rectangles.

TABLE I
THE HYPER-RECTANGLES OF THE GRID DEFINED BY THE COMPOSING INTERVALS

	INPUTS			
	x_1	x_2		x_m
G_1	$[x_1^1,..,x_1^2]$	$[x_2^1,..,x_2^2]$	...	$[x_m^1,..,x_m^2]$
G_2	$[x_1^2+\delta,..,x_1^3]$	$[x_2^1,..,x_2^2]$	...	$[x_m^1,..,x_m^2]$
...	...	...	...	...
G_{a_1}	$[x_1^{a_1-1}+\delta,..,x_1^{a_1}]$	$[x_2^1,..,x_2^2]$		$[x_m^1,..,x_m^2]$
G_{a_1+1}	$[x_1^1,..,x_1^2]$	$[x_2^{a_2-1}+\delta,..,x_2^{a_2}]$	...	$[x_m^1,..,x_m^2]$
...	...	...	...	...
G_p	$[x_1^{a_1-1}+\delta,..,x_1^{a_1}]$	$[x_2^{a_2-1}+\delta,..,x_2^{a_2}]$	...	$[x_m^{a_m-1}+\delta,..,x_m^{a_m}]$
	OUTPUTS			
	x_1	x_2		x_m
G_1	$[y_1^1,..,y_1^2]$	$[y_2^1,..,y_2^2]$	...	$[y_s^1,..,y_s^2]$
G_2	$[y_1^1,..,y_1^2]$	$[y_2^1,..,y_2^2]$	...	$[y_s^1,..,y_s^2]$
...	...	...	...	...
G_{a_1}	$[y_1^1,..,y_1^2]$	$[y_2^1,..,y_2^2]$	...	$[y_s^1,..,y_s^2]$
G_{a_1+1}	$[y_1^1,..,y_1^2]$	$[y_2^1,..,y_2^2]$	...	$[y_s^1,..,y_s^2]$
...	...	...	...	...
G_p	$[y_1^{b_1-1}+\delta,..,y_1^{b_1}]$	$[y_2^{b_2-1}+\delta,..,y_2^{b_2}]$	...	$[y_s^{b_s-1}+\delta,..,y_s^{b_s}]$

The dataset of Table I, viewed as a DEA input-output setup, is a typical Interval DEA dataset, having intervals in the place of crisp values. The interval units G_1, G_2 ,..., G_p are hyper-rectangles assessed by m inputs and s outputs. It is important to notice that, by their definition, the hyper-rectangles G_1, G_2 ..., G_p form a partition of the space and any unit $d_1, d_2, ..., d_n$ of the initial problem belongs to only one hyper-rectangle G_1, G_2 ..., G_p. If by J_k we denote the set of indexes of the units d_j that belong to the group G_k, the structure of groups G_k is described by the following relations:

i) $J_1, J_2, .., J_p \subset \{1,2,...,n\}$

ii) $J_i \cap J_j = \varnothing, \forall i \neq j$

iii) $J_1 \cup J_2 \cup .. \cup J_p = \{1,2,..,n\}$.

Note that when the definition of breakpoints is not feasible or not implied by the specific application, a k-means clustering procedure [6] can be also applied to define the G_k hyper-rectangles. Such a procedure is based on the Euclidean distance and can separate the units into relative homogeneous groups (units regarded as points in the multidimensional space are arranged in relative small distances from the center of the group while the group centers are relatively apart). The interval definitions for any groups G_k can be obtained by the minimum in inputs and maximum in outputs values of the included member units $d_k = (X_k, Y_k)$ belonging to group G_k.

For the compound units G_1, G_2 ..., G_p , an efficiency analysis is possible both by applying IDEA models to obtain an efficiency classification and by identifying the dominated units.

III. Estimation of Hyper-Rectangles Efficiency Scores Using IDEA

Interval DEA (IDEA), based on a dataset consisted on intervals as in Table I, can discriminate all the groups into classes of efficiency. In this section, we briefly present the basic models of IDEA and we demonstrate their use in the case of the hyper-rectangle grid structure. The terms and formulations are as follows.

Unlike the original DEA model, IDEA assumes that, in the place of the crisp input x_{ij} and output y_{rj} values, there exist bounded intervals $[x_{ij}^L, x_{ij}^U]$ and $[y_{rj}^L, y_{rj}^U]$, with strictly positive constant bounds. In such a setting, any unit j_0 is free to assign any value within the intervals, so its efficiency score lies between a minimum value $h_{j_0}^L$ and a maximum value, $h_{j_0}^U$. IDEA has been introduced by Cooper et al. [7]. Despotis and Smirlis [8] treated both interval and ordinal data using variable transformations and introduced a three-group efficiency classification of the units (E^{++}, E^+, E^-) instead of efficient and non-efficient partitions that typical DEA arranges. Wang et al. [9] proposed new, simpler models to use common production frontiers for all units.

