J. Basic. Appl. Sci. Res., 2(12)12031-12043, 2012

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Proposing a Mixed Model Based on Stochastic Data Envelopment Analysis and Principal Component Analysis to Predict Efficiency

Ali Yaghoubi[1], Mehdi Bashiri[2]

[1] Department of Industrial Engineering, Tehran Payame Noor University, Tehran, Iran.
[2] Department of Industrial Engineering, Shahed University, Tehran, Iran.

ABSTRACT

Data Envelopment Analysis is a technique based on linear programming methods to construct surface or non parametric boundary over the data. This boundary is used to evaluate proportional efficiency. In this paper, a mixed model was proposed based on Stochastic Data Envelopment Analysis (SDEA) and Principal Component Analysis (PCA) for predicting the similar units' efficiencies in an organization. It was tried to abate the most important shortcoming of DEA, involving: being unable to estimate the efficiency, the unreal distribution of weights of inputs and outputs of the model and variety in efficient branches. The following model covered the mentioned problems caused by entering the stochastic effect, and considering the effects of fuzzy weights on the inputs and outputs by the use of SDEA technique with fuzzy weights. PCA was used to determine efficiency mean for units with various risks. Finally, in order to reach a better understanding of the proposed model, it was applied to predict efficiencies for some Iranian Bank branches. The high correlation between real and predicted efficiencies was obtained which represented the validity of the proposed SDEA/PCA model.

KEYWORDS: Stochastic Data Envelopment Analysis, Principal Component Analysis, Efficiency, Decision making Unit, Fuzzy theory.

1. INTRODUCTION

Evaluating and comparing the performance of similar units of an organization is an important part of the responsibilities of organization management. One of the most important tools of relative performance comparing these units is a quantitative, precise and powerful approach called Data Envelopment Analysis (DEA). This technique is considered not only in performance evaluation but also in management; it is more precisely recognized in units under control. This method also has some major shortcomings the most important of which are impossibility of predicting efficiency, inability in determining acceptable risk level for the managers in the direction of achieving the predicted efficiencies in each unit, and unreal weight distribution to the inputs and outputs. For preparing the possibility of predicting efficiency and the level of its dependent risk, one can benefit from a mathematical model which is based on Stochastic Data Envelopment Analysis SDEA by entering stochastic effects, environmental factors like economic condition on the inputs and outputs of the units under control. Also for solving the problem of lack of unreal distribution of weights to inputs and outputs of the DEA model, we used fuzzy weights according to experts' suggestions for the weights of the model inputs and outputs. On the other hand, in this paper a mixed model was proposed based on Principal Component Analysis (PCA) and Stochastic Data Envelopment Analysis (SDEA) in order to predict efficiency mean for units with various risk levels.

Stochastic constraints programming is a very important and useful method in stochastic programming. Charnes and Cooper (1989) entered the chance constrained programming in research operation literature for the first time [4]. They, along with Rouds (1978), discussed data envelopment analysis for calculating efficiency. Sengupta et al. (1982) proposed the stochastic DEA models. In other words, these studies combined the models of data envelopment analysis with chance constraint programming (CCP) and used the obtained stochastic models for estimating efficiency and considering the measuring errors of input variables. The studies conducted on the weights of inputs and outputs in DEA are limited and the most important of them are the articles of Dayson and Thanassoulis (1988), Charnes et al. (1989), Roll and Golany (1993) and Jahanshahloo et al. (1997).

Land et al. (1993) proposed a model known as LLT. In this model, they considered both constraints of the envelopment form of CCR model as the stochastic variables. After proposing LLT model, Cooper et al. (1996) proposed a new model by applying Saimon Satisfactory model. This new model is a combination of the concept of satisfactory decision making with CCDEA models or data envelopment analysis with stochastic constraints. Jackson (2001) estimated the efficiency in free market using data envelopment analysis. Cooper et al. (2002) proceeded to

***Corresponding Author:** Uones Ghasembaklo, Department of Clinical Psychology, Payam-e-Noor University of Khoy.
Email: gasembaklo@gmail.com

the analysis of technical efficiency using stochastic constraints programming approach. Saati et al. (2003) presented a method for obtaining a common set of fuzzy inputs and outputs weights. They first suggested their model for deterministic data and then developed it for fuzzy data. Houang et al. (2005) proposed the combined model of SDEA and chance constraint programming. Cooper et al. (2006) presented the last proposed model in the SDEA ground. In this paper, they proposed the output-based BCC random model and improved it by applying the stated concepts in BCC model and the hypothesis of random inputs and outputs and normal distribution for them.

Adler et al. (2003) replaced parameters with input/output oriented using principal component analyzing (PCA). Bruce et al (2008) used DEA and PCA techniques for performance evaluation in internet bank industry using an approach similar to Cinca's approach. Shanmugam et al (2010) proposed a model at a crossroad of data envelopment and Principal component analysis to units ranking. Kao et al. (2011) proposed a two-stage approach of integrating independent component analysis (ICA) and data envelopment analysis (DEA) to overcome discrimination between efficient and inefficient decision-making units in the DEA.

The rest of this paper is organized as follows: a brief description is given about Stochastic DEA model and fuzzy set theory in section 2. Formulation proposed SDEA/PCA model is described in details in section 3. A practical example of application derived from this empirical study is documented in section 4. Discussion and future works are summarized in the last section (section 5).

3. Stochastic DEA and Fuzzy set theory:

3.1 Stochastic DEA introduction

Data Envelopment Analysis (DEA) assumes that there are n DMUs (Decision Making Unit, DMU) whose whole set is denoted by j (j=1,2,…,n). The performance of each DMU is characterized by its production process of m inputs (Xij for i=1,2,…,m) to yield s outputs (Yrj for r=1,2,…,s). It is also assumed that all DMUs have input and output vectors and all the components of these vectors are positive.

