Variables Selection Procedure for the DEA Overall Efficiency Assessment Based Plithogenic Sets and Mathematical Programming

Walid Abdelfattah

Department of Mathematics, College of Arts and Science, Northern Borders University, Rafha, Saudi Arabia.

Abstract

The appropriate selection of inputs and outputs also their count is a crucial step to achieve relevant results for any study involving the measurement of a set of decision-making units' overall efficiency using the data envelopment analysis methodology. In the literature, however, there is still no definitive standard to guide this selection. This article offers a novel two-phase procedure allowing, by solving an integer non-linear program in the first phase, to determine the most suitable number of both inputs and outputs to be used. And then, by following a developed plithogenic multi-criteria decision-making method in the second phase, to specify inputs and outputs that have the most contribution in improving efficiency and should be considered in the DEA model. A running example is applied throughout this article to make the proposed procedure more comprehensible. It consists in selecting variables for measuring the overall efficiency of the Northern Border University in Saudi Arabia. Obtained results consider at the same time the uncertainty aspect of data, also indeterminacy in experts' judgments towards inputs and outputs.

Keywords: Data envelopment analysis, overall efficiency, variable selection, plithogenic sets, universities' efficiency.

1. Introduction

Data envelopment analysis (DEA), first proposed by Charnes et al. (1978), has been used to measure the relative efficiency of a group of decision-making units (DMUs) regarding multiple inputs and multiple outputs. The original DEA essentially uses linear programming to find a best-practice frontier for efficient DMUs that envelops all other inefficient DMUs. It fundamentally assumes that inputs and outputs are measured with crisp, positive values on a ratio scale, and all the required data are available.

As its name indicates, DEA is highly sensitive to data; any noise or error in data measurement can easily cause non-applicable or insignificant results. Therefore, two keys that are important to the DEA's success involve (1) accurate measure of all factors, including inputs and outputs, (2) accurate selection of the variables to be included in the model. Regarded as a critical step routinely conducted before the DEA model is implemented, the second task is particularly complicated when the number of variables is greater than the number of DMUs. Decision-makers (DMs) may not be sure if some variables should be included in the model or not, so they generally pick inputs and outputs based on their prior knowledge or experience. It may lead to a miss specification of the DEA model.

In addition to the choice of variables that may strongly affect results, consideration of a non-suitable number of variables to be included could also yield unbiased results. In fact, when the number of variables is higher than what it should be, more DMUs could be classified as efficient, the DEA result will become then more selective. On the other hand, if the number of variables is limited, the implemented model would not be consistent to represent the real problem. An insightful discussion on the curse of dimensionality can be found in Fried et al. (2008).

In order to determine the relevant number of variables, Golany and Roll (1989) stated that the number of DMUs should be at least twice the number of the inputs and outputs being considered, i.e., $n \geq 2v$. where n is the number of DMUs and v is the number of variables. Bowlin (1998) suggested that each three DMUs

require one input and one output, i.e., $n \geq 3v$. To ensure a meaningful analysis, Dyson et al. (2001) stated that to achieve a reasonable level of discrimination, the minimum number of units should be twice the number of inputs and outputs, i.e., $n \geq 3(m + s)$, where m and s are the number of inputs and outputs, respectively. For a situation with three inputs and four outputs, Cooper et al. (2006) suggested $n \geq max\{m \times s, 3v\}$. Payrache et al. (2019) proposed to introduce cardinality constraints directly into the DEA program. The authors provided a rule of thumb for the choice of the maximal number of inputs and outputs which considers the convergence rate of the DEA estimator. The maximum number of variables that can be selected by the program should be determined as a function of the sample size.

Other researchers have focused on finding the most significant variables to be included in the DEA model rather than the optimal number of variables to be used. Banker (1996) lists three statistical tests to indicate the significance of an input or output variable to the production process. The null hypothesis is that the tested variable does not influence the production process. Pastor et al. (2002) considered two DEA formulations. The first formulation is different from the second by containing a testing variable called candidate. The impact of this variable on the efficiency is examined using a Binomial statistical test. Then a final decision is taken to add or eliminate the candidate variable. Ruggiero (2005) suggested a variable selection approach; An initial efficiency measure is attained from a set of known variables. Then efficiency is regressed against a set of selected variables. If these latter are shown to be relevant thus, they will be selected. The analysis is repeated until no further variable could be significant. Alongside these statistical methods, there are also methods based on an aggregate measure where a parametric aggregation function is specified, and then the inputs and outputs are grouped into indexes to be applied into the DEA model. In this context, the most popular approach is the DEA-PCA formulation proposed by Ueda and Hoshiai (1997) and Adler and Golany (2001) who used the principal components analysis (PCA) to decrease the number of inputs and outputs by substituting them with principal components. Other methods consist in analyzing the average change in efficiency scores by trying different model specifications (Norman and Stoker 1991, Valdmanis 1992, Sigala et al. 2004, and Wagner and Shimshak 2007). Nataraja and Johnson (2011) have discussed four approaches to guide variable specification in DEA; Efficiency Contribution Measure (ECM), Principal Component Analysis (PCA-DEA), a regression-based test, and bootstrapping for variable selection via Monte Carlo simulations. Li and al. (2016) proposed an approach to select the appropriate input and output variables set for evaluation via Akaike's information criteria (AIC) rule. This method mainly focuses on assessing the importance of a subset of original variables rather than testing the marginal role of variables one by one.