Following the approach of [8], we note that the maximum possible efficiency score $h_{j_0}^U$ is obtained when the evaluated unit j_0 is set to its most favourable position, that is, when it arranges its inputs to the lower bound and its outputs to the upper bound while all the rest units are set to their least

favourable position (inputs are set to the upper bound and outputs to the lower bound). Note that for any hyper-rectangle G_k, its most favorable position is the point with its lowest input and highest output i.e. $d_k^U = (X_k^L, Y_k^U)$ where $X_k^L \le \min X_i$ and $Y_k^U \ge \max Y_i, i \in J_k$. The point d_k^U may be a real unit in the initial problem or an artificial one. Similarly, the worst position of G_k is the point $d_k^L = (X_k^U, Y_k^L)$ defined as $X_k^U \ge \max X_i$ and $Y_k^L \le \min Y_i, i \in J_k$. The equal operand in the previous relations holds when there is a real unit in G_k that has the group minimum or maximum values.

Likewise the upper bound $h_{j_0}^U$, a lower bound $h_{j_0}^L$ of the efficiency score can be obtained for the unit j_0. For this lower bound, the evaluated unit sets its position to the worst artificial unit $d_{j_0}^L$ while all other to their best artificial unit d_j^U.

Interval DEA can be formulated either as a CCR or a BCC model. In the paper, we use the BCC formulation but similar form can be given for the CCR. Models (1) and (2) estimate for a unit j_0 the values for the efficiency scores $h_{j_0}^U$, $h_{j_0}^L$, respectively.

$$\begin{aligned}
&\max h_{j_0}^U = \textstyle\sum_{r=1}^{s} u_r y_{rj_0}^U - w_0 \\
&s.t. \\
&\textstyle\sum_{i=1}^{m} v_i x_{ij_0}^L = 1 \\
&\textstyle\sum_{r=1}^{s} u_r y_{rj_0}^U - \sum_{i=1}^{m} v_i x_{ij_0}^L - w_0 \le 0 \\
&\textstyle\sum_{r=1}^{s} u_r y_{rj}^L - \sum_{i=1}^{m} v_i x_{ij}^U - w_0 \le 0, j = 1,..,n; j \ne j_0 \\
&u_r, v_i \ge \varepsilon \quad \forall r, i
\end{aligned} \tag{1}$$

$$\begin{aligned}
&\max h_{j_0}^L = \textstyle\sum_{r=1}^{s} u_r y_{rj_0}^L - w_0 \\
&s.t. \\
&\textstyle\sum_{i=1}^{m} v_i x_{ij_0}^U = 1 \\
&\textstyle\sum_{r=1}^{s} u_r y_{rj_0}^L - \sum_{i=1}^{m} v_i x_{ij}^U - w_0 \le 0, \\
&\textstyle\sum_{r=1}^{s} u_r y_{rj}^U - \sum_{i=1}^{m} v_i x_{ij}^L - w_0 \le 0, j = 1,...,n, j \ne j_0 \\
&u_r, v_i \ge \varepsilon \quad \forall r, i
\end{aligned} \tag{2}$$

Note that (1) and (2) are typical DEA BCC models with crisp data (the values for $x_{ij}^L, x_{ij}^U, y_{rj}^L, y_{rj}^U$ are known), under estimation variables of which are the weights v_i, u_r. Based on the efficiency values $h_{j_0}^U$, $h_{j_0}^L$, a simple classification for the units into three classes is possible as follows (see [8]):

$$\begin{aligned}
&E^{++} = \{j \in \{1,..,n\} / h_j^L = 1\}, \\
&E^{+} = \{j \in \{1,..,n\} / h_j^U = 1 \; and \; h_j^L < 1\}, \\
&E^{-} = \{j \in \{1,..,n\} / h_j^U < 1\}
\end{aligned}$$

The set E^{++} consists of the units that are efficient in 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. The set E^{-} consists of the definitely inefficient units. For them, any combination of values within their input/output intervals results in inefficiency.

IV. Dominance Relation between Groups

In this section, the notion of unit domination in DEA is extended to the case of the groups G_1, G_2 ..., G_p. A DEA unit is dominated if another unit of the date set has lower input and higher output values. Any such dominated unit is inefficient. This relation is formally defined as follows:

Definition. In the context of DEA, a unit $d_q = (X_q, Y_q)$ is dominated by a unit $d_p = (X_p, Y_p)$, (symbolically $d_p \succ d_q$), if $X_p \le X_q$ and $Y_p \ge Y_q$ or equivalently $(-X_p, Y_p) \ge (-X_q, Y_q)$. The inequality symbol $\ge$ is applied to all the elements of the vectors.