DEA model:

$$\begin{aligned} Max \quad & E_k = \sum_{r=1}^{s} u_r \, y_{rk} \\ st: & \sum_{i=1}^{m} v_i \, x_{ik} = 1 \\ & \frac{\sum_{r=1}^{s} u_r \, y_{rj}}{\sum_{i=1}^{m} v_i \, x_{ij}} \le 1 \quad j = 1 \cdots n \\ & u_r, v_i \ge o \end{aligned} \qquad (1)$$

SDEA model:

$$\begin{aligned} Max \quad & E(E_k = \sum_{r=1}^{s} u_r \, \hat{y}_{rk}) \\ st: & \sum_{i=1}^{m} v_i \, x_{ik} = 1 \\ & P_r[\frac{\sum_{r=1}^{s} u_r \, \hat{y}_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \le \beta_j] \ge 1-\alpha_j \quad j = 1 \cdots n \\ & u_r, v_i \ge o \end{aligned} \qquad (2)$$

Where the above two models are designed to measure the performance (DEA efficiency) of the specific k-th DMU (in j) as Ek. The symbols (Vi and Ur) represent weight multipliers related to the i-th input and the r-th output, respectively. In Eq. (2), Pr stands for a probability and the superscript "^" presents that $\hat{y}_{rj}$ is a stochastic variable.

It is important to mention that this study is interested in future planning where the quantity of inputs can be controlled as decision variables, whilst being unable to control outputs, because these quantities depend upon external factors such as an economic condition. Hence, the inputs are considered as deterministic variables and the outputs are considered as stochastic variables. To describe the analytical structure of our SDEA model, it is compared with a traditional DEA model, often referred to as "DEA ratio form". Mathematically, the two models have the following formulations:

Model (1) is formulated under the condition that each DMU is evaluated by a ratio of its total weighted outputs to total weighted inputs. The original DEA model determine the ratio of all DMUs to be less than or equal to unity. Consequently, it belongs to an efficiency range between 0 and 100%. Meanwhile, Eq. (2) formulates the ratio to be less than or equal to βj (a prescribed value in the range between 0 and 100%) that represents an expected efficiency level of the j-th DMU. Cooper et al. (2006) consider the expected efficiency score as an "aspiration level" that is usually requested by an outside authority and/or a budgetary limitation. Since βj is set to be unity in Eq. (1), the deterministic model (1) can be considered as a special case of the SDEA model (2).

The other symbol αj stands for the probability that output/input ratio becomes more than βj with a choice of weight multipliers. Thus, αj is considered as a risk criterion representing utility of a manager. On the other hand, 1-αj shows the probability of attaining the requirement. Like βj, the risk criterion (αj) is also a described value that is

measured in the range between 0 and 1. When αj =0 in Eq. (2), it is certainly required that the out-put/input ratio becomes less than or equal to βj. Conversely, αj=1 omits the requirement under any selection of weight multipliers. The objective of Eq. (1) is formulated by $\sum_{r=1} u_r y_{rk}$ while that of Eq. (2) is expressed by $E(\sum_{r=1} u_r \hat{y}_{rk})$, where the symbol "E" stands for an expected value of the sum of weighted $\hat{y}_{rk}$.

3.2 Fuzzy sets theory

Theory of fuzzy sets is quite similar to man's attitude when facing uncertainties to express inaccurate words, such as ''approximately'', ''very'', ''nearly'', etc. as well as for consistency with subjective judgments of different people due to various interpretations from a subject. Zadeh (1965) introduced fuzzy sets theory for the first time, expressing it in the issue of decision-making. In fuzzy sets, membership degree of an element is between 0 and 1, while in classic sets, there are two states: an element with the degree 1 is inside the set, or it is not with degree 0. In order to elaborate on the said matter, consider the discussion in this paper, in which MC-ABC inventory classification is carried out using inventory managers' subjective judgments and introducing fuzzy concepts of prioritizing the criteria. To achieve these ends, fuzzy set, fuzzy numbers and linguistic variables should first be introduced (Chen, 2007).

Definition 2.1. A fuzzy set $\tilde{A}$ in a universe of discourse X is defined by a membership function $\mu_{\tilde{A}}(X)$ which associates $\forall x \in X$ a real number in the interval [0,1]. express $\mu_{\tilde{A}}(X)$ membership degree of x in $\tilde{A}$.

Definition 2.2. The α-cut of fuzzy set $\tilde{A}$ is a crisp set $\tilde{A}_\alpha = \{x | \mu_{\tilde{A}}(x) \geq \alpha\}$. The support $\tilde{A}$ is the crisp set $Supp(\tilde{A}) = \{x \| \mu_{\tilde{A}}(x) \geq 0\}$. $\tilde{A}$ is normal if and only if $Supp_{x \in X} \mu_{\tilde{A}}(x) = 1$.

Definition 2.3. A fuzzy subset $\tilde{A}$ of universe set X is convex if and only if $\mu_{\tilde{A}}(\lambda x + (1-\lambda) y) \geq \min(\mu_{\tilde{A}}(x), \mu_{\tilde{A}}(y))$, $\forall x, y \in X, \lambda \in [0,1]$, where min denotes the minimum operator.

Definition 2.4. $\tilde{A}$ is a fuzzy number if and only if $\tilde{A}$ is normal and convex fuzzy set of X.