Even though there are several different methods related to the selection of the suitable number of inputs and outputs that a DEA model should contain depending on the study, there is still no consensus in the literature about the decision on which variables to retain. Payrache et al. (2019) stated that the DEA itself does not provide guidance for the requirement of the production function and the input and output variables; instead, they are left to the user's choice, judgment, and expertise. However, several issues may arise when selecting variables, e.g., the unavailability of data, high dimensional production processes, and the inclusion of irrelevant inputs or outputs.

Therefore, there has been a great motivation for the DEA model to select proper inputs and outputs and a representative number of these variables. This research aims to present a reliable procedure to identify the most relevant variables that should be retained in a DEA overall efficiency model. Hence, this article proposes a two-phase procedure to help the variables selection in DEA. The first phase is devoted to determining the most suitable number of both inputs and outputs to be used, an integer non-linear program is developed for this purpose. whereas the second phase is dedicated to identifying which inputs and outputs should be considered in the final DEA model. This second phase is based on a plithogenic multi-criteria decision-making method that considers variables as alternatives and takes into account experts' and DMs' judgments about the importance of these variables. the proposed procedure is applied to a real-world running example throughout this article. It concerns the selection of inputs and outputs for measuring the overall efficiency of the Northern Border University in Saudi Arabia.

Recently introduced by Smarandache (2017) as a generalization of neutrosophy, plithogeny is shown to be powerful in handling data uncertainty. Furthermore, this theory could treat indeterminacy in experts' judgments towards inputs and outputs since a plithogenic set is a set of attributes described by values and

characterized by two major features which are appurtenance degree and contradiction degree. These features ensure results accuracy, which is the main contribution of this research. For more details on plithogenic theory and sets, readers can refer to Smarandache (2017, 2018).

The remainder of this article is organized as follows. In the next section, we briefly review the concept of plithogenic sets and some basic operations are introduced. In section 3, the two phases of the proposed procedure are presented and explained using a real-world running example. Section 4 provides the conclusion and future directions.

2. Methods
This section comprises some necessary basic concepts that will be used later in this article.

2.1. Plithogenic set

Smarandache (2017) has defined Plithogeny as the genesis or origination, creation, formation, development, and evolution of new entities from dynamics and organic fusions of contradictory and/or neutrals and/or non-contradictory multiple old entities. While plithogenic means what is pertaining to plithogeny. Plithogenic set is the most relevant tool allowing the implementation of plithogeny concept. It is defined as the generalization of the crisp, fuzzy, and neutrosophic set Smarandache (2018).

Definition 2.1 [Smarandache, 2018]

A plithogenic set (P, A, V, d, c) is the set that contains number of elements defined by a set of attributes $A = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}, m \geq 1$, each attribute has a value $V = \{v_1, v_2, \ldots, v_n\}, n \geq 1$, a degree of appurtenance $d(x, v)$ of the element x, with respect to some given criteria, and finally a contradiction (dissimilarity) degree $c(v_j, v_D)$ between each attribute j and the dominant attribute D.

As can be noted from the previous definition, appurtenance and contradiction degrees are the main elements that characterize the plithogenic set. In fact, these degrees assist to provide more accurate results under uncertainty.

Definition 2.2 [Smarandache, 2018]

Let the cardinal $card(V) \geq 1$.

Let $c: V \times V \rightarrow [0, 1]$ be the attribute value contradiction degree function between any two attribute values v_1 and v_2, denoted by $c(v_1, v_2)$, and satisfying the following axioms:

- $c(v_1, v_1) = 0$; the contradiction degree between the same attribute values is null.
- $c(v_1, v_2) = c(v_2, v_1)$; commutativity.

Definition 2.3 [Smarandache, 2017]

Let $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ be two plithogenic sets:

Plithogenic intersection:

$$(a_{i1}, a_{i2}, a_{i3}) \wedge_p (b_{i1}, b_{i2}, b_{i3}) = \left(a_{i1} \wedge_F b_{i1}, \frac{1}{2}(a_{i2} \wedge_F b_{i2}) + \frac{1}{2}(a_{i2} \vee_F b_{i2}), a_{i3} \vee_F b_{i3} \right); \qquad (1)$$

Where:

$$a_{ij} \wedge_F b_{ij} = \left(1 - c(v_D, v_j)\right) . t_{norm}(v_D, v_j) + c(v_D, v_j) . t_{conorm}(v_D, v_j); \quad j = 1, \ldots, 3; \qquad (2)$$

$$a_{ij} \vee_F b_{ij} = \left(1 - c(v_D, v_j)\right) . t_{conorm}(v_D, v_j) + c(v_D, v_j) . t_{norm}(v_D, v_j); \quad j = 1, \ldots, 3; \qquad (3)$$

Where:

$$t_{norm} = a \wedge_F b = ab; \quad t_{conorm} = a \vee_F b = a + b - ab; \qquad (4)$$

3. The Proposed Variables Selection Procedure VSP
The proposed VSP is supposed to be executed through two successive connected phases; The first phase is dealing with the determination of the most suitable maximum number of inputs and outputs to not exceed in

a given DEA study, it includes four steps. While the second phase is related to identifying the appropriate inputs and outputs to be considered, it also suggests four steps.