This dominance relation is now extended to the groups of units as follows: if the placement of a group (hyper-rectangle) is such that it is the best artificial unit, i.e. the one with lowest inputs and highest outputs (upper left corner in two dimensions) is worse, compared to any other group's worst artificial unit i.e. the one with highest inputs and lowest outputs (lower right corner in two dimensions), then the first group is dominated by the second. As every unit in a dominated group is also dominated, the clear conclusion is that it is inefficient. This means that every real unit belonging to the hyper-rectangle is inefficient. The formal definitions and proofs of the group domination relation are as follows:

Theorem 1. For two given groups $G_p = \{(X_k, Y_k), k \in J_p\}$ and $G_q = \{(X_q, Y_q), k \in J_q\}$, if $d_p^L \succ d_q^U$ then G_p dominates G_q ($G_p \succ G_q$).

Proof. It is sufficient to prove that every unit of G_p dominates every unit of G_q. Let $d_p = (X_p, Y_p)$ be a unit belonging to group G_p and $d_q = (X_q, Y_q)$ a unit belonging to group G_q. By the definition of the worst artificial unit, it is $d_p \succ d_p^L$. From the hypothesis, $d_p^L \succ d_q^U$, so it derives that $d_p \succ d_p^L \succ d_q^U$. Again by the definition of the best artificial unit it is $d_q^U \succ d_q$. From the last two relations follows that $d_p \succ d_p^L \succ d_q^U \succ d_q$. Thus, $G_p \succ G_q$, i.e. G_p dominates G_q. □

The dominated groups can be detected by simple comparisons of the elements of the artificial units d_k^L, d_k^U.

Note that the dominance relation is not sensitive to the returns to scale issue of DEA, so it cannot provide with efficiency classification information for the rest, undominated, units.

V. Exploitation of the IDEA and Dominance Relation-Numerical Example

The hyper-rectangle grid defined in the previous sections can be exploited in different ways. First, the conclusion for groups classified as E^- is that they contain inefficient units of the initial dataset. Such inefficient units can be excluded from the analysis as they do not affect the efficiency scores. In the case of large scale DEA problems, a sufficient reduction of the problem size is feasible when groups of E^- (and consequently all the units they include), may not participate in the analysis. The same outcome can be obtained by applying the dominance relation prior to the DEA modelling. Second, the grid is capable of predicting an efficiency score range for any new unit that enters the problem, without running the DEA models from the beginning. Indeed, prior of the DEA computational stage, using the values in inputs-outputs of a new unit, one can determine the group that this unit belongs. Then, by the IDEA concept, it is obvious that the efficiency score of this unit will be bounded by the efficiency values $h_{j_0}^U$, $h_{j_0}^L$ of the group that it belongs to. This means that if the group that the new unit belongs to is classified as E^{++}, the conclusion is that this unit is definitely efficient and if it is in the E^-, the unit will be definitely inefficient.

To illustrate the above, we provide the following simple arithmetic example: 34 units are assessed in terms of one input X and one output Y, the value ranges of which are [130 194], [50 90]. The breakpoints that we define for the input X are 130, 148, 170 and 200 and for the output Y 50, 72, 81, 90. Table II presents the nine groups (Groups G_1, G_2,..,G_9) formed by this segmentation, their population (column 2) and the low and upper efficiency scores and class (columns 6-8) obtained by IDEA models (1), (2).

TABLE II
Interval Segmentation and Efficiency

Group	Pop	x_{ij}^L	x_{ij}^U	y_{rj}^L	y_{rj}^U	$h_{j_0}^L$	$h_{j_0}^U$	Class
G1	3	130	148	50	70	0.493	0.978	E-
G2	5	148.1	169	50	70	0.432	0.858	E-
G3	3	169.1	194	50	70	0.376	0.752	E-
G4	4	130	148	70.1	81	0.691	1.000	E+
G5	4	148.1	169	70.1	81	0.603	0.993	E-
G6	5	169.1	194	70.1	81	0.527	0.870	E-
G7	3	130	148	81.1	89	0.884	1.000	E+
G8	3	148.1	169	81.1	89	0.702	1.000	E+
G9	9	169.1	194	81.1	89	0.614	0.956	E-

Fig. 1 presents graphically the coverage of the problem space by the parallelograms G_1,..,G_9 and the BCC efficient frontier in the initial dataset. The arrows indicate the dominance relation between the groups. In this case, all such relations are $G_4 \succ G_2$, $G_7 \succ G_5 \succ G_3$, $G_8 \succ G_6$, so all the dominated groups are G_2, G_3, G_5, G_6. To explain the dominance relation, take that of between G_4 and G_2 as example. G2 is dominated by group G4 because the worst case of G_4 (lower right corner point d_4^L) is better than the best case of G_2 (upper left corner d_2^U).