Definition 2.5. A triangular fuzzy number $\tilde{A}$ is defined with piecewise linear membership function $\mu_{\tilde{A}}(x)$ as follow:

$$\mu_{\tilde{A}}(x) = \begin{cases} \dfrac{x - a_1}{a_2 - a_1} & a_1 \leq x \leq a_2 \\ \dfrac{a_3 - x}{a_3 - a_2} & a_2 \leq x \leq a_3 \\ 0 & otherwise \end{cases} \quad (3)$$

And as a triplet (a1 , a2 , a3) is indicated, where a1, a3 the lower and upper bounds respectively, and a2 is the most likely value of $\tilde{A}$ (Shavandi, 2005).

4. Formulation Proposed SDEA/PCA model

4.1 Reformulation Stochastic DEA model

In this study, the constraints and objective of Eq. (2) are reformulated by CCP proposed by Cooper (2002). (Research by Cooper et al. (2006) shows how to incorporate the CCP technique into the DEA ratio form. In the SDEA models of these papers, both inputs and outputs are stochastic variables. Hence, our formulation presented in this study can be considered as a special case of their SDEA).

The constraints of Eq. (2), including the stochastic process, can be rewritten as follows:

$$P_r \left\{ \sum_{r=1}^{s} u_r \hat{y}_{rj} \leq \beta_j \left(\sum_{i=1}^{m} v_i x_{ij} \right) \right\} \geq 1 - \alpha_j \quad (4)$$

Eq. (4) is equivalent to:

$$P_r\left\{\frac{\sum_{r=1}^{s} u_r\,(\hat{y}_{rj}-\bar{y}_{rj})}{\sqrt{V_j}} \le \frac{\beta_j \sum_{i=1}^{m} v_i\,x_{ij}-\sum_{r=1}^{s} u_r\,\bar{y}_{rj}}{\sqrt{V_j}}\right\} \ge 1-\alpha_j \tag{5}$$

where $\bar{y}_{rj}$ is the expected value of $\hat{y}_{rj}$ and :

$$V_j=(u_1 \quad u_2 \quad \dots \quad u_s)\times\begin{pmatrix} v(\hat{y}_{1j}) & \operatorname{cov}(\hat{y}_{1j},\hat{y}_{2j}) & \dots & \operatorname{cov}(\hat{y}_{1j},\hat{y}_{sj}) \\ \operatorname{cov}(\hat{y}_{2j},\hat{y}_{1j}) & v(\hat{y}_{2j}) & \dots & \operatorname{cov}(\hat{y}_{2j},\hat{y}_{sj}) \\ \vdots & & & \vdots \\ \operatorname{cov}(\hat{y}_{sj},\hat{y}_{1j}) & \dots & \dots & v(\hat{y}_{sj}) \end{pmatrix}\times\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_s \end{pmatrix} \tag{6}$$

Here, Vj indicates the variance-covariance matrix of the j-th DMU in which the symbol "v" stands for a variance and the symbol "cov" refers to a covariance operator. To reformulate Eq. (5) by CCP, this study introduces the following new variable ($\hat{Z}_j$):

$$\hat{Z}_j=\frac{\sum_{r=1}^{s} u_r\,(\hat{y}_{rj}-\bar{y}_{rj})}{\sqrt{V_j}} \qquad j=1,\dots,n \tag{7}$$

which follows the standard normal distribution with zero mean and unit variance. Substitution of Eq. (7) in Eq. (5) produces:

$$P_r\left\{\hat{Z}_j \le \frac{\beta_j \sum_{i=1}^{m} v_i\,x_{ij}-\sum_{r=1}^{s} u_r\,\bar{y}_{rj}}{\sqrt{V_j}}\right\} \ge 1-\alpha_j \tag{8}$$

Since follows the standard normal distribution, the invariability of Eq. (8) is executed as follows:

$$\frac{\beta_j \sum_{i=1}^{m} v_i\,x_{ij}-\sum_{r=1}^{s} u_r\,\bar{y}_{rj}}{\sqrt{V_j}} \ge Z^{-1}(1-\alpha_j)\ ,\quad j=1,\dots,n \tag{9}$$

Here, Z stands for a cumulative distribution function of the normal distribution and Z^{-1} indicates its inverse function. The SDEA model (2) is obtained by replacing Eq. (4) by Eq. (9) and its resulting formulation becomes:

$$\begin{aligned} &Max\ E_K = E\Big(\sum_{r=1}^{s} u_r\,\hat{y}_{rk}\Big) \\ &st:\ \sum_{i=1}^{m} v_i\,x_{ik}=1 \\ &\beta_j \sum_{i=1}^{m} v_i\,x_{ij}-\sum_{r=1}^{s} u_r\,\bar{y}_{rj} \ge \sqrt{V_j}\,Z^{-1}(1-\alpha_j)\ ,\quad j=1,\dots,n \\ &u_r \ge o\ ,\ r=1,\dots,s\ ,\ v_i \ge o\ ,\ i=1,\dots,m \end{aligned} \tag{10}$$

This research assumes that a stochastic variable ($\hat{y}_{rj}$) of each output is expressed by:

$$\hat{y}_{rj} = \bar{y}_{rj} \pm b_{rj}\,\varsigma \quad , r = 1,...,s \quad , \quad j = 1,...,n$$