To better clarify the VSP steps, a running example will be adopted in the rest of this paper. This example seeks to determine the most relevant inputs and outputs suitable to measure the overall efficiency of the Northern Border University (NBU). The choice of this example is motivated by the fact that in recent years, a large number of researchers have focused on measuring and comparing universities' efficiency. Efficiency analysis allows universities' policymakers to define policies and guidelines to improve their performances especially with the recent increase of the number of universities, students, and limited funds allocated to universities Witte and López-Torres (2017). This increase in the number of researches has obviously caused the emergence of a wide range of inputs and outputs used as variables in the different studies conducted. Among these works, readers can refer to (Kao and Hung 2008, Agasisti and Johnes 2009, Kuah and Wong 2011, Visbal-Cadavid et al. 2017, Duan SX 2019,…). Mojahedian et al. (2020) have extracted detailed lists of the most used inputs and outputs in the literature to determine the efficiency of universities.

NBU is a Saudi Arabian university located in the Northern Borders Department, city of Arar. It was founded in 2007. It contains 16 colleges: 9 in Arar, 4 in Rafha, 2 in Treef, and 1 in Ewaqueela. It has a variety of majors in multiple levels, including bachelor's and master's degrees (Table 1).

Tab 1. Northern Border University' Colleges

ARAR	
DMU 1	Faculty of Medicine
DMU 2	College of Business Administration
DMU 3	Faculty of Home Economics (Girls)
DMU 4	Faculty of Education and Arts
DMU 5	College of Nursing
DMU 6	College of Science
DMU 7	Faculty of Applied Medical Sciences
DMU 8	Faculty of Applied Science A
DMU 9	College of Engineering
RAFHA	
DMU 10	Faculty of Arts and Science R
DMU 11	Faculty of Applied Science R
DMU 12	Faculty of Computing and Information
DMU 13	Faculty of Pharmacy
TREEF	
DMU 14	Faculty of Arts and Science T
DMU 15	Faculty of Applied Science T
EWAQUEELA	
DMU 16	Faculty of Arts and Science E

3.1. Phase I: Determining the most suitable number of inputs and outputs

Step 1: Decomposition of the overall efficiency

Depending on the goal of the study that should have been clearly set, the type of efficiencies which will be considered should be as well carefully selected so that the appropriate overall efficiency is assessed. According to the application domain, the study perspectives, and the available data, overall efficiency could be clustered into many types.

By exploring publications and works that have been done on the measurement of universities' overall efficiency, two to three types have been generally considered, namely, operations efficiency, teaching efficiency, and research efficiency. In the considered example, the overall efficiency will be evaluated from these three perspectives.

Step 2: Pre-selection of inputs and outputs for the study

In this sensitive step, the operations researcher is asked to explore types of inputs and outputs that have been used with similar study specifications; these data could be found in one or many resources like databases, literature reviews, experts' opinions… Subsequently, inputs and outputs that best match the study will be then selected. Noting that, inputs and outputs whose data are not available could be removed, ex: financial data are often not available in many applications for confidential reasons.

Back to the running example, figure 1 displays the summary of detailed lists of most used inputs and outputs in the literature to determine the efficiency of universities. these lists are extracted by Mojahedian et al. (2020).
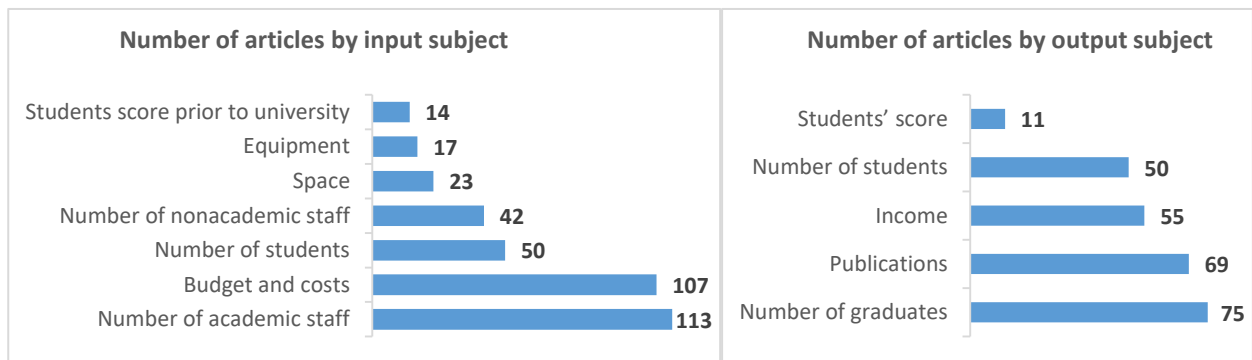

Fig 1. Number of articles by most used inputs and outputs in the efficiency of universities.

Step 3: Classifying the retained inputs and outputs

This step consists in associating each input and output retained in step 2 with its correspondent type of efficiency maintained in step 1. For the considered example, this step is summarized in table 2.