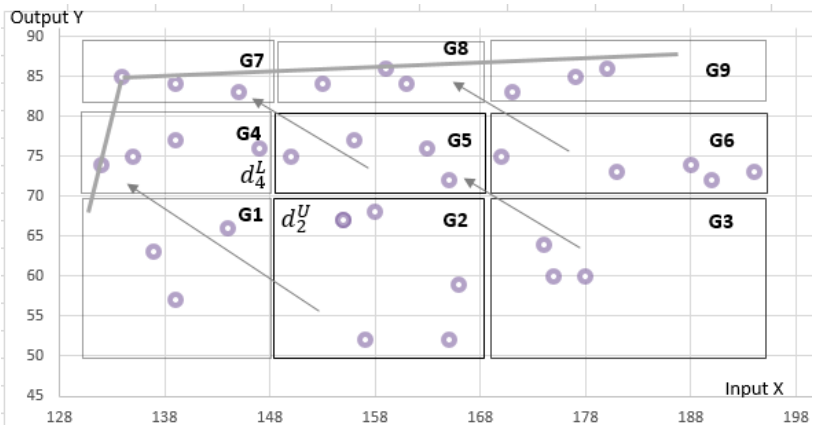

Fig. 1 The grid of the numerical example

Based on the above presented segmentation grid, the efficiency classification and prediction is possible. From the last column of Table II derives that groups G_1, G_2, G_3, G_5, G_6 and G_9 are classified as E, so they include inefficient units of the initial problem. Those groups contain 17 units of the initial DEA problem (50% of its size) and their exclusion from the analysis will result to a significant problem size reduction, without affecting the efficient units and their scores. The rest of the groups G_4, G_7 and G_8 in E^+, host parts the efficient frontier and may contain both efficient and inefficient units. Moreover, for any new unit that may enter the problem, the prediction of its efficiency status may be possible without solving the DEA problem from the beginning. For example, assume that a new unit (X,Y)=(158,75) will be assessed against all the existing in the initial dataset. From a direct comparison with the group breakpoints it is easy to detect that this unit belongs to group G_5 so the verdict is that this unit is inefficient and its score is bounded in [0,603 0,993], that of the group G_5 where it belongs to.

VI. Conclusion

In this paper, we presented a technique, acting as DEA preprocessor, for classifying and predicting efficiency. The technique is based on the formation of an interval dataset and the application of known Interval DEA models to access properly selected groups of the units. The efficiency status of the groups may be used to draw information about the efficiency of the units of the initial DEA problem. Moreover, the group structure presented can be used as an efficiency prediction template for units not included in the data set, assuming that their values for inputs and outputs do not violate the previously set ranges.

The presented technique is sensitive to the method which will be used to group the units of the problem. Its effectiveness depends on the number of groups and their population that will eventually be characterized as inefficient. A direction for the choice of the groups is to be such so to have relative large population and to be defined by interval values close to the maximum values for the inputs and to the minimum values for the outputs. Another indication for the inefficiency of the groups can be drawn by inspecting the interval dataset to find any existing dominance relations. An interesting outcome of the technique may also concern the class E^{++} which will include all the definitely efficient units. If the arrangement of the intervals and the groups is such to result to one or more classes E^{++}, then the identification of the efficient units will be possible, likewise the other DEA pre-processors.

Acknowledgement

This work has been partly supported by the University of Piraeus Research Center.

References

[1] A. Charnes, W. W. Cooper and Rhodes E. “Measuring the efficiency of decision making units”, European Journal of Operational Research 1978; 2; 429-444.
[2] J. H. Dulá, F. J. López. “Data envelopment analysis (DEA) in massive data sets”, Kluwer Academic Publishers; 2002; ISBN 1-4020-0489-3.
[3] Y. Chen, A. I. Ali, “Output-Input ratio analysis and DEA frontier”, Journal of Operational Research, 2002, 142:476-479.
[4] M. Shaheen, “A Pre-Processor for the CCR Model in DEA”, in INFORMS National Conference, Miami, FL. (Nov. 6, 2001).
[5] Ali I. A, “Streamlined computation for data envelopment”, European Journal of Operational Research, Volume 64, Issue 1, 8 January 1993, Pages 61-67.
[6] E. W. Forgy (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 21: 768–769.
[7] W. W. Cooper, K. S. Park and G. Yu. “IDEA and AR-IDEA: Models for dealing with imprecise data in DEA”, Management Science. 1999; 45; 597-607.
[8] D. K. Despotis, Y. G. Smirlis, “Data Envelopment with Imprecise Data”, European Journal of Operational Research 2002; 140; 24-36.
[9] Wang. Y-M. R. Greatbanks, Jian-Bo Yang. 2005. “Interval efficiency assessment using data envelopment analysis”. Fuzzy Sets and Systems 153:347–370.