Where is an $\bar{y}_{rj}$ expected value of $\hat{y}_{rj}$ and brj is it's standard deviation. (Section 5 of this paper describes how to determine the average and the standard deviation from his/her prediction of a decision maker(s). Cooper et al. (2006) proposed the assumption along with a practical rationale.) It is also assumed that a single random variable (ς) follows a normal distribution $N(0,\sigma^2)$. Under such an assumption, Vj becomes:

$$V_j = (u_1 \quad u_2 \quad ... \quad u_s) \times \begin{pmatrix} b_{1j}^2\sigma^2 & b_{1j}b_{2j}\sigma^2 & \dots & b_{1j}b_{sj}\sigma^2 \\ b_{2j}b_{1j}\sigma^2 & b_{2j}^2\sigma^2 & \dots & b_{2j}b_{sj}\sigma^2 \\ \vdots & & & \vdots \\ b_{sj}b_{1j}\sigma^2 & \dots & \dots & b_{sj}^2\sigma^2 \end{pmatrix} \times \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_s \end{pmatrix} \tag{11}$$

Incorporation of Eq. (11) into Eq. (10) provides:

$$\begin{aligned} & Max\, E(\sum_{r=1}^{s} u_r\, \hat{y}_{rk}) \\ & st: \sum_{i=1}^{m} v_i\, x_{ik} = 1 \\ & \quad \beta_j \sum_{i=1}^{m} v_i\, x_{ij} - \sum_{r=1}^{s} u_r\, \bar{y}_{rj} \geq (\sum_{r=1}^{s} u_r\, b_{rj}\sigma)\, Z^{-1}(1-\alpha_j) \;,\; j = 1,..,n \\ & \quad u_r \geq o\,,\; r = 1,...,\; s \;,\; v_i \geq o \;,\; i = 1,...,\; m \end{aligned} \tag{12}$$

Next, paying attention to $\hat{y}_{rj} = \bar{y}_{rj} \pm b_{rj}$ we reformulate the objective of Eq. (12) as follows:

$$Max: E(\sum_{r=1}^{s} u_r\, \hat{y}_{rk}) = E(\sum_{r=1}^{s} u_r(\bar{y}_{rj} \pm b_{rj}\varsigma)) = E(\sum_{r=1}^{s} u_r\, \bar{y}_{rk} \pm \sum_{r=1}^{s} u_r b_{rk}\varsigma) = \sum_{r=1}^{s} u_r\, \bar{y}_{rk} \tag{13}$$

It is assumed that the random variable (ς) follows a normal distribution N(0,1) in Eq. (13). Under such an assumption (so,σ=1), and because of E(ς)=0, consequently the SDEA model can be written in the following model:

$$\begin{aligned} Max \quad & E_k = \sum_{r=1}^{s} U_r\, \bar{y}_{rk} \\ & st: \sum_{i=1}^{m} v_i\, x_{ik} = 1 \\ & \quad \sum_{i=1}^{m} v_i(\beta_j\, x_{ij}) - \sum_{r=1}^{s} u_r\{\bar{y}_{rj} + b_{rj}\, Z^{-1}(1-\alpha_j)\} \geq o \quad , \quad j = 1,..,n \\ & \quad u_r \geq o\,,\; r = 1,...,\; s \;,\; v_i \geq o \;,\; i = 1,...,\; m \end{aligned}$$

4.2 Input/output weights determination with using fuzzy approach

Assuming $D = \{D_1, D_2, ..., D_r\}; r \geq 2$ be a group of r decision makers expressing r reciprocal judgment matrixes $\{R_{n\times n}^{(K)}; k = 1,...., r\}$ corresponding to pair wise comparisons for a set of n criterions $A = \{A_1, A_2, ..., A_n\}$, where $R_{n\times n}^{(K)} = \left(r_{ij}^{(K)}\right)$ is a positive squared matrix which validates $r_{ii}^{(K)} = 1, \quad r_{ji}^{(K)} = \frac{1}{r_{ij}^{(K)}} \succ 0$ for $i \neq j$.

The judgments represent the relative importance to the decision maker Dk of Ai compared to Aj. The comparison matrix given by the k-th decision maker is denoted as follows:

$$R_{n\times n}^{(K)} = \begin{pmatrix} 1 & \cdots & r_{1n}^{k} \\ \vdots & r_{ij}^{k} & \vdots \\ r_{n1}^{k} & \cdots & 1 \end{pmatrix} \tag{15}$$

With normalize the comparison matrix and using row geometric mean method, the weights of criterions are obtained (Saati et al., 2003).

In some study for determination importance of weights in DEA models, used group analytical hierarchy process (GAHP) technique. In order to using this technique with considering experts suggestions, first the inputs and outputs weights obtain with using questionnaire of even comparing and its distribution among individual of experts in order to know the subjective preferences of experts about the amount of the weights of each output and finally the preferred weight of each output was obtained by using eq. 15 as following. Then with considering this assumption that minimum and maximum of weights equal zero and double $(0, w_i, 2w_i)$. of weights respectively, we can show the weights in model as fuzzy number

If fuzzy weights put in model, the efficiency amount will obtain with considering experts suggestions for each units. But the solution of model may be infeasible by putting fuzzy weights in it, so we should define assurance region for obtained weights. Its better first, we define the broad assurance region for weights to prevent of infeasible solution that this region can be change with parameter α-cut. In other words, whatever α-cut close to 1, the experts' suggestions being applied more accurately and whatever α-cut close to 0, the obtained efficiency will be inaccurate. For example the assurance region for i-th input (Vi) with parameter α-cut as follows:

$$w_i - w_i(1-\alpha) \le v_i \le w_i + w_i(1-\alpha) \tag{16}$$

Eq. (16) can be rewritten as follows:

$$\begin{aligned} & v_i \ge w_i \alpha \\ & v_i \le 2w_i - w_i \alpha \end{aligned} \tag{17}$$