Tab 2. Classification of pre-selected inputs and outputs

Factors	Label	Subject
Operating inputs	Administrative staff	Number of nonacademic staff
	Technical support staff	Number of nonacademic staff
	College/Faculty area	Space
	Equipment	Equipment
Teaching Inputs	Academic staff	Number of academic staff
	Departments	Space
	Students	Number of students
	Available seats	Equipment
Research Inputs	Research staff	Number of academic staff / Budget
	Master programs	Number of available programs
	PhD programs	Number of available programs
	Funds	Budget and costs
Operating outputs	Operating projects	Building expansion projects
Teaching outputs	Graduated students	Number of graduates
Research outputs	Publications	Publications

Step 4: Determining the maximum number of inputs and outputs to be considered

Let consider c ($c = 1, \dots, C$; $C \geq 1$) components (Type of efficiency) of the overall efficiency identified in step 1.

Let w_c be the weight representing the importance of the component c according to the study objectives.

Let $\mathcal{I}$ and $\mathcal{O}$ be non-empty sets of all considered inputs and outputs in step 2, respectively.

$$\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_C\}; \mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_C\};$$

Where $\mathcal{I}_c$ and $\mathcal{O}_c$ are the non-empty sets of inputs and outputs, respectively, that belong to the component c, with the following properties:

- Each component should be covered by at least one input and one output:

$$card(\mathcal{I}_c) \geq 1; card(\mathcal{O}_c) \geq 1.$$

- No constraints on the intersection; When $\mathcal{I}_c \cap \mathcal{O}_c = \emptyset$ means that no input can act as an output and vice versa. When $\mathcal{I}_c \cap \mathcal{O}_c \neq \emptyset$ means that some inputs can act as outputs and vice versa.

- Some inputs/outputs could belong to more than one component.

To determine the appropriate number of inputs and outputs belonging to each component c, noted by x_c^* and y_c^*, respectively. This study proposes to use the following mathematical program (M1):

$$Max \sum_{c=1}^{C} w_c \left(\frac{card(I_c)}{card(I)} x_c + \frac{card(O_c)}{card(O)} y_c \right) \qquad (1.1)$$

$\quad$ Subject to: $\qquad$ (M1)

$$2 \sum_{c=1}^{C} (x_c + y_c) \leq n; \qquad (1.2)$$

$$x_c \leq card(I_c); \quad c = 1, \dots, C; \qquad (1.3)$$

$$y_c \leq card(O_c); \quad c = 1, \dots, C; \qquad (1.4)$$

$$x_c y_c \geq 1; \quad c = 1, \dots, C; \qquad (1.5)$$

$$\sum_{c=1}^{C} w_c = 1; \quad c = 1, \dots, C; \qquad (1.6)$$

$$x_c, y_c \text{ positive integers} \qquad (1.7)$$

This integer non-linear program (M1) can be interpreted as follow:

- The objective function (1.1) ensures not only to have the appropriate maximum number of inputs and outputs of each component that will be considered in the efficiency evaluation, but also ensures that these numbers are proportional to the total number of inputs/outputs retained in each component. In addition, the considered weights help to not obtain a multiple solution for the program and to consider priorities for the study purposes.
- The constraint (1.2) is in fact a common rule posited by Dyson et al. (2001) who stated that using significant numbers of inputs and outputs does not necessarily garner better results, the most important factor is the number of DMUs (noted by n in the constraint) should always be more than $2 \times (number\ of\ inputs + number\ of\ outputs)$. It is important to note that, depending on the study needs, this constraint could be replaced by any other common existing rules, Golany and Roll (1989), Bowlin (1998), Cooper et al. (2006), …etc.
- Constraints (1.3) and (1.4) are obvious for the guarantee that the number of inputs/outputs cannot exceed the total number of inputs/outputs considered in each component.
- The constraint (1.5) ensures two conditions; Each component should contain at least one input and one output, and, if an input of a given component is consumed, an output of the same component should be produced, and vice versa.
- The constraint (1.6) to remind that the sum of weights given to components do not exceed one.
- The last non-negativity constraint (1.7) is also obvious where the numbers of inputs and outputs should be positive integers.

By applying model (M1) on the considered example:

c = {1: "Operating efficiency", 2: "Teaching efficiency", 3: "Research efficiency"}

We assume $w_c = \{0.1, 0.5, 0.4\}$

$\mathcal{I} = \{\{\mathcal{I}_1\}: \text{"Operating inputs"}, \{\mathcal{I}_2\}: \text{"Teaching inputs"}, \{\mathcal{I}_3\}: \text{"Research inputs"}\}$

$\mathcal{I}_1 = \{\text{Administrative staff, Technical support staff, College/Faculty area in m}^2, \text{Equipment}\}$

$\mathcal{I}_2 = \{\text{Academic staff, Departments, Students, Available seats}\}$

$\mathcal{I}_3 = \{\text{Research staff, Master programs, PhD programs, Funds}\}$

$\mathcal{O} = \{\{\mathcal{O}_1\}: \text{"Operating outputs"}, \{\mathcal{O}_2\}: \text{"Teaching outputs"}, \{\mathcal{O}_3\}: \text{"Research outputs"}\}$

$\mathcal{O}_1 = \{\text{Operating projects}\}$

$\mathcal{O}_2 = \{\text{Graduated students}\}$

$\mathcal{O}_3 = \{\text{Publications}\}$

The optimal (maximum) number of inputs and outputs is obtained by using the Lingo 16.0 software as follows:

- Number of inputs related to the operations efficiency: $x_1^* = 1$;
- Number of inputs related to the teaching efficiency: $x_2^* = 3$;
- Number of inputs related to the research efficiency: $x_3^* = 1$;
- Number of inputs related to the operations efficiency: $y_1^* = 1$;
- Number of inputs related to the teaching efficiency: $y_2^* = 1$;
- Number of inputs related to the research efficiency: $y_3^* = 1$;

3.2. Phase II: Identifying the most relevant inputs and outputs for the study
Step 1: Incorporating the plithogenic importance weights of considered inputs and outputs

The importance weights of considered outputs are gained by a combination of interviews with the experts in problem field. These experts express their values judgments about the importance of the inputs and outputs by using the linguistic terms as in table 3. Next, based on expert's opinions, contradiction degrees are calculated between each attribute value with respect to the dominant attribute value in special, as well with respect to other attribute values.

Tab 3. Expressions for the importance weights of inputs and outputs

Expressions	Correspondent plithogenic number (T, I, F)
No Importance (NI)	(0.10, 0.70, 0.80)
Low Importance (LI)	(0.30, 0.40, 0.80)
Equal Importance	(0.50, 0.40, 0.60)
High Importance (HI)	(0.70, 0.30, 0.10)
Absolute Importance	(0.90, 0.10, 0.10)

Three experts are considered in the example; Expert 1 (E₁) represents the university operating deanship, Expert 2 (E₂) represents the university academic deanship, and Expert 3 (E₃) represents the university research deanship. Table 4 displays the importance expressions of inputs and outputs according to these experts.

Tab 4. Importance expressions of inputs and outputs according to experts

	i_{11}	i_{12}	i_{13}	i_{14}	i_{21}	i_{22}	i_{23}	i_{24}	i_{31}	i_{32}	i_{33}	i_{34}	o_{11}	o_{21}	o_{31}
Contradiction	0	0.	0.	0.	0	0.	0.	0.	0.	0	0	0.	0	0	0.
E_1	AI	HI	EI	HI	HI	EI	HI	LI	LI	LI	LI	NI	HI	EI	NI
E_2	AI	EI	HI	HI	AI	HI	AI	HI	EI	HI	HI	EI	AI	HI	HI
E_3	HI	HI	NI	HI	HI	EI	EI	NI	HI	AI	AI	HI	HI	AI	AI

Step 2: Aggregating experts' judgments

Based on tables 3 and 4, this step consists in finding decisions makers' weights using plithogenic aggregation given by equations (1, 2, 3, and 4).

Thus, aggregated weights $\omega_i = (T_i, I_i, F_i)$, for the considered inputs and outputs $i(1, \dots, 12)$, are obtained as follows:

$$T_i = (1 - c_i)[d_a(x, v_T) \wedge_F d_b(x, v_T)] + c_i[d_a(x, v_T) \vee_F d_b(x, v_T)]$$

$$I_i = \frac{1}{2}[d_a(x, v_I) \wedge_F d_b(x, v_I) + d_a(x, v_I) \vee_F d_b(x, v_I)]$$

$$F_i = (1 - c_i)[d_a(x, v_T) \vee_F d_b(x, v_T)] + c_i[d_a(x, v_T) \wedge_F d_b(x, v_T)]$$