With similar process, the assurance region for r-th output (Ur) as follows:

$$\begin{aligned} & u_r \ge w'_r \alpha \\ & u_r \le 2w'_r - w'_r \alpha \end{aligned} \tag{18}$$

The other reasons for using fuzzy weights in efficiency evaluation are increasing model flexibility and decreasing probability of model's infeasibility. Now with adding above fuzzy constraints to eq. 14, fuzzy SDEA model defines as follows:

$$\begin{aligned} & Max \sum_{r=1}^{s} u_r \, \bar{y}_{rk} \\ & st: \sum_{i=1}^{m} v_i \, x_{ik} = 1 \\ & \sum_{i=1}^{m} v_i(\beta_j \, x_{ij}) - \sum_{r=1}^{s} u_r\{\bar{y}_{rj} + b_{rj}\sigma Z^{-1}(1-\alpha_j)\} \ge o \quad , \quad j = 1,..,n \\ & v_i \ge w_1.\alpha \\ & v_i \le 2w_i - w_i.\alpha \\ & u_r \ge w'_r \alpha \\ & u_r \le 2w'_r - w'_r \alpha \\ & u_r \ge o, \; r = 1,...,s \;, v_i \ge o \;, i = 1,...,m \end{aligned} \tag{19}$$

For solving the above model first we consider $\alpha = 1$, if the solution is infeasible at least for one unit, we decrease amount of α and solve it again. The advantage of this work is increasing range of parameter's weights selection but the flaw is decreasing accuracy of experts' suggestions (Lai Young et al, 1992).

4.3 Estimation of Output

To determine $\bar{y}_{rj}$ and b_{rj}^2 of $\hat{y}_{rj}$ this study utilizes three different kinds of output estimate. A decision maker(s), who is involved in future planning, is asked to forecast the following three estimates on each output of the j-th DMU: 1) the most likely estimate (MLrj), 2) the optimistic estimate (OPrj), and 3) the pessimistic estimate (PErj). The ML is the most realistic estimate of $\hat{y}_{rj}$. From a statistical viewpoint, it is considered as the mode (the highest point) of the probability distribution for each output. The OP is aimed to be the unlikely but possible output quantity if everything goes well. It can be seen as an estimate of the upper bound of the probability distribution. The PE is intended to be the unlikely but possible output quantity if everything goes wrong. It is an estimate of the lower bound of the probability distribution.

Assuming that the data follows the beta probability distribution, this study converts the three estimates into the expected value and variance of each out- put. The expected value of its distribution is approximately:

$$\bar{y}_{rj} = (OP_{rj} + 4\,ML_{rj} + PE_{rj})/6 \tag{20}$$

The variance becomes:

$$b^2_{rj} = (OP_{rj} - PE_{rj})^2/36 \tag{21}$$

Where MLrj is a mode and ((OPrj + PErj)/2) shows a midrange between OPrj and PErj the expected value can be seen as a weighted arithmetic mean of the mode and the midrange. The mode has two-thirds of the entire weight. It is important to note that the above type of estimation is widely used in PERT/CPM (Program Evaluation and Review Technique/Critical Path Method). PERT/CPM is a management science technique for planning activity times and scheduling, while this study uses the technique to estimate the expected value and variance of each output. Using the proposed approach, future uncertainty regarding each output, which may fluctuate due to many economic factors, can be incorporated into our DEA formulation (Sabzehparvar, 2002).

4.4 SDEA/PCA model:

Suppose we have n independent homogeneous decision making units, that the purpose is evaluating and ranking units based on obtained efficiency (E_{lj}), (j = 1, 2,. . . ,n) and various risk levels (α_l) for them with Eq. (19) (l = 1, 2,. . . ,p). So, the efficiency matrix can be represented as follows:

$$D = \left[E_{lj}\right]_{n\times p} \tag{22}$$

In order to take PCA on the efficiency data, several steps are carried out as following:

Step 1: Calculate the sample correlation matrix

$$R = \left[r_{li}\right]_{p\times p} \tag{23}$$

Where $r_{li} = s_{li}/\sqrt{s_{ll}.s_{ii}}$ and sample covariance. $S_{li} = \frac{1}{n-1}\sum_{j=1}^{n}(E_{lj} - \bar{E}_l)(E_{ij} - \bar{E}_i)$

Step 5: Compute eigenvalues $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p$, $\sum_{l=1}^{p} \lambda_l = p$ and corresponding normalized eigenvectors . $\xi_1, \xi_2, ..., \xi_p$

Step 6: Compute and Select the principal components

$$PC = D\left[\xi_1, \xi_2, ..., \xi_p\right] = \left[PC_1, PC_2, ..., PC_p\right] \tag{24}$$

Herein, it is remarkable that any two different principal components are uncorrelated with each other, which shows that there is no information superposition between the two. The variance of any two principal components is λ_i. The relative importance of principal component PCi can be expressed by the proportion $\lambda_i / \sum_{j=1}^{p} \lambda_j = \lambda_i \big/ p$. Generally, the principle to choose the first M principal components may be to satisfy $\lambda_M \succ 1$, or cumulative contribution utility of dispersion $\frac{1}{P}\sum_{i=1}^{M} \lambda_i \geq \rho$, e.g. $\rho = 0.8$.