Tab 5. Weights of considered inputs and outputs

Inp / Outp	c_i	Step 2				Step 3	
		E_1	E_2	E_3	Aggregated neutrosophic weights	Weights crisp values	Normalized weights
i_{11}	0	(0.90, 0.10, 0.10)	(0.90, 0.10, 0.10)	(0.70, 0.30, 0.10)	(0.57, 0.20, 0.27)	0.70	0.2628
i_{12}	0.5	(0.70, 0.30, 0.10)	(0.50, 0.40, 0.60)	(0.70, 0.30, 0.10)	(0.65, 0.33, 0.23)	0.70	0.2633
i_{13}	0.75	(0.50, 0.40, 0.60)	(0.70, 0.30, 0.10)	(0.10, 0.70, 0.80)	(0.58, 0.53, 0.33)	0.57	0.2162
i_{14}	0.25	(0.70, 0.30, 0.10)	(0.70, 0.30, 0.10)	(0.70, 0.30, 0.10)	(0.53, 0.30, 0.18)	0.69	0.2577
i_{21}	0	(0.70, 0.30, 0.10)	(0.90, 0.10, 0.10)	(0.70, 0.30, 0.10)	(0.44, 0.25, 0.27)	0.64	0.2776
i_{22}	0.5	(0.50, 0.40, 0.60)	(0.70, 0.30, 0.10)	(0.50, 0.40, 0.60)	(0.55, 0.38, 0.48)	0.57	0.2458
i_{23}	0.25	(0.70, 0.30, 0.10)	(0.90, 0.10, 0.10)	(0.50, 0.40, 0.60)	(0.48, 0.30, 0.52)	0.56	0.2411
i_{24}	0.75	(0.30, 0.40, 0.80)	(0.70, 0.30, 0.10)	(0.10, 0.70, 0.80)	(0.53, 0.53, 0.37)	0.54	0.2356
i_{31}	0.25	(0.30, 0.40, 0.80)	(0.50, 0.40, 0.60)	(0.70, 0.30, 0.10)	(0.34, 0.35, 0.64)	0.45	0.2555
i_{32}	0	(0.30, 0.40, 0.80)	(0.70, 0.30, 0.10)	(0.90, 0.10, 0.10)	(0.19, 0.23, 0.84)	0.38	0.2135
i_{33}	0	(0.30, 0.40, 0.80)	(0.70, 0.30, 0.10)	(0.90, 0.10, 0.10)	(0.19, 0.23, 0.84)	0.38	0.2135
i_{34}	0.5	(0.10, 0.70, 0.80)	(0.50, 0.40, 0.60)	(0.70, 0.30, 0.10)	(0.50, 0.43, 0.40)	0.56	0.3175
o_{11}	0	(0.70, 0.30, 0.10)	(0.90, 0.10, 0.10)	(0.70, 0.30, 0.10)	(0.44, 0.25, 0.27)	0.64	1.0000
o_{21}	0	(0.50, 0.40, 0.60)	(0.70, 0.30, 0.10)	(0.90, 0.10, 0.10)	(0.62, 0.23, 0.35)	0.68	1.0000
o_{31}	0.5	(0.10, 0.70, 0.80)	(0.70, 0.30, 0.10)	(0.90, 0.10, 0.10)	(0.78, 0.30, 0.16)	0.77	1.0000

In order to show how neutrosophic aggregated weights are obtained in table 5, the example of the first input i_{11} is explained below:

$(AI) \wedge_p (AI) \wedge_p (HI) = (0.90, 0.10, 0.10) \wedge_p (0.90, 0.10, 0.10) \wedge_p (0.70, 0.30, 0.10);$

$(0.90, 0.10, 0.10) \wedge_p (0.90, 0.10, 0.10) = (0.81, 0.10, 0.19)$

$(0.81, 0.10, 0.19) \wedge_p (0.70, 0.30, 0.10) = (0.57, 0.20, 0.27)$

Step 3: Compute crisp and normalized values of weights

This step consists in using equation (5) to convert the weights obtained from the plithogenic aggregation, that have been done in the previous step, into their corresponding crisp values.

$$\omega_i{}^{crisp} = (2 + T_i - I_i - F_i)/n; \tag{5}$$

Where n represents the number of experts or decision-makers.

Then, the normalization can be done by simply dividing each crisp value by the sum of weights corresponding to the same set of attributes, equation (6).

$$\left\| \omega_i{}^{crisp} \right\| = \omega_1{}^{crisp} / \sum_{i \in \mathcal{I}_c} \omega_i{}^{crisp}; \tag{6}$$

The last two columns in table 5 show these operations.

As an example, the crisp weight value of the first input is obtained as follows:

$$\omega_1{}^{crisp} = (2 + T_1 - I_1 - F_1)/3$$
$$= (2 + 0.57 - 0.20 - 0.27)/3 = 0.70$$

And then the normalized weight for this input is derived:

$$\left\| \omega_1{}^{crisp} \right\| = \omega_1{}^{crisp} / \sum_{i \in \mathcal{I}_1} \omega_i{}^{crisp}$$
$$= 0.70/(0.70 + 0.70 + 0.57 + 0.69) = 0.2628$$

Step 4: Selecting final inputs and outputs to be used in the assessment

Based on normalized aggregate weights obtained in step 3 (phase II), variables ranking could be easily determined corresponding to these weights' values for each component. Then, this ranking is used to finally select only the inputs and outputs that should be considered in the study according to the maximum number of variables found in step 4 (phase I) for each component. For the kth component $k = (1, 2, \dots, c)$, this could be written explicitly as follows:

Inputs that will selected are the elements of a non-empty set $\mathcal{I}_k{}^S$, such that:

$$\mathcal{I}_k{}^S = \{i_{1k}, i_{2k}, \dots, i_{sk}\} \subseteq \mathcal{I}_k = \{i_{1k}, i_{2k}, \dots, i_{tk}\}; 1 \leq s \leq t; s = x_k^*; \tag{7}$$

Where x_k^* is the optimal number (obtained in step 4 (phase I)) of inputs to be considered in the efficiency assessment phase.

With the same way, outputs that will selected are the elements of a non-empty set $\mathcal{O}_k{}^S$, such that:

$$\mathcal{O}_k{}^S = \{o_{1k}, o_{2k}, \dots, o_{sk}\} \subseteq \mathcal{O}_k = \{o_{1k}, o_{2k}, \dots, o_{tk}\}; 1 \leq s \leq t; s = y_k^*; \tag{8}$$

Where y_k^* is the optimal number (obtained in step 4 (phase I)) of outputs to be considered in the efficiency assessment phase.