Shanmugam & Johnson (2010) pointed out that it was no longer appropriate to directly rate DMUs via PCA, for the output–input ratios did not meet a Gaussian distribution as required in PCA. However, under DEA structure, it does not need any distribution assumptions of factors (ratios, input, or outputs). Furthermore, all the chosen principal components can be treated as desirable outputs in DEA models, for the more prefer each DMUj, the bigger PCj l in the terms of lth one. Correspondingly, a dummy (one) can be seen as a virtual input for all DMUs. However, outputs of original DEA models need to be strictly positive, while the elements of the chosen principal components can be negative. So, a linear monotone increasing data transformation is made to the negative results of PCA to avoid being negative by the equation:

$$Z_{lj} = PC_{lj} + Q \tag{25}$$

Where $Q = -\min\left\{PC_{lj}\right\} + 1$ is a common choice to ensure that all transformed values are positive. As a result, all the elements of the chosen principal components are equally increased by the same increment. Pastor (2002) proved that the input-oriented BCC model was output translation invariant. Knox Lovell & Pastor (2006) further proved that an input-oriented CCR model with a single constant input (or dummy) coincided with the input-oriented BCC model. To evaluate operational efficiency of DMU0, a simplified input-oriented CCR model is proposed as follows:

$$\begin{aligned} &Max \quad W_0 = \sum_{l=1}^{M} P_l Z_{l0} \\ &st: \\ &\sum_{l=1}^{M} P_l Z_{lj} \leq 1, \qquad j = 1,2,....,n \\ &P_l - P_{l+1} \geq \varepsilon_l \qquad l = 1,2,...,M-1 \\ &P_l \geq 0 \qquad\qquad l = 1,2,...,M \end{aligned} \tag{26}$$

where P_l is the weight attached to the outputs Z_{lj},(j = 1, . . . ,n), and the weight constraints $P_l - P_{l+1} \geq \varepsilon_i$ represent the facts that the lth principal component carries the total dispersion more than the (l + 1)th one does.

5. Practical example

Consider a bank which has 10 branches and the bank supervisor is going to predict the efficiency of its branch and also the risk that every branch manager should accept for reaching the predicted efficiency from the respect of the allocated budget, for the future financial year. It should be said that in this system the branches inputs (personnel expenses, official expenses and the costs of the place renting (suppose the branches places are being leased)), are programmed at the end of each year for the future financial year in the framework of budgeting system. But about the branches output (the amount of granted facilities and the flow of inter-bank services), predicting is conducted based on PERT/CPM technique by using the output values of each branch in the last financial years. The branches

inputs and the estimated outputs are defined in table 1 by the supervisor as in the form of the optimistic estimate (OP), the most likely estimate (ML) and pessimistic estimate (PE). The supervisor is going to do the reforming measures for improving the efficiency of his under control set by estimating the efficiency and its dependent risk for each branch and for the future financial year. The expected efficiency of the supervisor is considered 1 for all the decision making units.

Table 1. The budgeted inputs and output estimates

Branch code	Inputs			Outputs					
	Official expenses	personal expenses	Rent costs	granted facilities ($\hat{y}_1$)			inter-bank services ($\hat{y}_2$)		
				OP	*ML*	*PE*	*OP*	*ML*	*PE*
1	24	46	298	5800	5027	4800	410	362	300
2	25	41	295	5920	4972	4910	430	356	310
3	32	40	300	5750	5019	4952	421	353	320
4	33	44	305	5610	5083	4823	412	354	313
5	27	46	296	5520	5088	4899	418	367	304
6	21	42	297	5742	5010	4962	429	347	310
7	19	38	301	5825	5017	4898	432	346	317
8	22	39	292	5912	4970	4992	409	353	326
9	24	45	294	5852	4994	4901	399	352	309
10	20	41	306	5712	5031	4925	415	349	311

5.1 Computational results

First, we estimate the expected value and related standard deviation of outputs for each branch of bank with PERT/CPM technique by using eq. 20 & 21. Results are obtained as shown in table 2.

Table 2. Mean and standard deviation of output estimates

Bank branch	$\overline{y}_{1j}$	$\overline{y}_{2j}$	b_{1j}	b_{2j}
1	5118	360	12.9	4.28
2	5120	361	12.9	4.47
3	5130	359	11.5	4.10
4	5128	357	11.4	4.06
5	5129	365	10.1	4.35
6	5124	355	11.4	4.45
7	5132	356	12.4	4.37
8	5131	358	12.3	3.71
9	5122	353	12.53	3.87
10	5127	354	11.4	4.16

Since the proposed model based on the normal distribution assumption for outputs, normal probability plot used with Stat Graphics plus 2.1 software for outputs of branches that results are shown in Fig. 1.

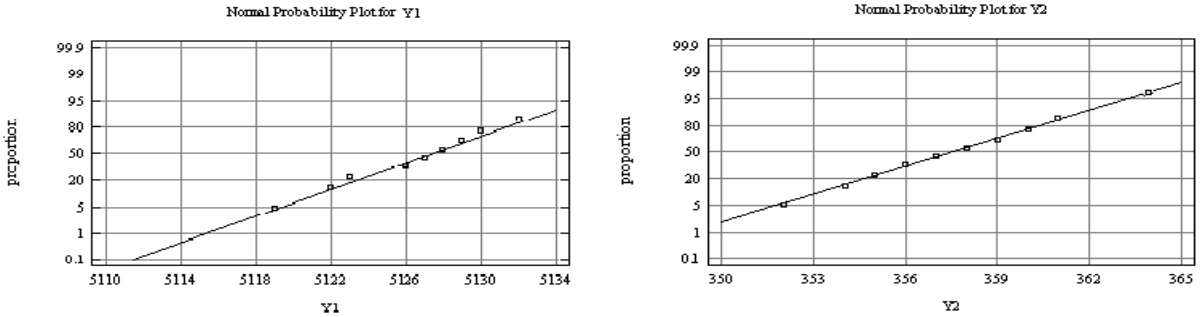

Figure 1. Normal plot for outputs (y_1 and y_2)

In this example, data sets of bank's outputs be positive and similar be negative for inputs inherently. Consequently, inputs and outputs values were normalized. Also in order to know the subjective preferences of bank

experts about the weights of each output, the questionnaires were distributed among four individual of deputies of the bank supervising and finally the preferred weight of each output was obtained by using eq. 15 as following:

Table 3. Outputs weights

Outputs	weight
granted facilities	**0.56**
inter-bank services	**0.35**

The bounded constraint for each output by using eq. 18 as following:

$$\left.\begin{array}{l} u_1 \geq 0.56\alpha \\ u_1 \leq 2 \times 0.56 - 0.56\alpha \end{array}\right\} \quad \text{Constraint for } (\hat{y}_1)$$

$$\left.\begin{array}{l} u_2 \geq 0.35\,\alpha \\ u_2 \leq 2 \times 0.35 - 0.35\,\alpha \end{array}\right\} \quad : \quad \text{Constraint for } (\hat{y}_2)$$

The proposed SDEA/PCA model in section 4, involves variables Wj, Pl and Zlj. The goal of the optimization of this model is estimate efficiency mean (Wj) for each branch of bank. Since Wk represent predicted efficiency mean for k-th branch of bank. First, we need to obtain various efficiency level (Ej) with various risk level (α) by running proposed model (eq. 19) with considering above constraints. We set the parameters β=1 in this model. The results of 10 trails on proposed model for each branch of bank presented in Table 4. These results have obtained according to different risk levels for branches.

Table 4. Amount of efficiency with various risk levels on proposed model Bank's branches

Bank branch	$\beta = 1$						
	$\alpha_1 = 0.05$	$\alpha_2 = 0.1$	$\alpha_3 = 0.2$	$\alpha_4 = 0.5$	$\alpha_5 = 0.8$	$\alpha_6 = 0.9$	$\alpha_7 = 0.95$
1	0.982	0984	0.985	0.99	0.992	0.995	0.996
2	0.953	0.955	0.957	0.962	0.967	0.971	0.972
3	0.965	0.967	0.969	0.975	0.983	0.986	0.989
4	0.972	0.975	0.979	0.985	0.992	0.995	0.997
5	0.991	0.993	0.994	0.998	1	1	1
6	0.952	0.954	0.959	0.965	0.97	0.975	0.977
7	0.96	0.964	0.969	0.976	0.984	0.988	0.989
8	0.934	0.937	0.941	0.949	0.955	0.96	0.962
9	0.958	0.961	0.964	0.969	0.973	0.978	0.98
10	0.961	0.965	0.969	0.975	0.979	0.984	0.986

Table 5 gives the results via the PCA process as suggested in Section 4. According to the principle ($\lambda_M \succ 1$), 2 eigenvalues (the others are omitted, for being smaller than 1) have been selected, whose cumulative contribution utility of dispersion is 91%.

Table 5. Eigenvalues and principal components for risk levels

Risk level	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
α_1	-0.376	0.502	0.508	-0.085	-0.208	0.384	0.388
α_2	-0.378	0.414	0.168	0.070	0.256	-0.724	0.251
α_3	-0.380	0.228	-0.411	0.262	0.374	0.525	-0.388
α_4	-0.381	0.067	-0.614	0.034	-0.620	-0.198	0.219
α_5	-0.379	-0.279	0.031	-0.829	0.003	0.067	-0.291
α_6	-0.378	-0.411	-0.071	0.079	0.515	-0.086	0.636
α_7	-0.375	-0.523	0.403	0.473	-0.319	0.032	-0.312
Eigenvalue	6.8745	3.1141	0.5557	0.4542	0.0008	0.0004	0.0003

After the transformation on the negative principal components by using eq. 25, the new transformed outputs for the first 2 principal components have been chosen as seen in the Table 6.

Table 6. The efficiency mean based upon the SDEA/PCA model

Bank branch	Z_{1j}	Z_{2j}
1	1.01968	3.62256
2	1.0904	3.61713
3	1.05373	3.6122
4	1.03027	3.61165
5	1	3.62701
6	1.07723	3.60131
7	1.05523	3.60742
8	1.12784	3.60872
9	1.073	3.61545
10	1.05939	3.61288

The results of predicted efficiencies mean were obtained with proposed SDEA/PCA model by using eq. 26, which table 7 represented it.

Table 7. The efficiency mean based upon the SDEA/PCA model

Branch	Efficiency (W_j)
1	0.994
2	0.97
3	0.985
4	0.992
5	1
6	0.971
7	0.986
8	0.958
9	0.975
10	0.98

In order to verify the our model performance, the real efficiencies were obtained with DEA model and real outputs (eq.1) for all of branches in finish the predicted financial period and results of these were compared with results of predicted efficiencies were obtained with proposed SDEA/PCA model, that table 8 represented it.

Table 8. Comparison between real efficiencies (with DEA Model) and predicted efficiencies (with proposed SDEA/PCA model)

Branch	granted facilities $(\hat{y}_1)$	inter-bank services $(\hat{y}_2)$	Real efficiency	Predicted efficiency
1	5132	371	0.995	0.994
2	5134	372	0.971	0.97
3	5144	370	0.986	0.985
4	5142	368	0.996	0.992
5	5143	376	1	1
6	5138	366	0.972	0.971
7	5146	367	0.987	0.986
8	5145	369	0.959	0.958
9	5136	364	0.976	0.975
10	5141	365	0.989	0.98

Correlation rate between real efficiencies and predicted efficiencies = 0.9792

The correlation rate between real efficiencies and predicted efficiencies is calculated with Stat Graphics plus 2.1 software. The high correlation rate (**0.9792**) has obtained represents the validity of proposed SDEA/PCA model.