Finally, the set of variables selected for the kth component could be represented by the following equation:

$$\mathcal{V}_k{}^S = \{\mathcal{I}_k{}^S, \mathcal{O}_k{}^S\}; \tag{9}$$

Note that in the case when variables from the same component give egalitarian weights, the selection could be done arbitrary. Table 6 shows the selected inputs and outputs for the considered example.

Tab 6. Selected inputs and outputs for each component

Factors	Max Number	Label	Weight	Rank	Selected
Operating inputs	1	Administrative	0.2628	2	-
		Technical support	0.2633	1	✓
		College/Faculty	0.2162	4	-
		Equipment	0.2577	3	-
Teaching Inputs	3	Academic staff	0.2776	1	✓
		Departments	0.2458	2	✓
		Students	0.2411	3	✓
		Available seats	0.2356	4	-
Research Inputs	1	Research staff	0.2555	2	-
		Master programs	0.2170	3	-
		PhD programs	0.2100	4	-
		Funds	0.3175	1	✓
Operating	1	Operating projects	1.0000	1	✓
Teaching	1	Graduated students	1.0000	1	✓
Research	1	Publications	1.0000	1	✓

For the operating efficiency:

$$\mathcal{V}_1^S = \{\mathcal{I}_1^S = \text{"Technical support staff"}, \mathcal{O}_1^S = \text{"Operating projects"}\}$$

For the teaching efficiency:

$$\mathcal{V}_2^S = \{\mathcal{I}_2^S = \text{"Academic staff, Departments, Students"}, \mathcal{O}_2^S = \text{"Graduated students"}\}$$

For the research efficiency:

$$\mathcal{V}_3^S = \{\mathcal{I}_3^S = \text{"Funds"}, \mathcal{O}_3^S = \text{"Publications"}\}$$

4. Discussion and Conclusion

The aim of this article is to suggest a solution for an important but unresolved problem in the DEA literature, namely the selection of inputs and outputs to be considered in the model. It proposes a two connected phases procedure that develops in the first phase a mathematical program allowing to determine the suitable number of variables to be included. Then, in the second phase, it offers a plithogenic multi-criteria decision-making method permitting to precisely identify, on the base of the first phase results, the most relevant variables that should be contained in the final DEA model. This proposed VSP counts some reliable advantages compared with previously proposed methods. In fact,

For the identification of the number of inputs and outputs that should be considered, the developed integer non-linear program is easy to solve and particularly flexible. This program allows to obtain not an exact number of variables that should be used, but instead the maximum number that should not be exceeded. It also permits the use of any of the thumb rules offered by the literature concerning the dimensionality of the study. Additionally, this program makes it possible to consider variables that play at the same time the role of input and output. Finally, it considers inputs and outputs that should represent each efficiency type when the overall efficiency is the matter.

On the other side, for the specification of which variable to be retained in the final DEA model, the second phase of the proposed VSP does not rely on any statistical tests compared with other works such as Banker (1996), Pastor et al. (2002), Ruggiero (2005). Moreover, it does not require any extra analysis methods such as in (Ueda and Hoshiai 1997, and Adler and Golany 2001) who use the principal components analysis (PCA) or as in (Norman and Stoker 1991, Valdmanis 1992, Sigala et al. 2004, and Wagner and Shimshak 2007) who require the analysis of the average change in efficiency scores.

As well, it takes into account experts' opinions and judgments regarding the inputs and outputs. This fact is not considered in Peyrache et al. (2019) and Nataraja and Johnson (2011). The need of incorporating value

judgment in real-life applications of DEA models is mainly due to the fact that Inputs and outputs are considered equally important in the DEA evaluation process. However, DMs usually have some preferences on the relative importance of the inputs and outputs. Especially, when the number of considered DMUs is small compared to the total number of inputs and outputs, classical DEA often identifies too many DMUs as efficient since it is then more likely that each DMU specializes in a specific input-output mix not directly comparable with that of the other DMUs. Hence, adding DMs' opinions on variables will certainly improve the differentiation between these DMUs in the DEA evaluation process.

Furthermore, using only plithogenic sets for this purpose could be considered novel. In fact, this concept is particularly used to consider indeterminacy and uncertainty of data and is generally associated with other methods to make multicriteria decision-making approaches; VIseKriterijumska Optimizacija I Kompromisno Resenje technique (VIKOR), Abdel-Basset et al. (2019), Analytic Hierarchy Process (AHP), Technique in Order of Preference by Similarity to Ideal Solution (TOPSIS), and VIKOR , Abdel-Basset et al. (2020), Multi-Attributive Ideal-Real Comparative Analysis (MAIRCA), Özçil et al. (2021). Consequently, it is shown in this article that plithogenic sets could stand by itself as a MCDM method.

Another main advantage that could be added to the proposed VSP is that it could applied to evaluate the overall efficiency of a set of DMUs and could be easily reduced to measure only one type of efficiency by considering only one component in the model (M1).

Finally, since in this article the proposed VSP is applied as a first attempt only to one real-life example, it would be more in-depth empirically analyzed in the future. As well as future directions, this procedure could be converted into a consistent algorithm making it more systematic from the data wrangling task to the desired results.

References

1- Abdel-Basseta M, El-hosenyb M, Gamala A, Smarandachec F (2019) A novel model for evaluation Hospital medical care systems based on plithogenic sets. Artificial Intelligence In Medicine. https://doi.org/10.1016/j.artmed.2019.101710.

2- Abdel-Basset M, Ding W, Mohamed R, Metawa N (2020) An integrated plithogenic MCDM approach for financial performance evaluation of manufacturing industries. Risk Management https://doi.org/10.1057/s41283-020-00061-4.

3- Adler N, Golany B (2001). Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to western Europe. European Journal of Operational Research 132(2):260–273.

4- Agasisti T, Johnes G (2009) Beyond frontiers: Comparing the efficiency of higher education decision-making units across more than one country. Educ Econ 17(1):59–79.

5- Banker R.D (1996). Hypothesis tests using data envelopment analysis. Journal of productivity analysis 7(2-3):139–159.

6- Bowlin W.F (1998) Measuring performance: an introduction to data envelopment analysis (DEA). J Cost Anal 15(2):3–27.

7- Charnes A, Cooper W.W, Rhodes E (1978) Measuring the efficiency of decision-making units. Eur J Opl Res 2:429–444.

8- Cooper W.W, Seiford L.M, Tone K (2006) Introduction to data envelopment analysis and its uses: with DEA-solver software and references. Springer Science & Business Media.

9- Duan SX (2019) Measuring university efficiency: An application of data envelopment analysis and strategic group analysis to Australian universities. Benchmarking: An International Journal, https://doi.org/10.1108/BIJ-10-2017-0274.

10- Dyson R.G, et al. (2001) Pitfalls and protocols in DEA. Eur J Oper Res 132(2):245–59.

11- Fried H.O, Lovell C.A.K, Schmidt S.S (2008) The Measurement of Productive Efficiency and Productivity Growth. Oxford University Press.

12- Golany B, Roll Y (1989) An application procedure for DEA. Omega 17(3):237–50.

13- Kao C, Hung HT (2008) Efficiency analysis of university departments: An empirical study. Omega 36(4):653–64.

14- Kuah CT, Wong KY (2011) Efficiency assessment of universities through data envelopment analysis. Procedia Computer Science (3):499–506.

15- Li Y, Shi X, Yang M, Liang L (2016) Variable selection in data envelopment analysis via Akaike's information criteria. Ann Oper Res. DOI 10.1007/s10479-016-2382-2.

16- Mojahedian MM, Mohammadi A, Abdollahi M, Kebriaeezadeh A, SharifzadehM, Asadzandi Sh, Nikfar Sh (2020) A review on inputs and outputs in determining the efficiency of universities of medical sciences by data envelopment analysis method. Med J Islam Repub Iran 34–42. https://doi.org/10.34171/mjiri.34.42.

17- Nataraja N.R, Johnson A.L (2011) Guidelines for using variable selection techniques in data envelopment analysis. European Journal of Operational Research 215(3):662–669.

18- Norman M, Stoker B (1991) Data Envelopment Analysis: The Assessment of Performance. John Wiley and Sons Chichester, England.

19- Özçil A, Tuş A, Öztaş GZ, Adalı EA, Öztaş T (2021) The Novel Integrated Model of Plithogenic Sets and MAIRCA Method for MCDM Problems. Proceedings of the INFUS Conference, Istanbul, Turkey, July 21–23, 2020. AISC 1197:733–741.

20- Pastor J.T, Ruiz J.L, Sirvent I (2002) A Statistical Test for Nested Radial DEA Models. Operations Research 50(4):728-735.

21- Peyrache A, Rose C, Sicilia G (2019) Variable selection in Data Envelopment Analysis. Eur J Opl Res. https://doi.org/10.1016/j.ejor.2019.09.028.

22- Ruggiero, J (2005) Impact Assessment of Input Omission on DEA. International Journal of Information Technology & Decision Making 04(03):359–368.

23- Sigala M, Airey D, Jones P, Lockwood A (2004) ICT Paradox Lost? A Stepwise DEA Methodology to Evaluate Technology Investments in Tourism Settings. Journal of Travel Research 43(2):180–192.

24- Smarandache F (2017) Plithogeny, Plithogenic Set, Logic, Probobility, and Statistics. Pons: Brussels, Belgium.

25- Smarandache F (2018) Plithogenic set, an extension of crisp, fuzzy, intuitionistic fuzzy, and neutrosophic sets-revisited. Neutrosophic Sets Syst (21):153–166.

26- Ueda T, Hoshiai Y (1997) Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. Journal of the Operations Research Society of Japan 40(4):466–478.

27- Valdmanis V (1992). Sensitivity Analysis for DEA Models: An Empirical Example using Public vs. NFP Hospitals. Journal of Public Economics 48(2):185–205.

28- Visbal-Cadavid D, Martínez-Gómez M, Guijarro F (2017) Assessing the efficiency of public universities through DEA: A case study. Sustainability 9(8),1416. https://doi.org/10.3390/su9081416.

29- Wagner J.M, Shimshak D.G (2007) Stepwise Selection of Variables in Data Envelopment Analysis: Procedures and Managerial Perspectives. European Journal of Operational Research 180(1):57-67.

30- Witte KD, López-Torres L (2017) Efficiency in education: a review of literature and a way forward. J Oper Res Soc 68(4):339–63.