6. DISCUSSION AND CONCLUSION

This study proposed a mixed model based on Stochastic Data Envelopment Analysis (SDEA) and Principal Component Analysis (PCA) for predicting efficiencies of similar units in an organization. It incorporated future information on unit's outputs into analytical framework. This model was tried to cover the most important shortcoming of Data Envelopment Analysis (DEA) by entering the stochastic effect of output variables and fuzzy weights for input/outputs weights in order to prevent the unreal distribution of weights with using of Stochastic DEA (SDEA) technique. Principal Component Analysis (PCA) was used to determine efficiency mean for units with various risks.

In order to reach a better understanding of the proposed SDEA/PCA model, it was applied to predict efficiencies for a number of Iranian Bank branches. In order to verify the proposed model performance, the real efficiencies were obtained with real outputs and DEA model for all of branches in finish the predicted financial period. This results were compared with the results of predicted efficiencies were obtained with proposed model. The high correlation (0.9792) between real and predicted efficiencies was obtained which represented the validity of the proposed SDEA/PCA model.

In this proposed model, normal distribution was applied for output stochastic variables. It is a straight forward matter to conduct a statistical test in the framework of SDEA/PCA analysis and the normal distribution. It is recommended to examine whether other distributions can be used for future analysis and apply them for forming the new models.

REFERENCES

1. Adler, N., & Golany, B. (2003). Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operations Research Society of Japan*, 46, pp. 66–73.
2. Bruce Ho, C., Wu, D.D. (2008). Online banking performance evaluation using data envelopment analysis and principal component analysis. *Journal Computers & Operations Research*, 25, pp. 125-135.
3. Charns, Cooper, Rouds, K, (1978). Data Envelopment Analysis, References and DEA solver software, Kluwer Academic Press.
4. Charnes A., W. W. Cooper, Q. L. Wei and Z. M. Huang, (1989), Cone-Ratio data Envelopment Analysis and Multi-Objective Programming, 20, pp. 1099-1118.
5. Chen, C. T. (2007). Extensions of the TOPSIS for group decision-making under fuzzy environment. *Fuzzy Sets and Systems*, 114, 1–9.
6. Cooper, WW., Huang, Z.M., Li, SX. (1996). Satisfying DEA models under chance constraints. *Annals of operation research*, 66, PP 279-296.
7. Cooper, WW., Deng, H., Huang, Z., and Li, SX. (2002). Chance constrained programming approach to technical efficiencies. *Journal of the Operational Research Society*, 53, PP 1347-1354.
8. Cooper, WW., Huang, ZM, Lelas, V., Li, SX., Olesen, OB. (2006). Chance constrained programming formulations for stochastic characterizations of efficiency and dominance in DEA. *Journal of productivity analysis*, 9, PP 53-79.
9. Dyson R. G. and E. Thanassoulis. (1988). Reducing Weight Flexibility in Data Envelopment Analysis. *Journal of Operational Research Society*, 39, PP. 563-576.
10. Huang, ZM., Li, SX (2005). Chance Constrained Programming and Stochastic DEA Models. *Proceeding of the Decision Sciences Institute*, PP 447-449.
11. Jahanshahloo G., M. Alirezaee, S. Saati and S. Mehrabian. (1997). The Role of Bounds on Multipliers in DEA; with an Empirical Study. *Journal of Sciences, Islamic Azad University*, 19, PP. 331-347.
12. Kao, L.J., Lu, C.J., Chiu, C., (2011). Efficiency measurement using independent component analysis and data

envelopment analysis. *Journal of the Operations Research*, 210, pp. 310–317.

13. Knox Lovell, C. A., & Pastor, J. T. (2006). Radial DEA models without inputs or without outputs. *European Journal of Operational Research*, 118(1), 46–51.
14. Lai Young-Jon, Hwang ching-lai. (1992). Fuzzy mathematical programming methods and Applications. *Springer-verlag*.
15. Land, K., Lovell, C.A.K., Thore, S (1993). Chance constrained data envelopment analysis. *Managerial and decisional economics*. 14, PP. 541-544.
16. Pastor, J. T. (2002). Translation invariance in data envelopment analysis: A generalization. *Annals of Operations Research*, 66, 93–102.
17. Premachandra, I. M. (2007). A note on DEA vs. principal component analysis: An improvement to Joe Zhu's approach. *European Journal of Operational Research*, 132, 553–560.
18. Roll Y. and B. Golany, (1993). Alternate Methods of Treating Factor Weights in DEA. *Omega*, 21, PP. 99-109.
19. Saati S., A. Memariani and G. R. Jahanshahloo, (2003). A Procedure for Finding a Common set of Weights in Fuzzy DEA. *Fuzzy Sets and Systems*,25, 235-245.
20. Sabzehparvar, M. (2002), Project control. Tehran, Khaniran Association.
21. Sengupta, JK. (1982). stochastic programming. *International journal of system science*, 7, PP. 822-835.
22. Shanmugam, R., Johnson, C.,(2010). At a crossroad of data envelopment and principal component analysis. *Omega*, 35(4), 2009, 351-364.
23. Shavandi, H. (2005). Theory of fuzzy sets in industrial engineering and management. Tehran, Gostaresh Oloom Payeh,
24. Zadeh, L. A. (1965). Fuzzy sets. Inform and Control, 8, 338–353.
25. Zhu, J. (2005). Data envelopment analysis vs. principal components analysis: An illustrative study of economic performance of Chinese cities. *European Journal of Operational Research*, 111, 50–61.